POLITECNICO DI TORINO

Laurea Magistrale in Ingegneria Informatica
Automation and Intelligent Cyber-Physical Systems

Tesi Magistrale

Costruzione e Validazione di Dataset di ricette e Ingredienti Culinari

Relatori

Prof. Maurizio MORISIO

Prof. Andrea BOTTINO

Candidato

Samuele PINO

26 Luglio 2024

Sommario

Il cibo riveste un ruolo indispensabile per la sopravvivenza umana. Oltre a fornirci energia, esso è anche una parte essenziale della nostra identità e cultura.

Nel corso del tempo però i modelli alimentari e la cultura della cucina stanno evolvendo. In passato, il cibo veniva principalmente preparato in casa, ma oggi spesso ci affidiamo ad altri, come servizi take-away, fast food e ristoranti.

Di conseguenza, l'accesso a informazioni dettagliate sui cibi consumati fuori casa è limitato, rendendo così difficile conoscere esattamente cosa stiamo consumando. In alcuni contesti, come la prevenzione di malattie, il fitness, il mantenimento di una dieta bilanciata e la corretta gestione di allergie e intolleranze specifiche, risulta molto importante monitorare quello che una persona mangia. In particolare è necessario identificare i diversi ingredienti che compongono un piatto e le loro quantità, grazie alle quali, è possibile calcolare i valori nutrizionali (grassi, proteine, carboidrati, ecc.) presenti in qualsiasi cibo consumato.

Un modo pratico per effettuare queste analisi sarebbe quello di scattare una foto del piatto che si intende consumare, utilizzando uno smartphone, e attraverso l'impiego di un classificatore automatico in grado di riconoscere la ricetta, riuscire a risalire agli ingredienti di cui è composta, e stimare in un secondo momento anche al quantità presente.

La presente tesi di laurea si inserisce in questo contesto, come un lavoro di creazione e validazione di una base dati, contenente numerose ricette e le rispettive liste di ingredienti. Questo database si presenta quindi come un ampio registro di ricette, da utilizzare come fonte di dati in una fase successiva al riconoscimento della ricetta, tramite fotografia. Una volta identificata la ricetta i valori nutrizionali potranno essere ottenuti interrogando propriamente questo database.

A partire da numerose ricette raccolte tramite Web Scraping sui principali siti di cucina italiani, i dati grezzi sono stati aggregati e in una struttura più ordinata, andando a creare il dataset di ricette vero e proprio.

Successivamente questo dataset ha affrontato delle fasi in cui i dati sono stati organizzati meglio, e puliti da impurità come duplicati, eterogeneità nelle chiavi, errori etc. Inoltre è stata effettuata una analisi per la ricerca e rimozione di pseudo-duplicati, ricette presentate con nome diverso, ma sostanzialmente identiche negli ingredienti. Per farlo è stato costruito uno script dedicato per questo scopo.

Infine è stato selezionato un database di ingredienti culinari che contenesse le relative informazioni nutrizionali. Questo database è poi stato aggiunto ed integrato nel progetto, permettendo di avere numerose informazioni nutrizionali per ogni ingrediente.

Indice

Elenco delle tabelle

Elenco delle figure

Acronimi

LD
: Levenshtein Distance

SW
: Sliding Window

WSW
: Weighted Sliding Window

LLM
: Large Language Model

Capitolo 1

Introduzione

1.1 Analisi del problema

Il cibo è uno degli aspetti fondamentali dell'esperienza umana. Il suo consumo è intrinsecamente legato alla nostra salute, alle nostre emozioni e alla nostra cultura. Come si dice "siamo ciò che mangiamo" e le attività connesse al cibo come cucinare, mangiare e discuterne, occupano una parte significativa della nostra quotidianità.

Nell'era digitale attuale, l'accesso a una vasta gamma di risorse multimediali ha facilitato la diffusione della cultura culinaria. I video, siti web e tutorial dedicati alla preparazione di ricette varie sono diventati sempre più diffusi. Queste risorse forniscono informazioni dettagliate e passo-passo su come realizzare piatti di ogni genere, consentendo alle persone di esplorare e sperimentare in cucina. Inoltre, i social media, permettono alle persone di condividere foto e video dei piatti che preparano e consumano. Una semplice ricerca su Instagram con l'hashtag 'food' porta a più di 500 milioni di post, evidenziando il valore innegabile che il cibo ha nella nostra società. Inoltre, è necessario sottolineare come, soprattutto negli ultimi anni, si è assistito ad una diffusione sempre più ampia del concetto di alimentazione salutare.

La consapevolezza dell'importanza di una dieta equilibrata e nutriente per il benessere generale è in costante aumento. Tale tendenza può essere indirettamente riscontrata attraverso i social network, per esempio, l'hashtag 'healtyfood' produce oltre 110 milioni di risultati su Instagram.

L'adozione di una dieta salutare non solo impatta positivamente sulla salute delle persone, ma ha anche importanti implicazioni a livello sociale ed ambientale,

promuovendo pratiche agricole sostenibili e il benessere degli animali. Per il mantenimento di uno stile di vita e di un regime alimentare più salutare è sempre più necessario quindi, tenere traccia dei diversi valori nutrizionali, (proteine, zuccheri, carboidrati), che caratterizzano in piatti consumati, operazione complessa che viene complicata ulteriormente quando si scelgono cibi preparati senza la cura domestica, come i piatti di ristoranti, fast food e take away, per i quali l'accesso a informazioni dettagliate sui cibi preparati è molto più limitato.

1.2 Obiettivi tesi

Per aiutare quindi le persone in questo processo di tracciamento dei valori nutrizionali dei cibi consumati, si vuole realizzare un'applicazione che, attraverso una fotografia o un'immagine fornita dall'utente, sia in grado di riconoscere il cibo rappresentato, gli ingredienti di cui è composto e il volume di quest'ultimi, per poterne calcolare i valori nutrizionali associati. L'idea si ispira ad un'applicazione già esistente, chiamata FatSecret, che fornisce delle statiche basate sui valori nutrizionali dei piatti consumati dall'utente, attraverso l'inserimento testuale del nome e delle quantità di ogni piatto.

La presente tesi di laurea si focalizza in particolare sull'analisi dei dati riguardanti le ricette, specialmente quelle più diffuse in Italia. La tesi stessa si sviluppa nelle seguenti fasi:

- Analisi preliminare, rimozione duplicati e creazione di una pipeline di preprocessing per migliorare la qualità dei dati, costruita ad hoc per le ricette italiane;
- Ricerca a rimozione di pseudo-duplicati, ricette differenti nel nome, ma pressochè identiche nelle liste di ingredienti.
- Ricerca, selezione e implementazione di un database di nutrienti per ogni ingrediente, con forte focus per il contesto della tradizione italiana;

1.3 Contributo Tesi

Per il raggiungimento di tali obiettivi, lo sviluppo della tesi è stato suddiviso in due macro fasi. La prima dedicata principalmente all'analisi e alla manipolazione del dataset di ricette. La seconda incentrata sull'analisi di alcuni database di ingredienti culinari, con successiva selezione del più idoneo e la sua implementazione

La prima fase è stata dedicata principalmente all'analisi, alla pulizia e alla manipolazione del dataset di ricette, con l'obiettivo di comprendere le caratteristiche principali delle stesse. Ciò ha richiesto la pulizia, la trasformazione dei dati unitamente alla ricerca e alla rimozione di ricette pseudo-duplicate. In particolare, sono state utilizzate tecniche di data preprocessing nel campo testuale contenente le informazioni sull'ingrediente (ad esempio name: "500 g di farina 00 per l'impasto), per rimuovere le informazioni ridondanti e normalizzare i dati, in modo da rendere più efficienti le analisi e gli utilizzi futuri.

La seconda fase invece è incentrata sull'analisi, sulla selezione e sulla successiva implementazione di un database di ingredienti culinari con relative informazioni nutrizionali.

La prima fase del progetto si articola più precisamente nei seguenti tre passaggi:

- Innanzitutto è stata eseguita una prima analisi esplorativa del dataset al fine di comprenderne meglio la struttura e la distribuzione degli ingredienti, ma soprattutto per stabilire numericamente alcune caratteristiche come le statistiche riguardanti la lunghezza delle liste di ingredienti o il numero di parole presenti nel campo testuale dell'ingrediente.

- Successivamente è stata costruita una pipeline di preprocessing specificamente progettata per le ricette italiane. Questa pipeline si occupa di preparare, modificare e analizzare i dati per renderli più strutturati e utilizzabili estraendo importanti informazioni come il nome o la quantità di ogni ingrediente.

  La fase di pulizia ha previsto la standardizzazione dei formati, la gestione dei valori mancanti o errati, al fine di garantire la qualità del dataset. La fase di manipolazione e analisi, ha invece comportato l'analisi testuale delle ricette per estrarre in modo preciso e uniforme gli ingredienti delle ricette e le quantità corrispondenti.

- Una volta raffinato il dataset ed estratto le informazioni necessarie degli ingredienti, esso è stato sottoposto alla ricerca di pseudo-duplicati, ricette sostanzialmente identiche a livello di ingredienti, ma con nome differente. Per individuare queste coppie di ricette, sono state analizzate tutte le coppie di ricette, andando a confrontare le loro liste di ingredienti, testando inoltre diversi metodi per questa ricerca.

Nella seconda fase invece è stata effettuata la ricerca di alcuni database di ingredienti, che riportassero anche informazioni nutrizionali a riguardo. I candidati individuati sono stati successivamente analizzati secondo alcune metriche, come l'attinenza agli ingredienti italiani, il livello di dettaglio, l'autorevolezza e altro. Questa ampia analisi comparativa, in funzione delle metriche decise, ha permesso di individuare il candidato ideale e di implementare questo database nel lavoro.

1.4 Panoramica Dettagliata del Report

Il report si articola come descritto dallo schema riportato in figura 1.1. Il colore verde è stato impiegato per indicare gli input esterni al lavoro. In questo caso, il database di valori nutrizionali, il dataset di ricette e il file di categorie personalizzate creato. L'arancione indica invece gli output effettivi generati da questo lavoro di tesi, sia di codice (come la pipeline o lo script per la rimozione degli pseudo-duplicati) sia di dati.

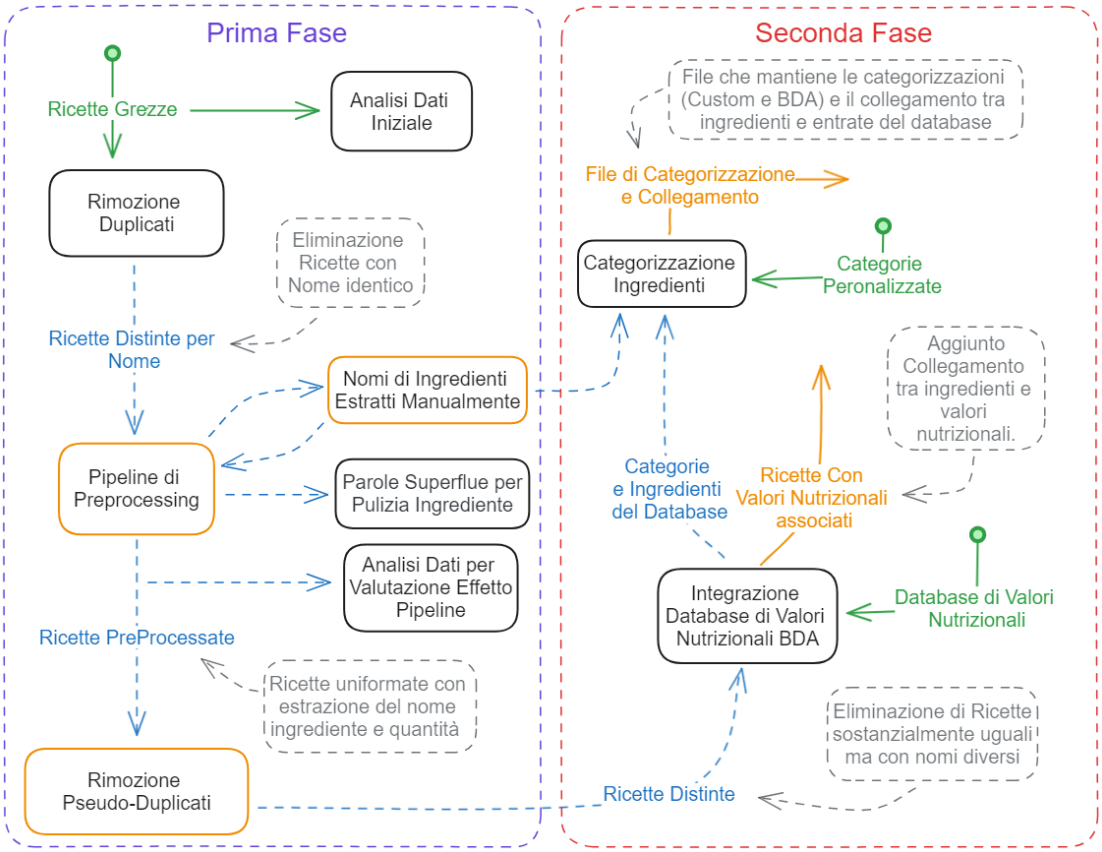


Figura 1.1: Schema del presente report

1.4.1 Analisi Dati Iniziale

Innanzitutto è stata eseguita una prima analisi esplorativa del dataset al fine di comprenderne meglio la struttura e la composizione.

Il dataset iniziale infatti, presentava delle ricette, ognuna composta da diversi campi, come il nome della ricetta, la lista degli ingredienti, le istruzioni di preparazione, la portata o le immagini associate alla ricetta stessa.

Dopo aver schematizzato adeguatamente la struttura del dataset, ulteriori analisi hanno riguardato principalmente la copertura che ogni ingrediente ha rispetto al numero di ricette totali a livello percentuale. In altre parole quante volte appare un determinato ingrediente, come ingrediente di una ricetta all'interno del dataset.

Questo tipo di analisi del dataset ha semplificato e permesso alcune analisi manuali dei dati andando a volgere l'attenzione su quegli ingredienti del dataset più impattanti a livello percentuale.

Un secondo tipo di analisi mira invece a stabilire numericamente alcune caratteristiche delle liste degli ingredienti, come le statistiche riguardanti la lunghezza delle stesse o il numero di parole presenti nel campo testuale dell'ingrediente.

Questa seconda analisi ha permesso invece di stabilire dei termini di paragone per poter confrontare queste statistiche con quelle misurate nuovamente sui dati usciti dalla pipeline di preprocessing.

1.4.2 Rimozione Duplicati

Notoriamente la rimozione dei duplicati dal dataset è essenziale per ottenere risultati di analisi accurati e robusti, poiché ciò permette di eliminare sovrastime di ingredienti e distorsioni, garantendo un dataset pulito e coerente che riflette fedelmente la struttura dei dati.

In questo caso, per rimozione dei duplicati, si intende quelle ricette che appaiono con lo stesso identico nome.

1.4.3 Pipeline di Preprocessing

La prima macro fase della pipeline ha come scopo quello di manipolare i dati, eliminare gli errori e le impurità, e si articola in alcune fasi intermedie.

Una prima fase intermedia, ha come scopo quello di andare ad uniformare le varie chiavi delle diverse tabelle ad uno standard unico.

Successivamente avviene una analisi testuale che estrae la quantità insieme all'unità di misura per ogni ingrediente.

Il passo successivo è quello di una pulizia del campo ingrediente, rimuovendo le parole superflue, seguito poi da una fase di estrazione del nome dell'ingrediente utilizzando una tabella di ingredienti (estratti manualmente dal dataset) e la distanza Levensthein (LD), come metrica di similarità tra stringhe.

Una volta ottenuti i vari ingredienti, vengono gestite le eccezioni rappresentate da quegli ingredienti che non hanno una quantità ben definita in grammi o millilitri.

Essi sono etichettati come 'gen', 'err' o 'qb' e sono reintegrati nel dataset con un processo specifico.

Infine la quantità di ogni ingrediente è poi normalizzata sul peso totale in grammi della ricetta, e viene aggiunta la percentuale coperta dal singolo ingrediente sui 100 grammi di ricetta.

Un esempio di un ingrediente in ingresso nella pipeline come [Nome_ing: "500 g di farina 00 per l'impasto] risulterebbe così in uscita: [name: "farina", quantità: 500, unità: 'g', norm_on_100g: 38.9 ...].

1.4.4 Ricerca e Rimozione Pseudo-Duplicati

Gli pseudo-duplicati sono qui definiti come ricette che, nonostante riportino nomi diversi, presentano gli stessi ingredienti e sono fondamentalmente identiche o estremamente simili.

Questo tipo di duplicati può essere particolarmente insidioso, poiché non vengono solitamente rilevati dalle tecniche di rimozione dei duplicati tradizionali, che si basano sulla comparazione esatta dei valori dei campi all'interno delle ricette. La presenza di pseudo-duplicati può quindi portare una generale distorsione dei risultati delle analisi successive, al pari della presenza di duplicati semplici.

Le analisi effettuate sulle ricette, coinvolgono principalmente le rispettive liste di ingredienti, poiché una semplice analisi di similarità sui titoli o sulla portata può facilmente portare a errori nell'identificare questi pseudo-duplicati.

Per individuare queste coppie di ricette pseudo-duplicate, sono state analizzate tutte le possibili combinazioni di due ricette, andando quindi a confrontare le loro liste di ingredienti. Per farlo sono stati testati alcuni approcci effettuando alcuni esperimenti. Ad essere esaminati sono stati tre algoritmi di confronto degli ingredienti e due diversi ordinamenti delle liste di ingredienti.

1.4.5 Analisi Database di Ingredienti e Integrazione

Dopo aver selezionato quattro database di ingredienti, sono state decise alcune metriche per valutarli al meglio così da poter scegliere accuratamente il database da cui attingere i valori nutrizionali corrispondenti per ogni ingrediente.

Le metriche in questione sono: completezza su ingredienti italiani, livello di dettaglio delle informazioni, accessibilità (diretta o legale e tecnica), autorevolezza, prezzo, dimensione e aggiornamento.

Una volta fatto ciò, per arricchire le ricette con queste nuove informazioni, è stato integrato il suddetto database di valori nutrizionali con tutti gli ingredienti presenti nelle ricette del nostro dataset, attraverso un file di collegamento utilizzato come supporto. Questa integrazione consente di aggiungere, indirettamente, nuove

feature alle ricette, come ad esempio il contenuto di calorie, proteine, grassi e carboidrati.

Capitolo 2

Analisi Dati Iniziale e Rimozione Ricette Duplicate

In questo primo capitolo sono riportate le analisi preliminari effettuate sul dataset delle ricette, unitamente ai risultati corrispondenti. Dopo aver rimosso le ricette duplicate dal dataset, sono state effettuate in particolare due analisi

Una prima analisi generale del dataset, volta a comprenderne la struttura e i dati, nonché la distribuzione interna degli ingredienti rispetto alle ricette.

Una seconda analisi riguarda invece lo studio delle distribuzioni di alcune caratteristiche, precisamente la lunghezza delle liste di ingredienti, presenti all'interno delle ricette, e il numero di parole presenti nel campo testuale del singolo ingrediente, a descriverlo.

2.1 Presentazione del Dataset Iniziale

Il dataset di partenza era composto approssimativamente da 10.000 ricette, acquisite attraverso un approccio di data mining e web scraping. In particolare, gli strumenti impiegati per raccogliere in modo efficiente le informazioni necessarie, sono stati Parse Hub e Google Image Scraper.

La struttura era analoga a quella utilizzata per il dataset Recipe1M, benchmark principale per i modelli state of the art nel campo del Recipe retrieval.

Ciascuna ricetta nel dataset era composta da diversi campi, come il nome della ricetta, gli ingredienti, le istruzioni di preparazione, la portata e le immagini associate alla ricetta.

In particolare, i siti web selezionati per la raccolta dei dati sono stati Cook Around, Fatto in casa da Benedetta, Giallo Zafferano e Misya. Questi fonti web sono state selezionate per la loro popolarità e affidabilità, nonché per la varietà e la quantità di ricette disponibili.

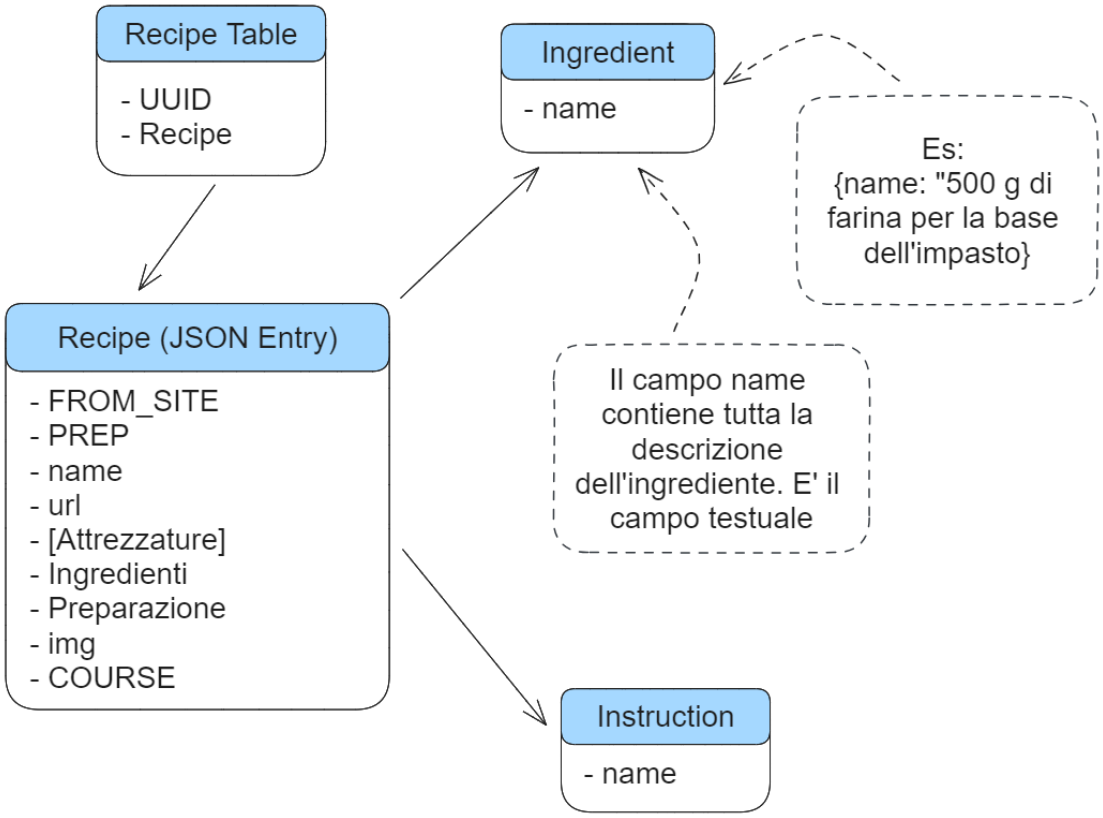


Figura 2.1: Schema UML dei Dati iniziali

In particolare le ricette riportavano dei campie eterogenei, come "name" talvolta indicato con "titolo" o "Title" o "Ricetta", o ancora, "Ingredienti" indicato alle volte come "ingr". Negli ingredienti, il campo testuale "name", non era sempre tale, ma si presentava anche come "ingr", "Name" etc.

2.1.1 Analisi Statistica degli Ingredienti

L'analisi statistica riguardante il numero di ingredienti e il numero di parole che compongono il singolo ingrediente, è stata effettuata per poter confrontare successivamente i dati qui ottenuti con quelli ricalcolati subito dopo alcune fasi di pulizia e manipolazione dati, affrontate nelle sezioni successive (vedi sezione 3).

Distribuzione del numero di ingredienti per ricetta

La prima analisi statistica condotta sul dataset di ricette culinarie ha riguardato il numero di ingredienti necessari per ogni ricetta.

Per ogni ricetta è stato calcolato il numero di ingredienti presenti, e successivamente si è calcolato la frequenza con cui ricette con lo stesso numero di ingredienti appaiono nel dataset. Le analisi hanno rivelato una distribuzione non uniforme del numero di ingredienti per ricetta.

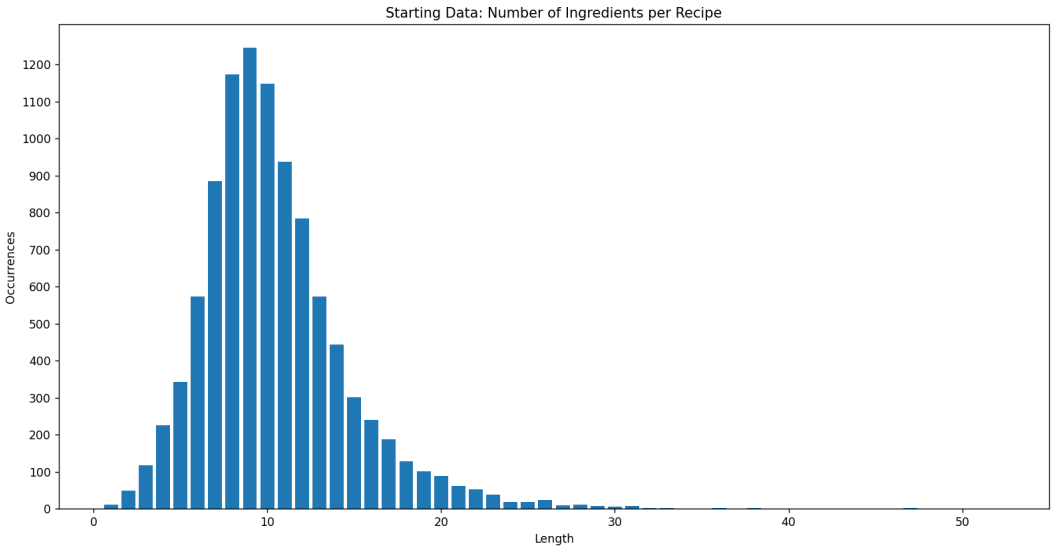


Figura 2.2: Distribuzione del numero di ingredienti per singola ricetta. Dall'istogramma notiamo come la maggior parte delle ricette presentino un numero di ingredienti vicino a 10.

In particolare, la maggior parte delle ricette (circa l' 85.5%) presenta un numero di ingredienti compreso tra 5 e 15 (estremi compresi), mentre solo il 10.4% delle ricette presenta più di 15 ingredienti. Solo il 4.1% sono invece ricette con un numero di ingredienti sotto i 5.

La media del numero di ingredienti per ricetta è risultata essere di 10,4, con una deviazione standard di 13.2. La moda risulta 9 mentre la mediana 10.

Questa distribuzione suggerisce che la maggior parte delle ricette sono relativamente semplici e richiedono un numero limitato di ingredienti, mentre solo alcune ricette più complesse ed elaborate necessitano di un numero più elevato di ingredienti.

Numero di parole presenti campo testuale degli ingredienti

La seconda analisi effettuata riguarda invece il numero di parole presenti nel campo testuale dell'ingrediente. Questa metrica è importante perché permette di definire quante parole sono usate per descrivere un singolo ingrediente. I campi testuali che

descrivono ingredienti con troppe parole contengono molto probabilmente parole superflue, che non riguardano l'ingrediente stesso ma delle caratteristiche secondarie (come temperatura, colore, presenza o meno di noccioli etc.), sostanzialmente inutili o poco rilevanti per quelli che sono qui gli scopi del lavoro.

In altre parole questa metrica permette di identificare in maniera qualitativa una prima misura di rumore riguardante il campo testuale degli ingredienti.

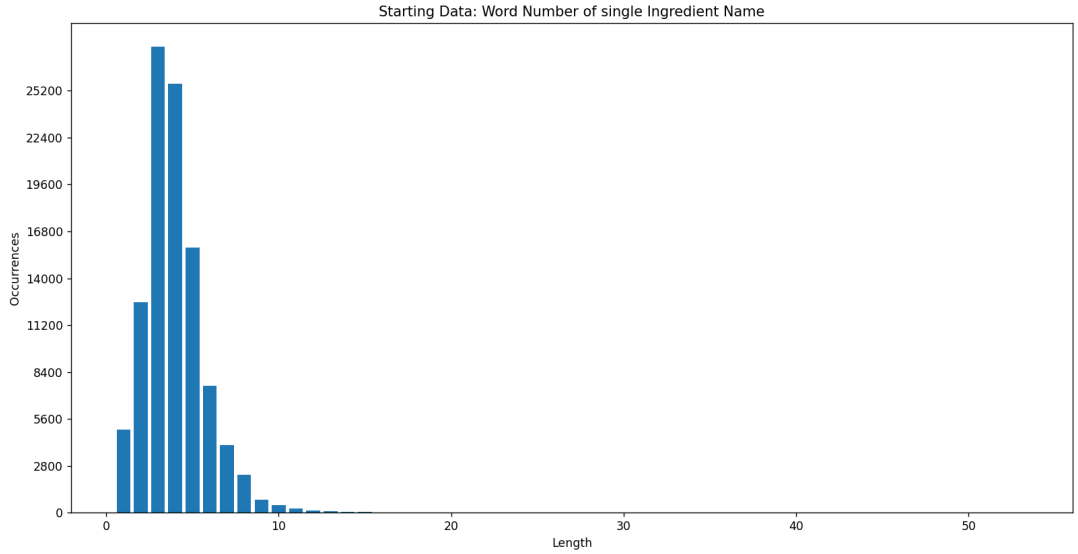


Figura 2.3: Distribuzione del numero di parole utilizzate per descrivere un singolo ingrediente nel rispettivo campo testuale. Dall'istogramma si nota che la maggior parte degli ingredienti è descritto da 3 o 4 parole.

L'analisi della lunghezza del campo testuale degli ingredienti ha rivelato che la maggior parte degli ingredienti (circa l' 84.7%) presenta una lunghezza compresa tra 1 e 5 parole, estremi compresi. Solo il 15.3% degli ingredienti presenta una lunghezza superiore a 5 parole. Di questi solo l'1.8% degli ingredienti contengono un numero elevato di parole uguale o superiore a 9 parole nel proprio campo testuale.

Considerando che a livello teorico un ingrediente dovrebbe essere definito da meno parole possibili al fine di ottimizzare le eventuali analisi testuali successive e soprattutto al fine di essere abbastanza generale da non distinguere come ingredienti distinti gli stessi ingredienti con una sola parola di differenza (un esempio di un ingrediente estremamente comune nel dataset: "Olio extravergine" e "Olio extravergine d'oliva"), questi risultati non soddisfano completamente i nostri standard.

È comunque necessario sottolineare che il numero di parole che descrivono un ingrediente e da ritenere non superflue, dipende sia dal livello di dettaglio che si vuole mantenere nella distinzione degli ingredienti (ad esempio "Tonno al naturale" e "Tonno sott'olio" riguardano entrambi lo stesso ingrediente "Tonno" ma sicuramente a livello di valori nutrizionali ci sono delle differenze dovute alla presenza o meno di olio nel metodo di conservazione) ma anche dalla presenza di parole descrittive di caratteristiche secondarie, che potrebbero riguardare maggiormente la preparazione invece che impattare a livello nutrizionale (alcuni esempi potrebbero essere "Peperoni rossi o gialli" a livello nutrizionale il colore è trascurabile, mentre magari nella preparazione potrebbe influenzare l'impiattamento o altro).

La media della lunghezza degli ingredienti è risultata essere di 3.97 parole, con una deviazione standard di 11.16. La moda risulta 3 mentre la mediana 4.

Questa distribuzione suggerisce che la maggior parte degli ingredienti sono descritti in modo conciso e semplice, mentre solo alcuni ingredienti richiedono una descrizione più dettagliata.

Questa prima analisi statistica riguardante gli ingredienti suggerisce che il dataset presenta una struttura abbastanza uniforme, ma non sufficiente da soddisfare gli standard necessari previsti per utilizzarlo come fonte di informazioni nutrizionali all'interno del progetto.

2.1.2 Analisi composizione

La comprensione approfondita della composizione del dataset iniziale si è rivelata una fase cruciale nel processo di analisi, poiché ha consentito di identificare i principali ingredienti che avrebbero ricoperto la maggior parte delle ricette.

Questo approccio ha quindi permesso l'impiego di una analisi manuale dei dati testuali (fatta ad esempio per la pulizia del campo 'name', per l'estrazione del nome degli ingredienti veri e propri o ancora per la gestione delle eccezioni del campo 'quantità' e 'unit' come 'qb' e 'gen', vedi sezione 3.2.1), andando a volgere l'attenzione su quegli ingredienti del dataset più impattanti a livello percentuale.

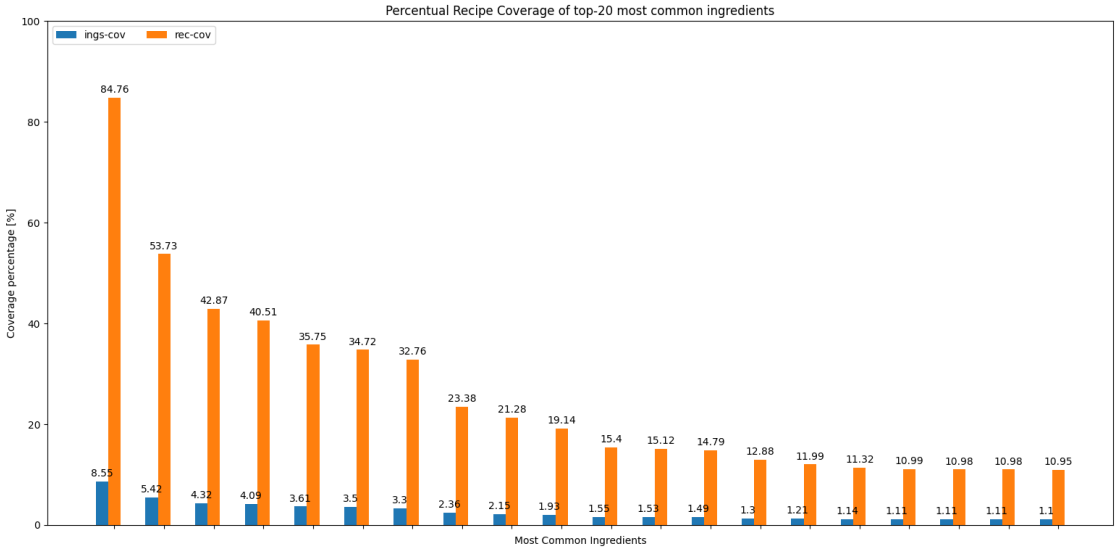


Figura 2.4: Copertura degli ingredienti rispetto alle ricette (arancione) e Copertura degli ingredienti rispetto al totale degli ingredienti(blu).

In figura 2.4, è riportato un grafico a barre che indica con quale percentuale un ingrediente appare nelle ricette, e nel totale degli ingredienti. In altre parole in quale percentuale delle ricette è presente l'ingrediente, e quale percentuale ricopre rispetto ad un insieme di tutti gli ingredienti estratti dalle ricette, considerando anche i duplicati. Ad esempio si può considerare un insieme di ingredienti come [olio: 78, sale: 109, ...] e considerare la percentuale rispetto al totale.

Queste informazioni sono state riportate solo per i primi 20 ingredienti per copertura delle ricette.

La lista dei 20 ingredienti più comuni nel dataset è la seguente, ordinata così come nel grafico in figura 2.4: sale, olio extravergine d'oliva, farina, pepe, uovo, zucchero, burro, latte, spicchio di aglio, pomodoro, lievito, prezzemolo, panna, parmigiano reggiano, olio, cipolle, tuorlo d'uovo, cioccolato, costa di sedano, zucchero velo.

2.2 Schema dei Dati

I seguenti schemi descrivono accuratamente le strutture in cui sono organizzati i dati grezzi del dataset iniziale nelle diverse fase. È mostrato lo schema dei dati dopo aver attraversato le diverse fasi della pipeline di Preprocessing.

Infatti l'obiettivo principale della pipeline di preprocessing è quello di trasformare i dati di partenza in dati che rispecchiano una organizzazione equivalente allo schema

riportato in figura 2.5 che descrive ad alto livello la struttura desiderata e le relazioni esistenti tra i diversi elementi.

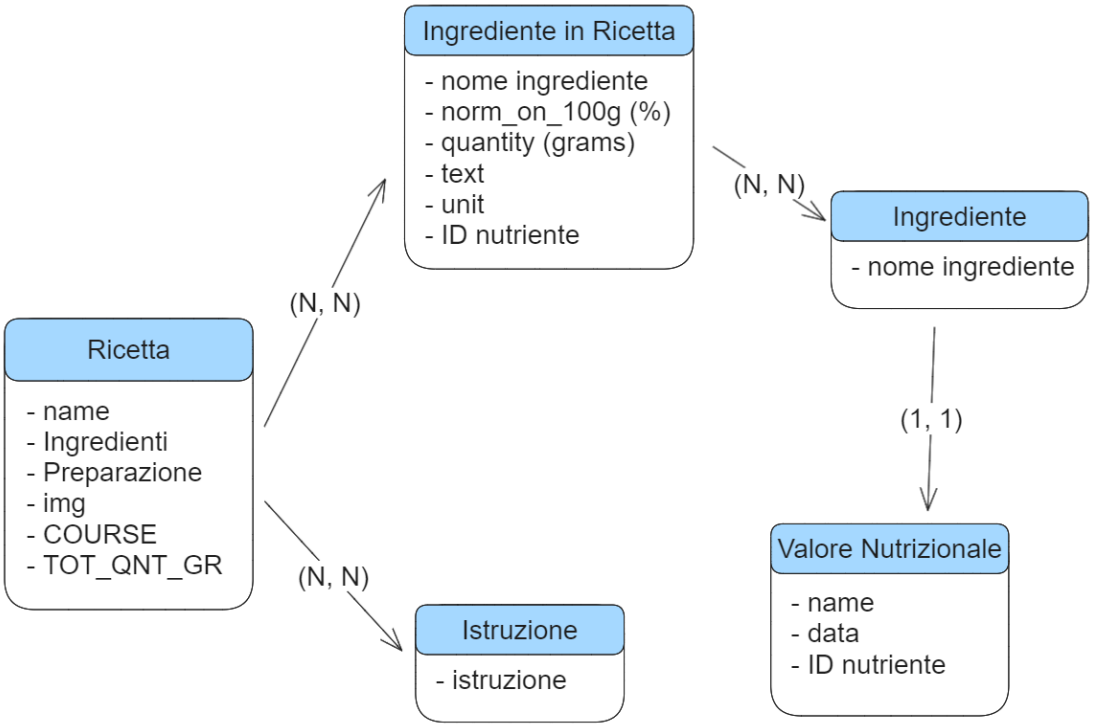


Figura 2.5: Schema UML desiderato come output della pipeline di Preprocessing, ad alto livello

Come si può notare le entità principali sono Ricette, Ingredienti, Istruzioni e Nutrienti.

L'entità "Ricetta" rappresenta le diverse ricette disponibili nel dataset, ciascuna caratterizzata dai propri attributi come il nome della ricetta, la lista di ingredienti, quella di istruzioni per la preparazione, un'immagine rappresentativa (tramite url), il tipo di portata (ad esempio antipasto, primo, secondo o dolce), e la quantità totale in grammi della ricetta stessa. Ogni ricetta può includere molti ingredienti all'interno della lista. Una ricetta può avere una lista di ingredienti ma ogni ingrediente non è vincolato ad una singola ricetta. La stessa relazione N a N esiste anche verso l'entità "Istruzione".

L'entità "Ingrediente in Ricetta" indica la struttura completa che rappresenta l'ingrediente con i dati, relativamente a quella ricetta. Esso è contenuto nella lista di ingredienti della ricetta. Contiene tutte le caratteristiche di ciascun ingrediente, inclusi il nome "name", la percentuale rispetto alla quantità totale della ricetta ovvero "norm_on_100g", calcolata quindi su 100 grammi, la quantità in grammi

indicata per la ricetta, una descrizione testuale di partenza che contiene le informazioni originali da cui sono state estratte le altre ("text"), e l'unità di misura specificata per valutare la quantità.

L'entità "Ingrediente" invece riporta esclusivamente il suo nome. Esso può essere presente in più liste di ingredienti, all'interno delle ricette e con dati diversi. Inoltre questo nome permette il collegamento univoco con i relativi dati nutrizionali contenuti in "Valore Nutrizionale".

L'entità "Valore Nutrizionale" comprende informazioni dettagliate sui nutrienti che caratterizzano l'ingrediente, come il nome dell'ingrediente, i dati specifici relativi ad esso e un identificativo. Ogni Valore Nutrizionale è associato esclusivamente al nome dell'ingrediente e non alla struttura.

Infine, l'entità "Istruzione" contiene la singola istruzione per la preparazione delle ricette, con ogni istruzione associata direttamente a una specifica ricetta.

A livello implementativo invece, lo schema riportato in figura 2.6, descrive più accuratamente anche le strutture e le chiavi che esistono propriamente nell'implementazione nel codice.

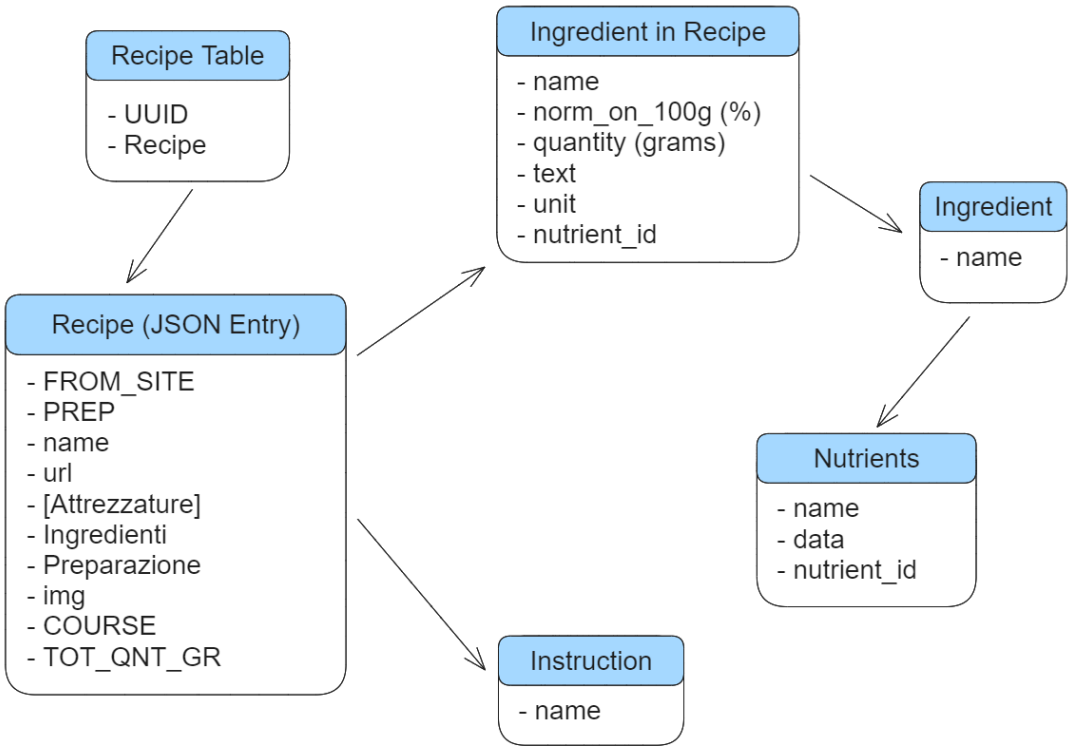


Figura 2.6: Schema UML implementativo

Come si nota dallo schema implementativo in figura 2.6 esiste una struttura dati più generale utilizzata per organizzare le ricette chiamata "Recipe Table". Questa struttura è una tabella (gestita in fase implementativa come un dizionario) che associa un ID univoco ad ogni ricetta. Infatti come chiavi presenta gli ID della ricetta, mentre il valore collegato ad un ID è rappresentato dalla ricetta stessa. La generazione dell'ID univoco è ottenuta utilizzando la libreria UUID.

Ogni ricetta è a sua volta una tabella con diverse chiavi. Queste chiavi organizzano tutti i dati della ricetta e permettono di accedervi individualmente. Tra queste chiavi quelle di maggiore importanza sono "name" che indica il nome estratto dell'ingrediente (ad esempio "farina", "olio d'oliva" etc.) e la chiave "Ingredienti" che permette invece di ottenere la lista di tutti gli ingredienti che compongono la ricetta. Le altre chiavi invece riportano:

- "**FROM_SITE**": indica da quale sito proviene la ricetta. Serve a mantenere le informazioni legate alla creazione del dataset tramite strumenti di Web Scraping effettuato per la raccolta delle ricette.
- "**PREP**": indica se la ricetta prevede una preparazione o meno.
- "**url**": riporta l'url con cui poter accedere alla pagina della ricetta.
- "[**Attrezzature**]": è riportato tra parentesi quadre per indicare il carattere opzionale di questa chiave, dato che non tutte le ricette la possiedono. Se presente, riporta una lista di attrezzature necessarie per la realizzazione della ricetta, sottoforma di lista di stringhe.
- "**Preparazione**": riporta una lista di istruzioni che descrivono la procedura per realizzare correttamente la ricetta.
- "**img**": riporta l'url dell'immagine associata alla ricetta.
- "**COURSE**": indica la portata associata alla ricetta, ad esempio "primo", "antipasto" etc.
- "**TOT_QNT_GR**": riporta la quantità totale della ricetta misurata a partire dagli ingredienti e dalla loro analisi.

Successivamente, ogni ingrediente appartenente alla lista di ingredienti, possiede poi i propri dati organizzati nelle seguenti chiavi:

- "**name**": riporta una stringa contenente il nome dell'ingrediente estratto e pulito da parole superflue.

- "**norm_on_100g**": indica la percentuale che tale ingrediente ricopre rispetto al peso totale calcolato della ricetta
- "**quantity**": riporta il peso dell'ingrediente così come estratto dal campo testuale originale dell'ingrediente, sottoforma di float.
- "**text**": è il campo testuale originale dell'ingrediente, così come è stato salvato nel processo di creazione del dataset tramite Web Scraping.
- "**unit**": riporta l'unità di misura con cui interpretare la quantità nella chiave "quantity". È anche utilizzata per etichettare alcuni ingredienti come quantità particolari (vedi sotto sezione 3.2.1).
- "**nutrient_id**": riporta l'ID utilizzato per collegare l'ingrediente alla entrata corrispettiva all'interno del database dei nutrienti scelto, e permettere così una successiva estrazione delle informazioni nutrizionali.

Per quanto riguarda le strutture "Instruction" esse riportano esclusivamente le istruzioni, ordinate in una lista al'interno della struttura "Recipe". Ogni singola istruzione della lista è rappresentata da una stringa.

"Ingredient" invece non è propriamente una struttura a se stante ma è riportata così per far capire la distizione con "Ingredient in Recipe". Infatti il campo "name" di "Ingredient" riporta solo il nome relativo all'ingrediente stesso, che è indipendente dalla struttura riportata internamente nella ricetta, che ne riporta il nome come valore del campo "name".

"Nutrients" invece riporta le informazioni nutrizionali dell'ingrediente, ed è collegato anche direttamente con ogni "Ingredient in Recipe", tramite il rispettivo identificativo.

2.3 Data Cleaning

La prima fase di un processo di creazione di una pipeline di data cleaning è sicuramente quella di rimozione dei duplicati.

Per molti aspetti, nel caso in cui i nomi delle ricette coincidano esattamente l'operazione può risultare banale, ma quando si tratta di ricette prese da diversi siti, identificare due ricette uguali ma con nomi diversi può risultare più complesso.

In questo ultimo caso ci troviamo di fronte a una categoria nota come 'pseudo-duplicati'. Questi si manifestano quando due ricette sono sostanzialmente identiche o estremamente simili, ma si presentano con nomi diversi, sia a livello di Titolo che a livello di ingredienti. Riconoscere e trattare tali pseudo-duplicati richiede una strategia più sofisticata, in quanto la semplice comparazione dei nomi non risulta più sufficiente.

2.3.1 Rimozione Duplicati Con stesso titolo

Durante la fase di preparazione dei dati, una delle operazioni fondamentali è stata la rimozione dei duplicati dal dataset di ricette. Questo passaggio è cruciale per garantire la qualità dei dati, evitando duplicati che potrebbero distorcere tutte le analisi e le applicazioni successive.

La rimozione dei duplicati è stata eseguita andando a condurre prima un'analisi del dataset per identificare eventuali righe duplicate. Questo è stato fatto confrontando le caratteristiche chiave delle ricette, come il titolo e gli ingredienti. Eseguendo una lettura totale del dataset e utilizzando una struttura dati per raccogliere tutte le ricette, esse sono state inserite in una tabella con chiave il loro titolo, e come valore il numero di occorrenze di quel titolo. Successivamente sono state rimosse tutte quelle ricette che avevano come titolo una delle chiavi con numero di occorrenze superiore a 1. Durante queste analisi è stato identificato circa il 9.5% dei dati iniziali come duplicati, che su approssimativamente 10800 ricette, rappresentano poco più di 1000 duplicati.

Una volta identificati i duplicati, è stata presa la decisione su quale riga mantenere e quale eliminare. Questa si stata basata principalmente sul mantenere la ricetta più ricca di informazioni, e corrette cercando di evitare errori di battitura negli ingredienti o altro. Essendo circa 1000 duplicati e il numero di ridondanze medie di poco superiore a 2 questo processo di selezione è stato affrontato manualmente.

Infine, dopo aver completato il processo di rimozione dei duplicati, è stata condotta una revisione finale per verificare l'efficacia dell'operazione. Sono stati eseguiti controlli incrociati per assicurarsi che tutte le righe duplicate fossero state correttamente eliminate e che il dataset risultante fosse privo di duplicati indesiderati.

Capitolo 3

Pipeline di Preprocessing: Struttura, Funzione e Risultati

In questo capitolo vengono descritti dettagliatamente i vari passaggi, gli algoritmi e le implementazioni utilizzati per trasformare i dati grezzi di partenza, in dati di maggiore qualità.

Lo scopo di questa prima fase è quello di ottenere dati raffinati che siano pronti per essere utilizzati sia per addestramento di nuove reti (mantenendo le immagini delle ricette), sia come database di ricette utile per ricavare i valori nutrizionali delle singole ricette da utilizzare anche direttamente nell'applicativo.

È necessario sottolineare che la struttura della pipeline di preprocessing complessiva è generica, pertanto utilizzabile anche su altri dati, inoltre è fondamentale precisare che le strutture dati utilizzate per l'elaborazione testuale (quali per esempio tabelle di pulizia ed estrazione degli ingredienti, vedi sezione 3.4 e 3.3, o nella tabella per la gestione delle quantità particolari 3.2.1) sono state create appositamente per la lingua italiana e su misura per i dati qui utilizzati. Infatti queste strutture sono costituite da stringhe di testo che riportano parole italiane.

Dopo aver rimosso i duplicati ed avere effettuato le analisi statistiche preliminari, il passo successivo è stato quello di manipolare i dati andandoli a prepararli per essere usati nelle fasi successive dell'applicativo.

In questa fase verranno quindi riorganizzati i dati, andando a raffinarli il più possibile, eliminando il rumore testuale presente in generale all'interno dei campi ingrediente (considerando quindi tutti i campi, i nomi e i valori), nonché più in

particolare nei campi testuali degli ingrediente stessi, quindi nel campo inizialmente chiamato "name".

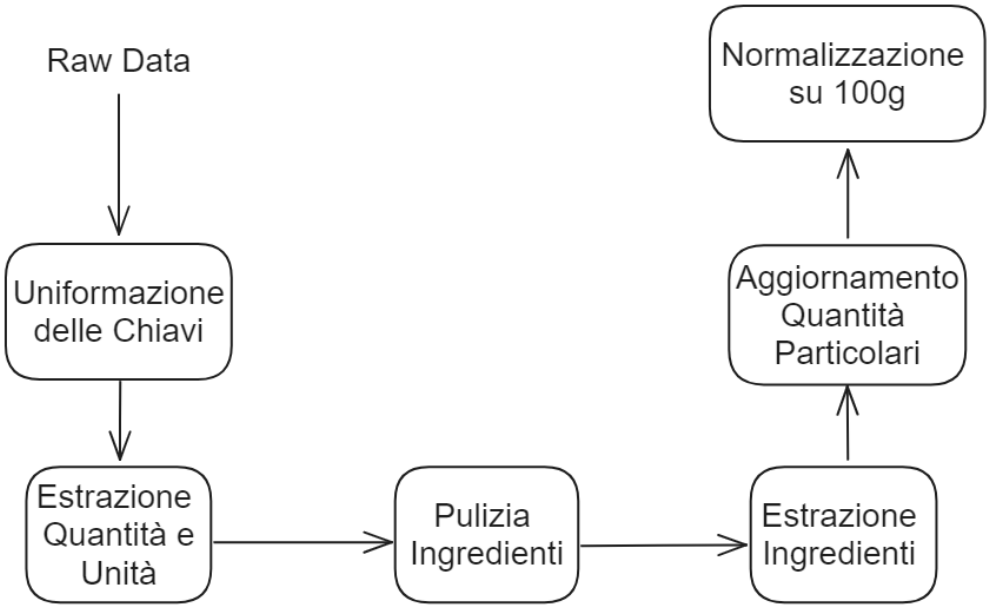


Figura 3.1: Schema Pipeline di Preprocessing

3.1 Uniformazione delle Chiavi

Il primo passo è stato quello di uniformare i nomi delle chiavi di ogni ricetta e anche la sua struttura.

Nei dati di partenza, sono presenti diverse chiavi per indicare in realtà lo stesso campo, pertanto dopo aver scelto una definizione standard delle varie chiavi (sia della tabella ricette sia di quelle più interne come la tabella degli ingredienti, vedi 2.5), tutte le ricette che si discostavano da quest'ultima sono state modificate di conseguenza.

Al termine di questa fase iniziale, tutte le ricette avranno le stesse chiavi per indicare gli stessi campi e saranno quindi pronte per passare alla fase successiva.

Un esempio eclatante è il campo testuale che descrive il singolo ingrediente, il campo 'name' della rispettiva tabella Ingrediente. Queste tabelle potevano riportare fino a 3 chiavi distinte per accedere alla medesima informazione riguardo l'ingrediente: 'name', 'ingr' e 'Ingr'. Tutte queste chiavi sono riportate all'unica definita dallo schema uml: 'name'.

Come schema di riferimento rimane lo schema implementativo in figura 2.6.

3.2 Estrazione delle Quantità e Unità di misura

Il passo successivo è quello di andare ad analizzare il campo 'name' di ogni ingrediente per estrarre le informazioni necessarie per stabilire la quantità utilizzata nella ricetta. A questo scopo è necessario estrarre un numero che indichi la quantità e un simbolo che indichi l'unità di misura relativa. Per fare ciò viene sfruttata la libreria standard di python delle espressioni regolari 're'.

Sono stati costruiti alcuni pattern specifici per estrarre numeri opportuni (ad esempio, in 'Farina 00 - 50g' si evita di registrare la coppia di zeri), ma anche frazioni o range. In particolare le frazioni come 1/3 o 'mezza' sono riportate al loro valore, mentre per i range viene preso come quantità la media (i.e. '4-5' diventa 4.5).

La stessa cosa viene fatta per le unità di misura, interpretate come i caratteri successivi alla quantità numerica. Per evitare errori di battitura o altro, si calcola la Levensthein Distance (LD) tra i caratteri estratti per l'unità di misura e delle stringhe raccolte in una lista di unità di misura standard.

La stringa più vicina secondo questa metrica, viene utilizzata come unità di misura per quell'ingrediente.

Le unità utilizzate sono le seguenti: g, gr, ml, l, lt, kg.

Tutte le unità di misura sono poi riportate a gr, l, kg e ml. Successivamente viene effettuata una prima conversione: kg in g, l in ml e poi una seconda conversione dai ml ai g utilizzando una densità media stimando dei liquidi più comuni usati in cucina molto vicina a quella dell'acqua: $1.005 g/ml$. Il motivo di questa scelta è stato la volontà di considerare anche altri liquidi come latte o miele, liquidi che riportano una densità maggiore dell'acqua ma sono usati meno. Inoltre anche l'olio è molto presente ed avendo una densità inferiore a quella dell'acqua tende a controbilanciare quella più alta ma meno comune dei liquidi più densi sopracitati.

3.2.1 Casi particolari

Nel caso in cui non venga trovata nessuna corrispondenza tra i caratteri che seguono la quantità e le unità di misura standard della lista, l'unità di misura assegnata agli ingredienti è una tra le seguenti etichette: 'qb', 'gen' o 'err'.

La prima etichetta 'qb' indica che la quantità non è specificata ma nel testo dell'ingrediente è presente l'indicazione 'q.b.', 'qb' o 'q.b'. Questa sigla è utilizzata per indicare che l'ingrediente è da impiegare in quantità arbitrarie a discrezione di chi preparerà la ricetta. Nel nostro caso risulta problematico da analizzare essendo non più un numero ma una variabile da stimare in virtù dell'ingrediente specifico.

L'etichetta 'gen' indica invece che una quantità è stata rilevata, ma non una unità di misura relativa. Questo avviene principalmente nei casi in cui si indicano un numero di elementi dello stesso ingrediente (ad esempio '4 zucchine', '3 uova'). In questo caso si ha un numero ma non una quantità effettiva e anche qui si torna a lavorare con delle variabili da stimare, seppur con più informazioni, dipendenti strettamente dall'ingrediente specifico.

L'ultima etichetta è invece 'err' che indica la completa assenza di numeri e unità di misura. La si trova in un numero ristretto di ingredienti che riportano solamente il nome dello stesso. In questo caso il problema si riduce ad una pura speculazione sulla quantità e necessita di ricerche specifiche per ottenere una stima adeguata della quantità.

Queste 3 etichette sono gestite diversamente e in un secondo momento. Una volta terminata questa seconda fase, ogni ingrediente ha ora delle chiavi 'quantity' e 'unit', dove si registra la quantità dell'ingrediente i-esimo in una ricetta, e l'unità di misura ad esso associata.

3.3 Pulizia Ingrediente

Una volta estratta la quantità e l'unità di misura, bisogna inziare a prepare i dati per estrarre il nome vero e proprio dell'ingrediente. Questo passaggio è necessario poiché la maggior parte degli ingredienti presenta inizialmente un campo di testo chiamato 'name' in cui una breve stringa descrive l'ingrediente, la quantità ma a volte anche la motivazione della scelta, dettagli del suo utilizzo o diverse alternative.

Alcuni esempi sono "Burro per ungere la teglia", "scorza di limone o arancia" ma altri ancora riportano addirittura nomi di utensili, tratti di istruzioni per la preparazione o alternative esageratamente prolisse.

Dopo aver quindi riscontrato queste molteplici impurità nel dataset, si procede quindi con questa fase importante, in particolare per alleggerire e semplificare l'analisi successiva dell'estrazione dell'ingrediente.

Il primo obiettivo è quello di rimuovere tutti quei falsi 'ingredienti' che, sorprendentemente, contenevano in realtà attrezzi da cucina ('padella', 'frullatore', 'coltelli'), o titoli per categorie di ricette ('Idee per Frittate', 'Ricette invernali','sfiziosità').

Il secondo è invece quello di andare ad eliminare tutte quelle parole superflue utilizzate per andare a definire l'ingrediente, quindi molti aggettivi, preposizioni

etc sono rimossi. Alcuni esempi come 'per', 'fredde', 'denocciolati' sono parole considerate superflue, mentre altre parole che possono specificare dettagli influenti sul futuro calcolo dei valori nutrizionali dell'ingrediente sono mantenute, come ad esempio 'sott'olio', 'sotto aceto' o 'sotto sale' etc.

Per eliminare i falsi ingredienti e le parole superflue, per prima cosa è stata costruita manualmente una tabella di parole inutili per andare a definire l'ingrediente.

Successivamente sfruttando questa tabella, costruita analizzando attentamente i dati, le parole individuate sono rimosse dal campo 'name' dell'ingrediente stesso (ie: fredde, verdi, denocciolate etc.).

Occorre sottolineare che il campo 'name' ora contiene la stringa originale depurata dalle parole superflue, ma la stessa stringa originale è stata precedentemente salvata in un nuovo campo chiamato 'text', potenzialmente utile per eventuali analisi successive.

Inoltre alcune parole e simboli attivano l'annullamento completo del campo 'name' (che viene sostituito con una stringa vuota) poiché indicano che quell'ingrediente non è tale ma è o un attrezzo da cucina o una categoria di ricette.

Questo ridotto pool di parole è stato costruito seguendo l'esperienza maturata dall'analisi personale di questi dati, avendo colto alcuni pattern ricorrenti (ad esempio la presenza di una ' / ' indica che l'ingrediente è una lista di attrezzi separati da /, o che la presenza della parola stampo/stampi indica l'attrezzo come in 'stampo da muffin').

Nell'eventualità che un ingrediente contenga solo parole inutili, queste sono rimosse e l'ingrediente risulterà una stringa vuota, che sarà eliminata nelle fasi successive.

3.4 Estrazione Ingrediente

Una volta puliti adeguatamente tutti gli ingredienti, il campo 'name' risulta come una stringa più ridotta, contenente poche parole significative per l'identificazione dell'ingrediente vero e proprio.

L'approccio qui utilizzato è stato quelli di creare manualmente una tabella di ingredienti, chiamata tabella di riferimento. Questa tabella è stata costruita maualmente partendo dagli ingredienti presenti all'interno del dataset, dopo averne analizzato gli ingredienti. Per ogni ingrediente di ogni ricetta, viene eseguita successivamente una funzione di confronto che utilizza la LD per calcolare il match migliore tra la stringa di input, ovvero il campo 'name' dell'ingrediente in esame, e una delle entrate della tabella. Il nome dell'ingrediente sarà quindi aggiornato

con l'entrata più simile, in caso di match, altrimenti verrà mantenuto il nome precedente.

Lo scopo di questa fase è quello di uniformare gli ingredienti utilizzando delle denominazioni comuni, col fine duplice di aggregare i nomi degli ingredienti e di risolvere contemporaneamente gli errori di battitura più comuni, ed eventualmente risolvendo anche alcuni elementi grezzi rimanenti dalla fase di pulizia. Così facendo viene quindi ridotto il numero di ingredienti diversi da diverse migliaia a poco più di 800.

Bisogna evidenziare che per considerare una corrispondenza come valida, si utilizza una soglia minima per la distanza di Levensthein; se non viene trovata una corrispondenza con similarità superiore o uguale alla soglia il match sarà considerato nullo e il campo testuale dell'ingrediente sarà mantenuto immutato.

Un secondo punto da evidenziare è che la qualità di questa fase dipende molto sia dalla scelta della soglia di similarità minima, sia anche dalla popolazione di ingredienti contenuti nella tabella di riferimento. La prima dipendenza sarà analizzata con alcuni dati successivamente.

Per quanto riguarda la dipendenza dalla tabella di riferimento utilizzata, non verrà approfondita in questa analisi. Ciò è dovuto al fatto che la tabella di riferimento può essere liberamente modificata, aggiungendo o rimuovendo voci, per adattarsi in modo più preciso e dettagliato al dataset in esame. Pertanto, poiché la tabella di riferimento non è fissa ma può essere personalizzata in base alle esigenze specifiche, non è considerato rilevante analizzarne il ruolo in questo contesto.

3.4.1 Analisi Dipendenza dalla Soglia di Similarità

Come anticipato, l'algoritmo dipende dalla scelta di una soglia di similarità. Di seguito verranno analizzati alcuni risultati ottenuti scegliendo una soglia sbilanciata, o troppo bassa o troppo alta.

In particolare più la soglia è bassa, maggiore sarà la flessibilità dell'algoritmo ad associare ad un ingrediente una entrata della tabella di riferimento. La tabella 3.1 riporta alcuni esempi di associazioni errate tra i campi testuali degli ingredienti (*target_string*), successivamente alla fase di pulizia, e le entrate della tabella di riferimento (*match*) con relativo punteggio di similarità calcolato secondo la distanza di Levensthein (*score*).

target_string	**match**	**score**
farfalle	sale	68
feta greca	frutta fresca	70
brandy	rane	68
gassata	pasta	67
ammoniaca	soia	68
pioppini	pipe	68
pleurotus	aperol	60
brandy	rane	68
baguette	ragu	68
mezze maniche	mais	68
caserecce	pere	68

Tabella 3.1: Estrazione Ingredienti con Soglia Similarità bassa, pari 55

Nella tabella 3.1, sono raccolte alcune associazioni fatte dall'algoritmo di estrazione degli ingredienti. La soglia utilizzata in questa analisi è pari a 55. Le associazioni qui riportate sono considerate errate, pocihè gli ingredienti non corrispondono. Inoltre esse sono scelte appositamente per evidenziare quali errori sono possibili in caso di utilizzo di una soglia troppo bassa per questa fase della pipeline di preprocessing.

Per ottenere i risultati da cui sono estratti i dati riportati in tabella 3.1 è stata utilizzata una soglia di similarità bassa, pari a 55. Come è possibile notare dalla tabella stessa già a valori prossimi a 65 circa le associazioni iniziano a risultare sbagliate.

D'altro canto se la soglia è troppo alta, l'algoritmo sarà troppo rigido nella ricerca di corrispondenze e di conseguenza si andranno a perdere alcune associazioni che per l'umano potrebbero essere scontate. A questo proposito nella tabella 3.2 sono riportate alcune associazioni scartate da un utilizzo di una soglia di similarità troppo alta.

Nella tabella 3.2 sono raccolte alcune associazioni fatte dall'algoritmo di estrazione degli ingredienti. La soglia utilizzata in questa analisi è pari a 85. Le associazioni qui riportate sono associazioni non considerate, poiché la similarità, definita dal valore *score* è inferiore alla soglia stabilita. Il motivo per cui queste associazioni sono state riportate è per evidenziare quali associazioni, da considerarsi corrette, sarebbero scartate da un algoritmo di estrazione degli ingredienti che utilizza una soglia troppo alta.

target_string	match	score
orate	orata	80
cavolfiore viola	cavolfiori	81
broccoletti	broccoli	84
broccoletti	broccoli	84
bietoline campestri	bietole	77
H] po' d'erba cipolina	erba cipollina	79
sarago	saraghi	77
cipolline borrettane	cipolla	77
salsicce verzini	salsiccie	80
polpi	polpo	80
scologni	scalogno	75
alcune ciliegine candite	ciliegie	79

Tabella 3.2: Estrazione Ingredienti con Soglia Similarità alta, pari 85

Per ottenere i risultati da cui sono estratti i dati riportati in tabella 3.2 è stata utilizzata una soglia di similarità alta, pari a 85. Come è possibile notare dalla tabella stessa attorno a valori di similarità prossimi a 75 alcuni plurali verrebbero scartati. Ancora la presenza di alcune forme alterate o la presenza di alcuni aggettivi potrebbe portare l'algoritmo a non aggregare correttamente gli ingredienti sotto lo stesso nome.

3.4.2 Scelta Empirica della Soglia di Similarità

Dopo un'attenta analisi, e considerando anche i dati riportati qui sinteticamente nelle tabelle 3.1 e 3.2, la soglia utilizzata per le analisi successive è stata definita pari a 77. Questo valore sembra essere empiricamente il valore che ottimizza entrambe le problematiche relative ad una adozione di una soglia di similarità o troppo bassa o troppo alta. In particolare tende a minimizzare il numero di associazioni errate che deriverebbero da una soglia troppo bassa, e al contempo minimizzare il numero di associazioni corrette ma scartate da una soglia troppo alta.

L'analisi è state effettuata seguendo passo passo l'esecuzione di questa fase, e salvando su un file un output dell'algoritmo simile a quello riportato nelle tabelle 3.2 e 3.1. L'esperimento è stato eseguito diverse volte per cogliere la soglia migliore.

3.4.3 Distanza di Levenshtein come misura di Similarità

Per individuare la corrispondenza migliore tra il campo testuale di un ingrediente e una entrate della tabella di riferimento, e costruire così una associazione più corretta possibile è stata scelta come metrica la distanza di Levenshtein.

La Levenshtein distance (LD) è una metrica molto diffusa nell'analisi testuale, che indica la differenza tra due stringhe. Un tipico esempio di utilizzo è l'individuazione automatica di errori di battitura, una funzionalità molto diffusa nei principali software di editing di testo. In particolare questa funzione misura il numero minimo di operazioni da effettuare che sono necessarie per trasformare la prima stringa nella seconda. Tra le operazioni considerate troviamo inserimenti, cancellazioni o sostituzioni di singoli caratteri.

Questa funzione rappresenta in sostanza la quantità di editing necessaria per far si che le due stringhe coincidano.

3.5 Aggiornamento quantità particolari

Alcuni ingredienti non hanno grammi come unità di misura ma sono identificati da delle etichette utili per gestire in maniera diversa il calcolo della loro quantità, così come descritto nella sottosezione 3.2.1.

L'approccio utilizzato per la gestione di questi casi è quello di andare a dividere questi ingredienti in due macro gruppi: quelli trascurabili ma presenti (quindi segnare che c'è una quantità minima di pochi grammi ad esempio) e ingredienti non trascurabili, aventi un peso decisamente impattante sulla ricetta stessa (es: 5 zucchine, 2 cipolle, 3 cucchiai di olio...). In questo secondo caso viene utilizzata una quantità media standard per gli ingredienti più comuni e poi moltiplicata per la quantità registrata. La quantità di quest' ultimo gruppo sarà quindi nuovamente aggiornata come:

$$quantity = old\_quantity * peso\_medio\_singolo$$

Per quanto riguarda l'unità di misura associata:

$$unit = "gr"$$

Per implementare questa soluzione, è stata costruita una tabella specifica per un ristretto sotto insieme degli ingredienti più comuni che appaiono con le etichette 'qb', 'gen' ed 'err'. La tabella associa ad ogni nome dell'ingrediente due chiavi, 'qb' e 'gen'. In particolare la prima chiave viene utilizzata sia per i 'qb' sia per gli 'err'. Il motivo di utilizzo anche per le etichette 'err' è legato al fatto che anche solo mantenere la presenza di quell'ingrediente potrebbe essere utile anche senza una buona stima della quantità, in particolare si pensi ad allergie o intolleranze.

Questa tabella associa quindi una quantità stimata ad ogni ingrediente, differenziando i casi di 'qb' e 'gen'.

È necessario specificare che la tabella utilizzata in questa fase è costruita attenendosi agli ingredienti presenti nella tabella utilizzata tra le fasi precedenti per l'estrazione degli ingredienti dal testo, vedi sotto sezione 3.4.

3.5.1 Analisi di distribuzione delle Quantità e Quantità particolari

In questa sotto sezione vengono analizzate le distribuzioni dei dati ottenuti relativamente ai campi quantità. In particolare sono analizzate le distribuzioni dei valori di quantità e le distribuzioni delle quantità particolari come 'err', 'qb' o 'gen' sia prima che dopo la fase di Aggiornamento delle quantità particolari.

Lo scopo di queste analisi è quello di indagare ed eventualmente sottolineare gli effetti che questa fase di aggiornamento, seppur effettuata parzialmente sugli ingredienti.

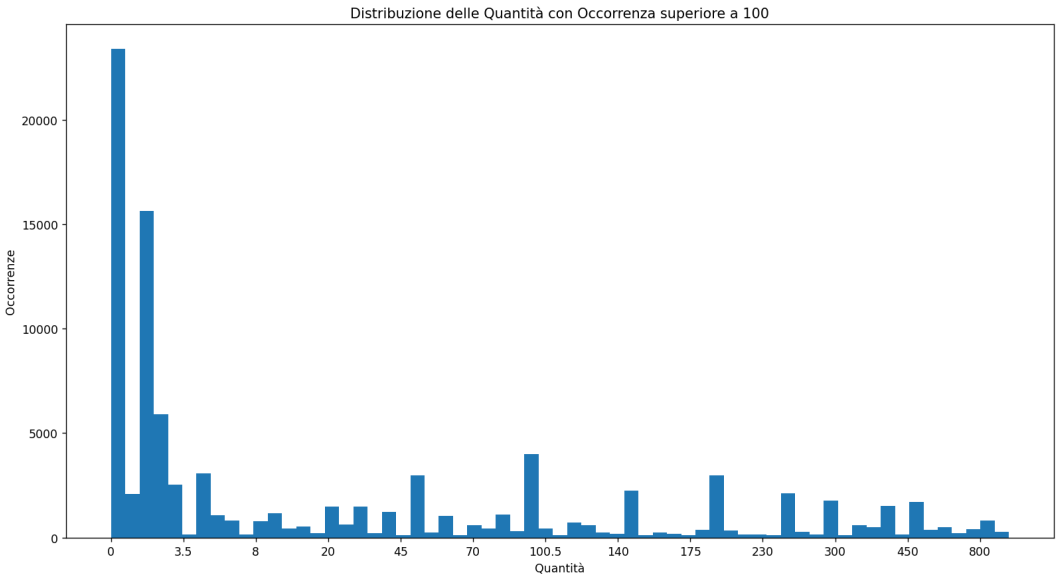


Figura 3.2: Distribuzione delle occorrenze delle quantità superiori a 100 occorrenze successivamente alla prima fase di estrazione delle quantità

Come si può notare la distribuzione delle quantità più comuni estratte dal dataset, ovvero quelle che appaiono più di 100 volte, è abbastanza irregolare. Presenta un picco iniziale su 0, dovuto alla presenza di tutti quegli ingredienti la cui unità di misura non è stata correttamente estratta (vedi sotto sezione 3.2.1).

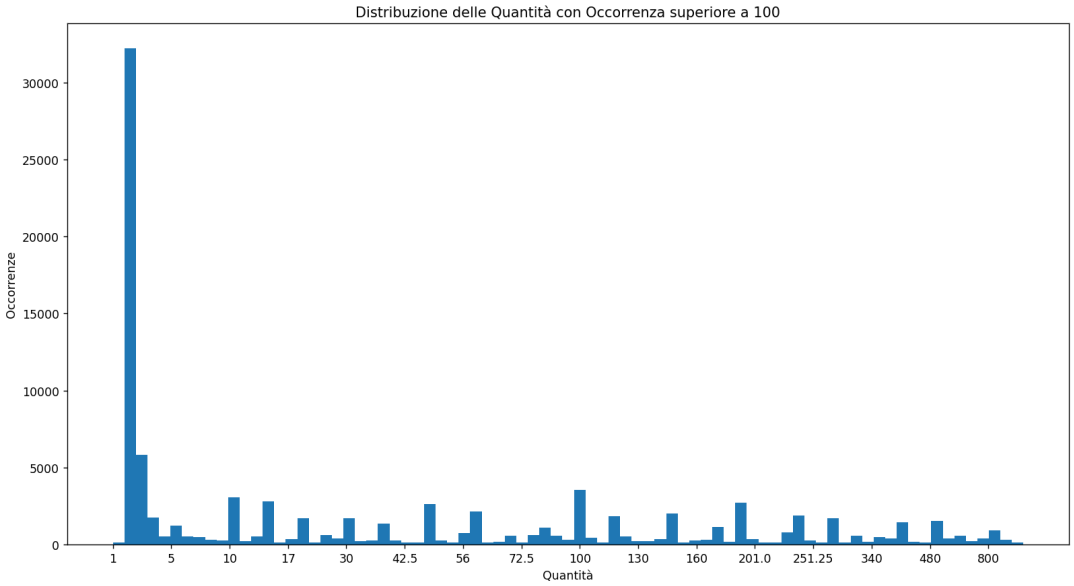


Figura 3.3: Distribuzione delle occorrenze delle quantità superiori a 100 occorrenze successivamente alla fase di aggiornamento delle quantità particolari

Nella seconda distribuzione possiamo notare come la fase di aggiornamento abbia spostato il picco da 0 a 2. Il motivo risiede nella implementazione di questa funzione che attribuisce alle quantità con unità 'qb' o 'err' una quantità predefinita.

Infatti queste etichette riportavano come quantità iniziale (prima della fase di aggiornamento) il valore 0.

Questa quantità predefinita dipende da ingrediente a ingrediente ed è assegnata tramite una tabella costruita manualmente. Questa tabella copre principalmente gli ingredienti più comuni presenti nelle categorie 'qb' e 'gen'. Nonostante sia un aggiornamento parziale delle quantità, essendo effettuato sugli ingredienti più diffusi i risultati sono evidenti soprattutto per quantità piccole legate alla presenza delle etichette 'qb' ed 'err'.

3.6 Seconda Analisi statistica degli Ingredienti

In questa sezione, viene presentata una analisi statistica dei dati dopo l'applicazione delle tecniche di preprocessing descritte nelle sezioni precedenti.

Una prima analisi effettuata dopo la fase di Pulizia dell'ingrediente, dove le parole superflue sono rimosse, e una analisi successiva effettuata invece dopo la fase di Estrazione dell' Ingrediente.

Lo scopo di questa analisi è quello di valutare l'efficacia delle tecniche di preprocessing utilizzate, nel migliorare la qualità e la consistenza dei dati. A tal fine, vengono confrontate le statistiche ottenute dall'analisi dei dati grezzi precedentemente effettuata (vedi sezione 2.1.1) con quelle ottenute dall'analisi dei dati preprocessati.

Questo confronto consente di determinare l'impatto del preprocessing sulla struttura e sulla qualità dei dati, e di identificare eventuali miglioramenti o cambiamenti nella loro distribuzione. In questo modo, è possibile verificare se le tecniche di preprocessing hanno raggiunto il loro obiettivo, ovvero quello di migliorare la qualità e l'utilizzabilità dei dati per applicazioni future.

In particolare le metriche analizzate di seguito sono il numero di ingredienti presente per ogni ricetta (per verificare quanti falsi ingredienti sono stati rimossi) e il numero di parole nel campo testuale dell'ingrediente che descrivono l'ingrediente stesso (per verificare le variazioni di rumore testuale nel campo testuale dell' ingrediente).

3.6.1 Distribuzione del numero di ingredienti per ricetta

Come precedentemente osservato, la distribuzione del numero di ingredienti per ricetta nel dataset di ricette culinarie presenta caratteristiche specifiche. In questa sezione, l'obiettivo è quello di verificare se e come le tecniche di preprocessing applicate al dataset hanno influenzato tale distribuzione.

In questa prima analisi è stato calcolato il numero di ingredienti per ricetta e la loro frequenza.

Analisi dopo pulizia ingrediente

Riprendendo le statistiche considerate nelle analisi sui dati grezzi iniziali, queste non riportano differenze sostanziali in nessun campo.

Le statistiche della distribuzione possono essere trovate nella tabella 3.3.

Analisi dopo Estrazione ingrediente

Dopo l'estrazione dell'ingrediente, il numero di ricette con un numero di ingredienti compreso tra 5 e 15 (estremi compresi) rappresenta l'85.7% del totale, mentre solo il 8.9% delle ricette presenta più di 15 ingredienti. Il 5.4% sono invece ricette con un numero di ingredienti inferiore a 5.

Le statistiche della distribuzione possono essere trovate nella tabella 3.3.

3.6.2 Numero di parole presenti campo testuale degli ingredienti

La seconda analisi, che riguarda invece il numero di parole presenti nel campo testuale dell'ingrediente, risulta decisamente più influenzata dalle trasformazioni e dalla pulizia effettuate tramite le fasi di preprocessing precedentemente descritte.

Analisi dopo pulizia ingrediente

L'analisi di questa metrica ha rivelato che la quasi totalità degli ingredienti (circa il 99.44%) presenta in questa fase una lunghezza testuale compresa tra 1 e 5 parole. Solo lo 0.53% degli ingredienti presenta una lunghezza superiore a 5 parole. Di questi solo lo 0.03% degli ingredienti contengono un numero elevato di parole uguale o superiore a 9 parole nel loro campo testuale. Andando ad analizzare più dettagliatamente l'intervallo preponderante di 1-5 parole, si riporta che il 92.5% degli ingredienti ha un numero di parole tra 1 e 3, e circa il 7% tra 4 e 5 parole, estremi compresi.

Le statistiche della distribuzione possono essere trovate nella tabella 3.4.

Analisi dopo Estrazione ingrediente

Questa analisi riporta invece che l'esatta totalità degli ingredienti presenta, successivamente a questa fase, una lunghezza testuale compresa tra 1 e 5 parole. il 99.9% rientra nell'intervallo 1-3, e nessun ingrediente viene descritto da più di 4 parole.

Le statistiche della distribuzione possono essere trovate nella tabella 3.4.

3.6.3 Performance ed Effetti della Pipeline: Confronto statistiche dati grezzi con dati preprocessed

Nelle figure successive sono riportati i grafici rispettivamente delle distribuzioni del numero di parole per campo testuale e del numero di ingredienti per ricetta. Nelle tabelle seguenti sono invece riassunte le principali statistiche estratte dalle distribuzioni sopracitate.

Questi risultati si sono rivelati conformi alle aspettative esistenti, in particolare riguardo questa ultima fase di estrazione dell'ingrediente, poiché il campo testuale dell'ingrediente viene rimpiazzato dall'entrata, ottenuta da una tabella specifica, che più risulta vicina al campo testuale pulito dalle parole superflue.

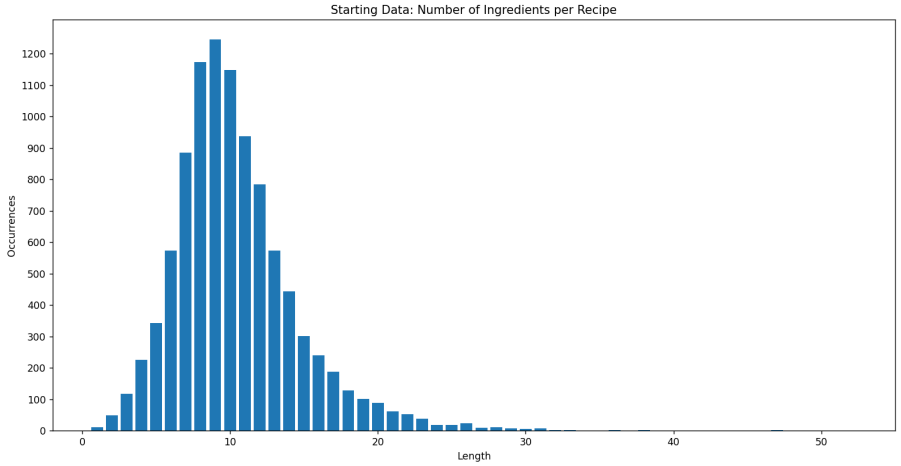
Pertanto la variazione della distribuzione dei dati ottenuti dopo questa fase, è estremamente dipendente dalla distribuzione dei dati utilizzati nella tabella. In altre parole le statistiche sul numero di parole che descrivono gli ingredienti sono estremamente simili ai campi testuali utilizzati nella tabella.

Dati Analizzati	**Media**	**Max**	**Stdev**	**Moda**	**Mediana**
Dati di partenza	10.43	52	13.18	9	10
Pulizia Ingrediente	10.43	52	13.18	9	10
Estrazione Ingrediente	9.92	49	12.88	9	9

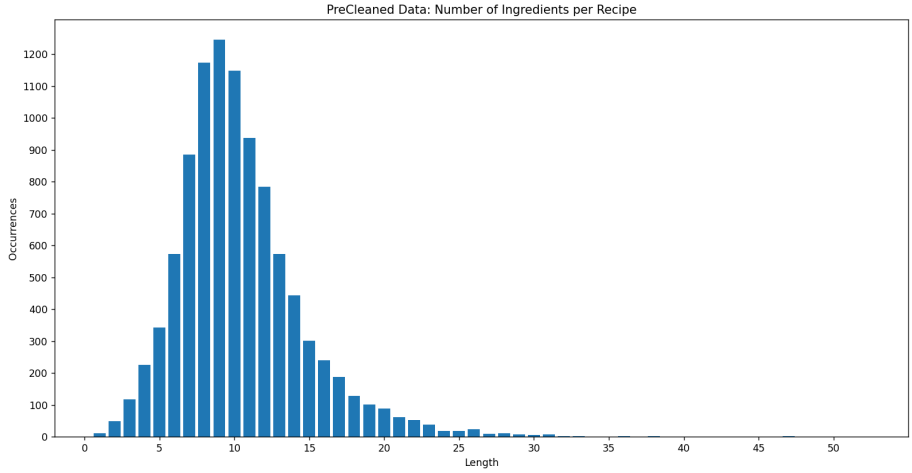
Tabella 3.3: Statistiche misurate relative al numero di ingredienti per ricetta

Dati Analizzati	**Media**	**Max**	**Stdev**	**Moda**	**Mediana**
Dati di partenza	3.97	53	11.16	3	4
Pulizia Ingrediente	1.82	13	3.89	1	1
Estrazione Ingrediente	1.33	7	2.3	1	1

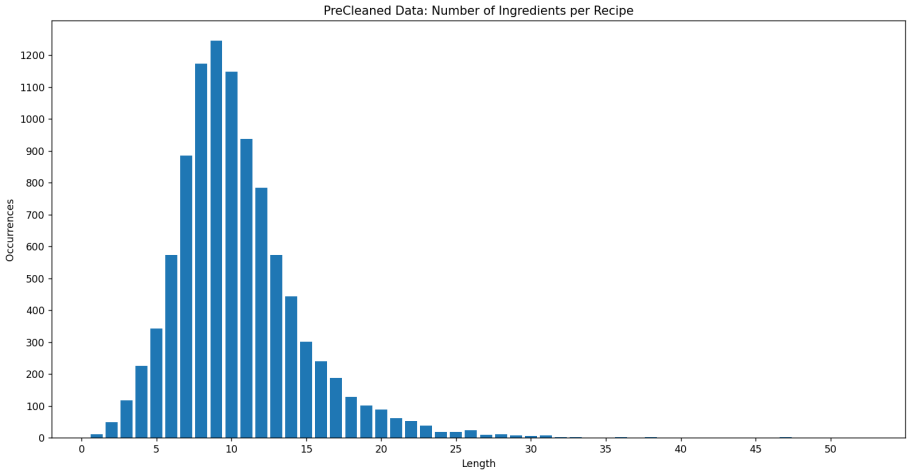
Tabella 3.4: Statistiche misurate relative al numero di parole presenti nel campo testuale del singolo ingrediente



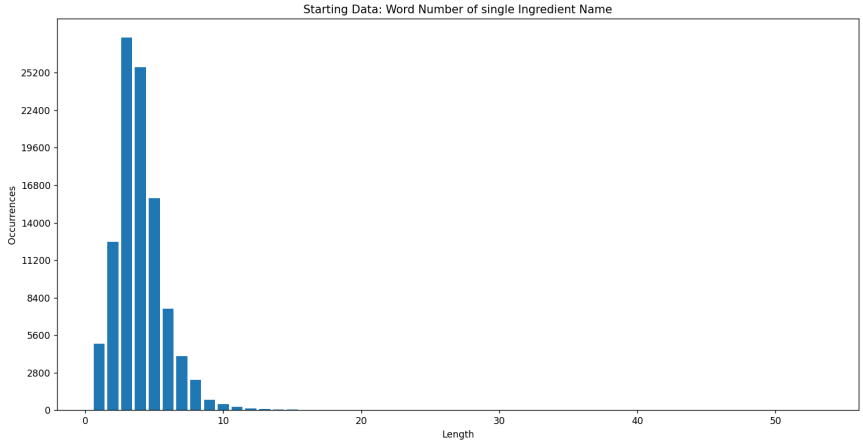
(a) Dati di partenza



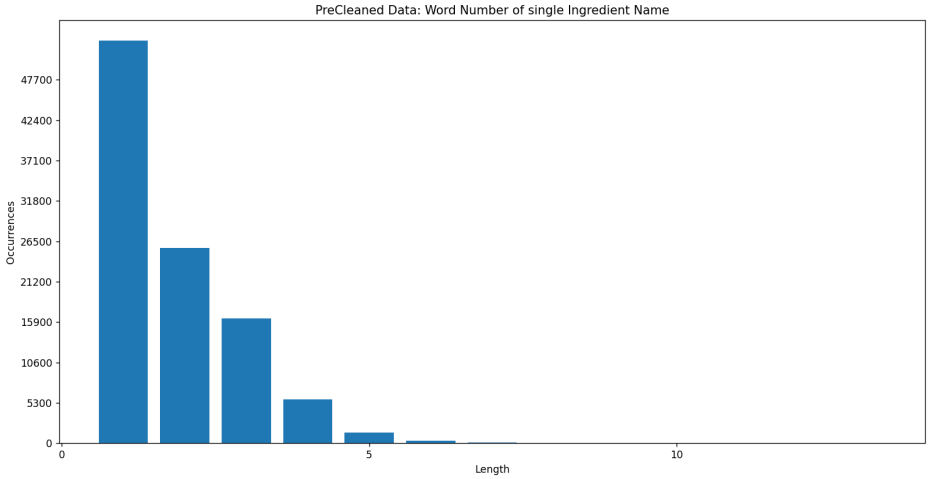
(b) Dopo pulizia ingrediente



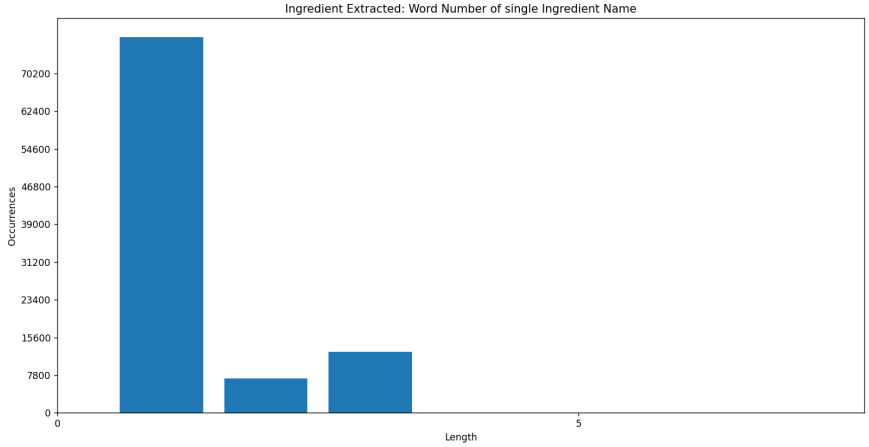
(c) Dopo Estrazione ingrediente



(a) Dati di partenza



(b) Dopo pulizia ingrediente



(c) Dopo Estrazione ingrediente

3.7 Normalizzazione delle quantità degli ingredienti

Una volta gestite le quantità e il nome degli ingredienti, in questa ultima fase, viene aggiunto un nuovo campo denominato $norm\_on\_100gr$ ad ogni ingrediente di una ricetta.

Questo campo rappresenta la percentuale della quantità che l'ingrediente ricopre rispetto alla quantità totale in grammi della ricetta. Quest'utlima è stata calcolata sommando le quantità di ogni ingrediente che la compone. La normalizzazione delle quantità degli ingredienti è stata effettuata dividendo la quantità di ogni ingrediente per la quantità totale in grammi della ricetta e moltiplicando il risultato ottenuto per 100. In questo modo, ogni ingrediente è stato espresso come una percentuale della quantità totale della ricetta.

In ultima battuta ora ogni ingrediente presente nelle ricette del dataset è un dictionary simile a:

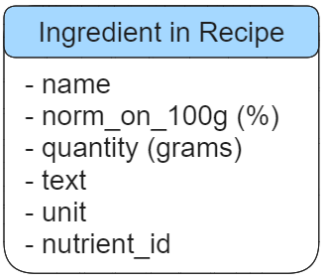


Figura 3.6: Schema UML della struttura Ingrediente in Ricetta, al termine della pipeline di preprocessing.

Capitolo 4

Ricerca e Rimozione Pseudo-Duplicati

4.1 Introduzione

In questo capitolo sono analizzati nel dettaglio gli esperimenti e i relativi risultati ottenuti per quanto riguarda la rimozione di pseudo-duplicati potenzialmente presenti all'interno del dataset di ricette.

Gli pseudo-duplicati sono qui definiti come coppie di ricette il cui titolo non corrisponde appieno ma la ricetta risulta fondamentalmente la stessa, per ingredienti e preparazione, seppur con alcune differenze trascurabili.

Una prima analisi è stata effettuata nelle fasi iniziali del processo di pulizia e raffinamento del dataset, impiegando una rete neurale per effettuare text-similarity.

Notati i risultati poco utili di questa prima analisi, le analisi successive sono state effettuate invece alla fine del processo di pulizia, andando a lavorare con dati più puliti ed elaborati.

Nelle pagine successive è descritta inizialmente la prima analisi effettuata con la rete neurale. Successivamente sono spiegate gli altri approcci che impiegano invece soluzioni algoritmiche principalmente concentrate sugli ingredienti, tralasciando il titolo e la portata.

4.2 Prima Ricerca degli Pseduo-Duplicati

4.2.1 Preambolo

Questa prima analisi per evidenziare la presenza di pseudo-duplicati, è stata eseguita subito dopo la rimozione dei duplicati semplici dal dataset.

In queste fasi iniziali, il dataset era ancora grezzo e per tanto la maggior parte degli ingredienti delle ricette era affetto da un grosso quantitativo di rumore internamente ai campi testuali.

Il motivo degli scarsi risultati ottenuti da questi esperimenti preliminari, può essere attribuito con una buona probabilità a questo rumore iniziale. È necessario ricordare che con rumore qui si intendono parole da considerarsi superflue al fine dell'identificazione dell' ingrediente.

Queste parole, presenti nel campo testuale dell'ingrediente stesso, non portavano nessuna informazione riguardante la natura dell'ingrediente in esame, e pertanto è probabile che portassero in errore la rete neurale qui usata.

4.2.2 L'esperimento con la rete neurale

Per effettuare una analisi di similarità tra le varie combinazioni di coppie di ricette, al fine di verificare l'esistenza di pseudo-duplicati, è stata impiegata una rete neurale che effettuasse l'embedding testuale. Successivamente veniva calcolato la distanza tra i due vettori di embedding e questa distanza è utilizzata per andare a definire dei valori di similarità.

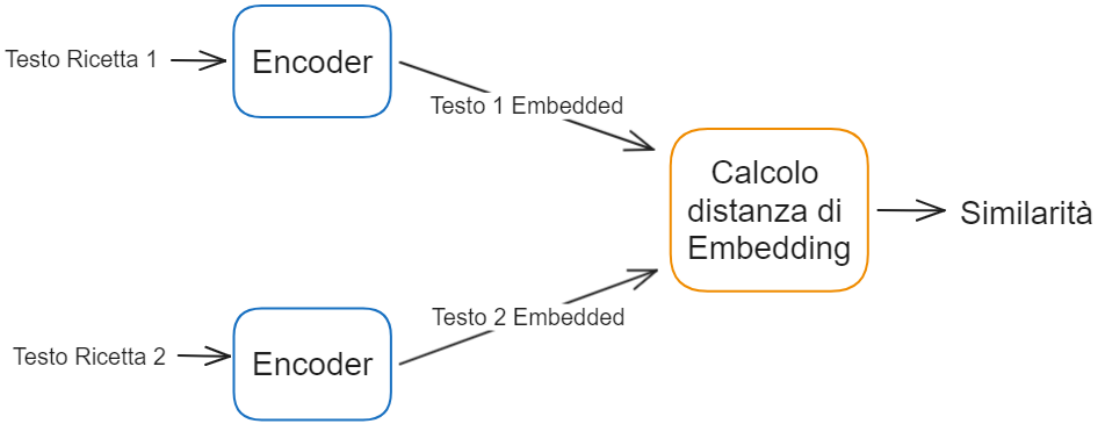


Figura 4.1: Schema di analisi di similarità testuale.

Come si nota nella figura 4.1, lo schema è generico ed è applicato separatamente dapprima sui Titoli delle ricette e successivamente sulle portate. Nell'ultima fase di questo esperimento, sarà utilizzato sugli ingredienti in diversi modi.

Per prima cosa occorre notare che per misurare la similarità tra due ricette, confrontare anche i loro rispettivi ingredienti porta a dei risultati più completi.

Per misurare quindi la similarità tra due ricette nel modo più meticoloso possibile, è necessario confrontare Titolo, Portata e Ingredienti.

Proprio per questo motivo, l'analisi viene effettuata in primo luogo su Titolo e Portata, e successivamente sugli ingredienti.

Per questo scopo al fine di misurare la similarità tra le varie coppie di ricette è stata utilizzata una rete neurale per fare text-similarity; la rete neurale è stata implementata usando la libreria di python 'spacy', dedicata ad operazioni di Natural Language Processing.

Per ottenere la lista di tutte le coppie ricetta-ricetta possibili sono state estratte tutte le combinazioni senza ripetizione di lunghezza due. Essendo una funzione fattoriale, la dimensione di questa lista si aggira intorno alle 50 milioni di entrate, con un dataset di partenza di poco meno di 9800 ricette circa.

L'idea è di effettuare prima una analisi composita sul nome della ricetta (title) e sulla portata (course) e successivamente selezionare tramite una threshold, decisa in maniera empirica, quali di queste ricette potessero rappresentare uno pseudo-duplicato.

Le ricette selezionate in questa prima fase, sono quindi analizzate nuovamente a livello dei singoli ingredienti che le compongono.

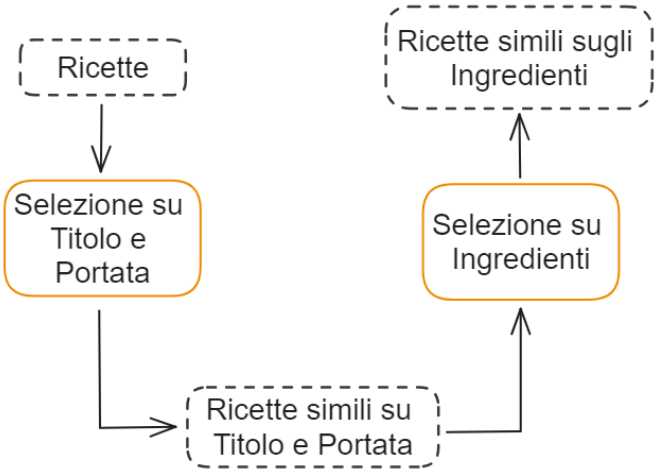


Figura 4.2: Schema riassuntivo della sequenza in cui sono effettuate le varie analisi di ricerca degli pseudo-duplicati sulle ricette.

4.2.3 Analisi di Simlarità su Titolo e Portata (Title and Course)

Per ogni coppia possibile di ricette, è state effettuata una prima analisi di similarità composita sul titolo e sulla portata (chiamata analisi TnC acronimo per Title and Course). L'idea semplice alla base è che se due ricette hanno nome simile e portata molto simile, se non identica, possono potenzialmente essere la stessa ricetta.

La metrica di similarità composita utilizzata è una interpolazione lineare delle similarità tra i titoli e delle similarità tra le portate. Le similarità testuali sono calcolate tramite una funzione della libreria spacy che calcola direttamente la similarità (similarità coseno) a partire dalle rappresentazioni dei due testi, esistenti internamente alla rete.

$$similarity = w_{title} \cdot sim_{title} + (1 - w_{title}) \cdot sim_{course}$$

Le due similarità nella formula, sim_{title} e sim_{course} sono pesate nell'interpolazione tramite due parametri, w_{title}, che indica il peso attribuito alla similarità calcolata tra i titoli delle ricette, e w_{course} che indica invece quello delle portate.

Il peso w_{title} è stato fissato empiricamente a 0.4. Il titolo pertanto influirà per il 40 % del valore di similarità mentre la portata per il 60%.

Il peso maggiore dato alla Portata è giustificato, poiché i titoli delle ricette sono spesso poco indicativi e possono essere molto soggetti a caratteristiche culturali che variano da regione a regione, pur potendo mantenere informazioni utili riguardo la ricetta stessa. Tendenzialmente invece, se due ricette condividono la stessa portata è più probabile che, se condividono anche titoli simili, le ricette siano vicine tra loro.

4.2.4 Analisi di Similarità sugli Ingredienti

Successivamente è stata eseguita una seconda analisi, separatamente a quelle precedente su Titoli e Portata, che confrontava gli ingredienti presenti nelle ricette.

Specialmente nel caso dell'impiego della rete neurale questa analisi appare decisamente pesante a livello computazionale e pertanto è stata effettuata solo su quelle coppie di ricette che superavano una certa soglia nella prima analisi sui Titoli e Portata.

La motivazione principale, oltre a quella legata intrinsecamente all'algoritmo che sfrutta cicli annidati, è che l'embedding degli ingredienti è fatto durante l'esecuzione, il che richiede molte risorse computazionali. L'idea cardine di questo algoritmo è quella di prendere la prima ricetta (r1) e confrontare ogni suo ingrediente (i) con tutti gli ingredienti (j) della seconda ricetta (r2).

Viene salvato come similarità da prendere in considerazione per quel singolo ingrediente i-esimo, la similarità massima ottenuta in questo confronto tra l'ingrediente i-esimo e tutti gli ingredienti j-esimi. Successivamente si passa al prossimo ingrediente i+1.

Successivamente queste similarità per ogni ingrediente i-esimo venivano salvate in una lista, e al termine del ciclo su tutti gli ingredienti i-esimi, veniva fatta una media delle similarità, presa come valore rappresentante della similarità tra r1 e r2

Per ogni coppia di ingredienti (i,j) viene presa la similarità massima poiché ci si aspetta che se entrambe le ricette condividono un ingrediente, la similarità tra l'ingrediente i-esimo e j_k-esimo sarà alta mentre per gli altri ingredienti j-esimi sarà bassa. Se invece la similarità tra gli ingredienti i-esimi e tutti gli ingredienti j-esimi di r2 è bassa, andranno ad abbassare la similarità nella media calcolata successivamente. La similarità è qui calcolata come la media aritmetica delle similarità massime tra tutte le coppie di ingredienti, ovvero la loro somma totale diviso la lunghezza della lista di ingredienti minore.

Inoltre prendere la similarità massima, era necessario perché gli ingredienti di due ricette potenzialmente simili, possono essere in un ordine diverso. Un semplice algoritmo di ordinamento non funziona poiché gli ingredienti in questa prima fase erano ancora grezzi e non contenevano solamente il nome dell'ingrediente. In questa fase l'ordinamento alfabetico dei nomi degli ingredienti era poco significativo per via del rumore che caratterizzava questi campi testuali.

4.3 Seconda Ricerca di Pseudo Duplicati dopo la fase di Preprocessing

In una fase successiva alla pulizia dei dati, è stato deciso di effettuare una seconda ricerca per identificare potenziali pseudo-duplicati.

Questa ricerca è effettuata su dati notevolmente più raffinati e puliti e in particolare caratterizzati da poco rumore nei campi testuali degli ingredienti.

Questo nuovo tentativo di ricerca si basa principalmente su un metodo algoritmico per confrontare due ricette sulla base degli ingredienti che le compongono. L'idea alla base è che due ricette sono tanto più simili quanto più simili sono gli ingredienti che le compongono, considerando anche le proporzioni degli stessi rispetto alla quantità totale della ricetta.

L'algoritmo qui utilizzato è frutto di diversi esperimenti effettuati con configurazioni differenti, nonostante l'idea alla base rimanga il confronto delle liste di ingredienti delle due ricette.

Nelle sotto sezioni seguenti saranno descritti i vari esperimenti effettuati che riepilogano i passaggi e le motivazioni che hanno portato all'evoluzione sia dell'algoritmo di confronto degli ingredienti, sia alla funzione di calcolo della similarità definitivi.

A grandi linee si possono definire due approcci distinti, differenziati principalmente per la chiave di ordinamento utilizzata per ordinare le liste degli ingredienti delle ricette.

Le due chiavi di ordinamento sono per nome (*name*), quindi un ordinamento alfabetico degli ingredienti, e per quantità di ingrediente ($norm\_on\_100g$), quindi per quale percentuale di peso (calcolata con la quantità del singolo ingrediente rispetto alla quantità totale della ricetta normalizzata a 100 grammi) l'ingrediente impatta sulla quantità totale calcolata della ricetta.

Questa seconda chiave di ordinamento è considerata quella con più significato, perché anche la percentuale ricoperta da un ingrediente rispetto alla sua ricetta, porta con se informazioni riguardo alla ricetta stessa. Per questo motivo è considerato l'ordinamento migliore.

È possibile effettuare una distinzione ulteriore di questi due approcci, andando a differenziare sulla base del metodo di confronto utilizzato per contare il numero di ingredienti condivisi dalle due ricette prese in esame alla ricerca di possibili pseudo-duplicati.

Tra i metodi esplorati è stato inizialmente utilizzato un confronto parallelo, andando a scorrere la lista di lunghezza minore e controllando se l'elemento corrispondente della lista maggiore fosse il medesimo. Questo approccio è stato il primo utilizzato poiché a seguito di un ordinamento sembrava essere la soluzione più intuitiva per individuare elementi in comune.

Successivamente è stato testato un confronto con sliding window, andando ad utilizzare una finestra di dimensioni parametrizzate, che permettesse di aumentare la tolleranza per la ricerca di una corrispondenza tra le due liste di ingredienti.

L'idea dietro a questa implementazione è che se due ingredienti che porterebbero ad individuare una corrispondenza si trovano in due posizioni leggermente diverse all'interno delle due liste, è sbagliato non considerare anche quella corrispondenza. Ciò permette di effettuare un confronto meno superficiale che tollera anche dei

leggeri sfasamenti tra ingredienti uguali all'interno delle liste di ingredienti.

In particolare questo appare più importante nell'approccio in cui si utilizza come chiave di ordinamento la percentuale di peso. Il motivo è che andando ad ordinare a livello numerico, potrebbero verificarsi degli sfasamenti tra ingredienti dovuti a differenze molto piccole se non trascurabili, esistenti tra i due numeri rappresentanti le quantità normalizzate.

poiché questo numero dipende dalle quantità registrate dagli ingredienti stessi (estratte durante le fasi precedenti della pipeline di preprocessing), dal momento in cui esiste una grande variabilità (vedi sotto sezione 3.5.1), nonché una discreta quantità di errori nella definizione delle quantità stesse degli ingredienti, già presenti nei dati iniziali, è necessario mantenere una tolleranza meno stringente per rendere questa analisi più robusta rispetto a queste problematiche.

L'ultimo approccio testato è una versione più complessa dell'approccio precedente con sliding window che implementa anche un concetto di peso per quanto riguarda la distanza registrata tra due corrispondenze individuate all'interno della finestra di tolleranza.

Un esempio potrebbe essere che se nella lista minore è preso in esame un ingrediente i_2 (dove 2 indica l'indice di un ingrediente i preso dalla lista minore), considerando una finestra di tolleranza con dimensione 2, se la corrispondenza individuata nella lista maggiore si trova alla posizione j_2, poiché i due ingredienti si trovano allo stesso indice questa corrispondenza avrà un peso maggiore in particolare se confrontata invece con una corrispondenza j_4. Questo a causa della maggiore distanza tra gli indici 2 e 4 rispettivamente dell'ingrediente i e dell'ingrediente j.

Con questo metodo, si intende mantenere una tolleranza più ampia per i motivi descritti nei paragrafi precedenti riguardo all'approccio sliding window normale, ma si pone un'enfasi più accentuata sulla distanza calcolata tra gli indici rispettivi degli ingredienti che portano alla corrispondenza individuata.

È facile intuire che questo approccio complica anche il calcolo della similarità tra due ricette.

4.3.1 Spiegazione Dettagliata degli Algoritmi di confronto utilizzati

In questa sotto sezione saranno analizzati i vari algoritmi di confronto utilizzati per effettuare le analisi di similarità delle ricette. Saranno descritti nel loro

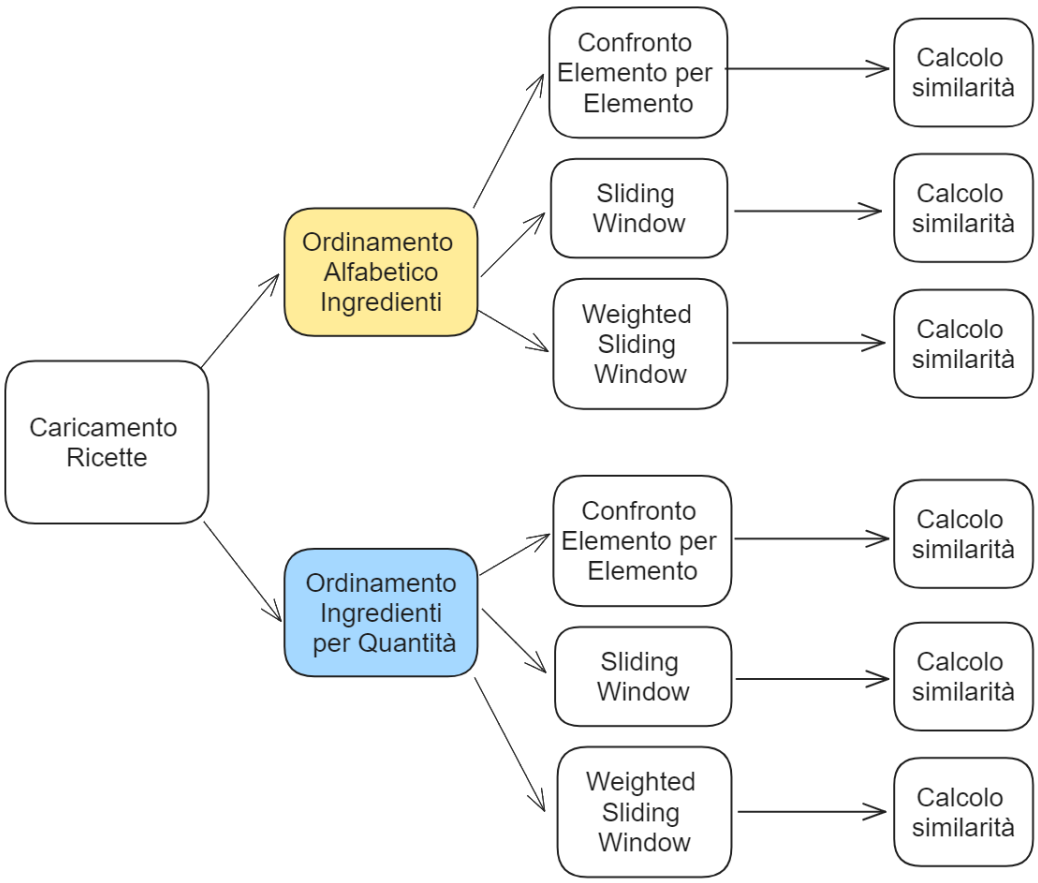


Figura 4.3: Schema esplicativo dei vari approcci implementati e sulla loro distinzione nelle diverse fasi dell'algoritmo.

funzionamento, e nelle motivazioni che giustificano il loro utilizzo, nonché nei loro limiti di applicazione e difetti.

Confronto parallelo

Per confrontare le due liste di ingredienti, si è utilizzato come primo metodo il Confronto parallelo. Questa funzione prende in ingresso le due liste contenenti i campi testuali degli ingredienti, e confronta gli elementi corrispondenti, uno per uno. Se gli elementi nella stessa posizione nelle due liste coincidono, viene contata la corrispondenza, e aggiunta al totale delle corrispondenze trovata, valore poi utilizzato per il calcolo della similarità.

Tra i pregi sono da citare la semplicità di implementazione (possibile con un

semplice ciclo su entrambe le liste in esame) e l'intuitività dell'approccio stesso. I difetti sono invece la scarsa flessibilità nei confronti di qualunque forma di sfasamento che può esistere tra due elementi uguali. Infatti se due elementi identici si trovano in indici diversi all'interno delle due liste questi non saranno contati.

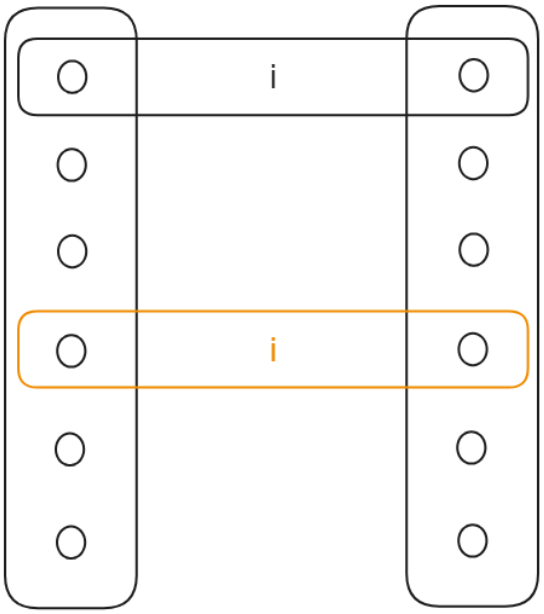


Figura 4.4: Schema confronto parallelo

Il presente schema in figura 4.4, mostra come si comporta la funzione di confronto durante la prima iterazione (in nero) nonché durante la quarta (in arancione). Ogni ingrediente i-esimo è direttamente, ed esclusivamente, confrontato con il corrispettivo ingrediente della seconda lista.

4.3.2 Confronto con Sliding Window

Per spiegare dettagliatamente il funzionamento della seguente funzione, occorre innanzitutto definire gli ingredienti della lista minore come elementi i e quelli della lista maggiore come elementi j. Inoltre si definisce il parametro ws come la dimensione del raggio di ricerca a partire dall'elemento i_k.

Per ogni elemento della lista di lunghezza minore i_k, la funzione apre una finestra di dimensione $1 + 2ws$ centrata sull'indice j_k della lista di lunghezza maggiore. In questa finestra viene confrontato il singolo elemento i_k con i diversi j_k appartenenti alla finestra precedentemente descritta, tra $j_k - ws$ e $j_k + ws$.

Questo secondo approccio è stato impiegato con lo scopo di aumentare la tolleranza agli sfasamenti possibili all'interno dell'ordine degli elementi delle due liste di ingredienti. Infatti con l'impiego di questa funzione, un leggero sfasamento minore o uguale a ws non rappresenta un problema nell'individuare delle corrispondenze.

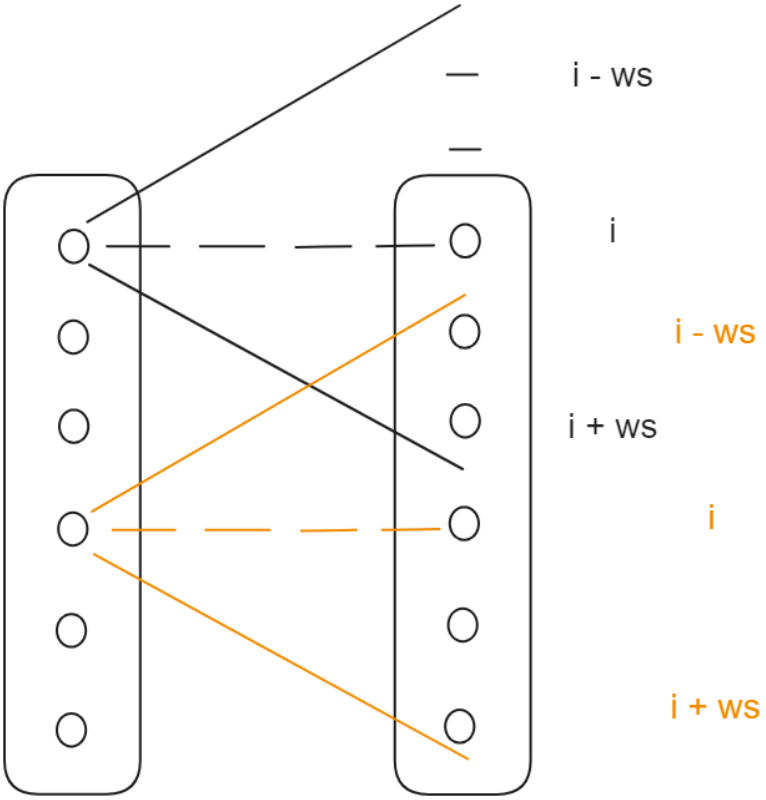


Figura 4.5: Schema confronto con Sliding Window

Il presente schema in figura 4.5, mostra come si comporta la funzione di confronto durante la prima iterazione (in nero) nonché durante la quarta (in arancione). La finestra di ricerca per una corrispondenza di ingredienti, ha in questa funzione, una dimensione maggiore rispetto alla precedente, permettendo valori di sfasamento relativo degli elementi maggiore. In particolare nella figura la dimensione della finestra è pari a 2.

La funzione prende in input le due liste di ingredienti da analizzare, e la dimensione del raggio della finestra definita precedentemente come ws . Essa scorre la lista minore, e cerca per ogni elemento i_k la corrispondenza nella finestra sugli elementi j_k come schematizzato dall'immagine 4.5.

È interessante inoltre notare che nell'ottica di una funzione di confronto di sliding window, la funzione di confronto parallelo ricopre il ruolo di caso degenere in cui la dimensione del raggio ws della finestra di ricerca ha dimensione pari a 0.

Confronto con Weighted Sliding Window (WSW)

Questo ultimo approccio nasce come ulteriore ottimizzazione dell'approccio precedente con SW. L'idea di questa modifica è quella di aumentare si la tolleranza rispetto agli sfasamenti che possono presentarsi tra elementi condivisi da entrambe le liste, così come già l'approccio SW faceva, ma di considerare comunque questo sfasamento pesandolo in funzione alla sua ampiezza.

Come la funzione precedente, si definiscono gli elementi della lista minore come elementi i e quelli della lista maggiore come elementi j. Inoltre, anche qui il parametro ws rappresenta la dimensione del raggio di ricerca a partire dall'elemento i_k.

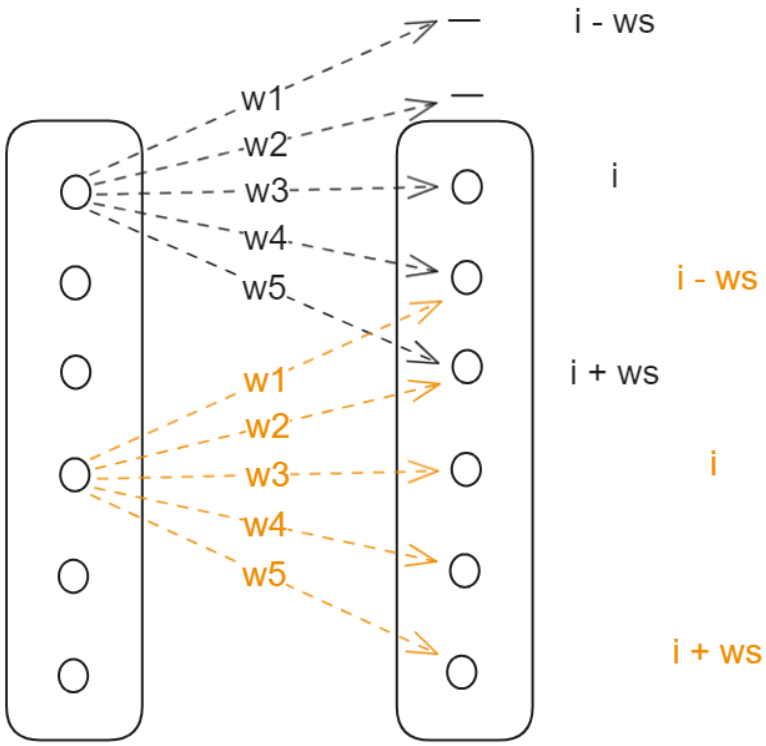


Figura 4.6: Schema confronto con Weighted Sliding Window

Il presente schema riportato in figura 4.6, mostra come si comporta la funzione di confronto durante la prima iterazione (in nero) nonché durante la quarta (in arancione). In questa funzione di confronto è invece lampante la presenza di pesi indicati come w_k. Ogni peso è definito a priori, con un valore minore quanto più il

peso è associato ad indici distanti dall'indice in esame i. In particolare nella figura la dimensione della finestra è pari a 2.

Per ogni elemento i_k della lista più corta, la funzione apre una finestra di dimensione $1 + 2ws$ centrata sull'indice j_k della lista più lunga. All'interno di questa finestra, l'elemento i_k viene confrontato con i vari j_k compresi tra $j_k - ws$ e $j_k + ws$.

A differenza della funzione precedente però, se la funzione individua una corrispondenza tra l'elemento i_k e un elemento j_k appartenente alla finestra, i suo conteggio non sarà più unitario, ma sarà prima moltiplicato per un parametro di peso w_r.

I parametri w_r funzionano come dei veri e propri pesi, che danno più o meno valore alla corrispondenza individuata in funzione della distanza esistente tra gli indici k, dell'elemento i_k, e r dei pesi w_r.

Così facendo è possibile tollerare sfasamenti entro una distanza ws ma non considerarli tutti in ugual modo. Inoltre con una funzione di peso è possibile aumentare anche la dimensione della finestra, senza temere di rendere il confronto una classica ricerca di elementi della lista minore in quella maggiore. Per fare ciò va però opportunamente definita la funzione di peso da implementare.

Infatti la distribuzione dei pesi può essere delle più varie, sia simmetrica che non, sia decrescente che crescente, e conseguentemente pone delle basi interessanti per approcci di confronti anche in altri ambiti al di fuori del presente.

Anche in questo caso è interessante sottolineare che questo approccio generalizza ulteriormente l'approccio precedente. Infatti l'algoritmo di confronto SW non è altro che un algoritmo WSW con una funzione di peso pari a 1 per ogni elemento della finestra in esame. In sostanza quest'ultimo approccio racchiude i due precedenti e se opportunamente configurato permette di essere usato per testare anche il confronto parallelo e il confronto con SW.

Nelle implementazioni riportate successivamente la scelta della distribuzione dei pesi è stata fatta con obiettivo principale la semplicità. Infatti le distribuzioni scelte sono [0.8, 1, 0.8] per una finestra con $ws = 1$ e [0.5, 0.7, 1, 0.7, 0.5] per una finestra con $ws = 2$. La distribuzione è simmetrica, per comprendere sfasamenti sia in avanti che indietro, ed è decrescente in valore rispetto al centro della finestra per attribuire una forma di "penalità" a quelle corrispondenze trovate a distanze maggiori.

Andando a modificare il valore ritornato da questa funzione attraverso i pesi, anche il calcolo della similarità ne sarà influenzato. Infatti se prima durante un confronto tra due liste di ingredienti, l'algoritmo individuava 3 corrispondenze, il valore ritornato era pari a 3. Adesso, a parità di dimensione della finestra, il valore ritornato potrebbe essere nella peggiore delle ipotesi 3×0.8 o 3×0.5 ovvero i casi in cui le corrispondenze siano trovate il più distanti possibili.

Questa variazione dipende chiaramente dalla scelta della distribuzione dei pesi, e in questo caso può ridurre dal 20% al 50% il valore di similarità restituito da questa funzione di confronto in funzione anche del valore del parametro ws.

4.3.3 Evoluzione dell'Algoritmo: Primo Approccio al Problema

Il primo approccio studiato, nato come una prima bozza della soluzione di questo problema, è stato un confronto parallelo degli ingredienti, alla ricerca di corrispondenze allo stesso indice nelle due liste. In questa prima fase il metodo per effettuare l'ordinamento era basato sull'ordinamento alfabetico.

L'idea dietro questo primo approccio era che, se due ricette condividono gli stessi ingredienti, se quest'ultimi sono posizionati in ordine alfabetico, eseguendo un confronto parallelo, sarebbe stato ottimale trovare delle eventuali corrispondenze a parità di indice.

Inizialmente viene caricata la tabella delle ricette, e per ogni ricetta, si va ad ordinare la lista degli ingredienti.

Una volta caricata la tabella delle ricette, vengono generate tutte le possibili coppie di ricette, ottenute come combinazioni di lunghezza due di tutte le ricette, senza reinserimento per evitare di confrontare una ricetta con sè stessa.

Queste combinazioni sono raccolte in una lista di circa 48 milioni di elementi, ovvero tuple di lunghezza due contenenti gli identificativi delle due ricette da confrontare. Si ricorda che gli identificativi sono le chiavi per risalire ad una ricetta all'interno della tabella sopracitata.

Per una maggiore efficienza in termini di memoria, la lista viene costruita a tempo di esecuzione sfruttando il generatore offerto direttamente dalla funzione "combinations" nativa della libreria "itertools" di python.

Successivamente si apre un ciclo effettuato sulla lista contenente le coppie di ricette da analizzare.

Qui viene innanzitutto estratta per ogni ricetta della coppia la lista degli ingredienti. Si ricorda che questa lista è una lista di dizionari, e ogni dizionario

contiene chiavi associate a diverse informazioni dell'ingrediente in esame, come "name", "quantity" o "norm_on_100g".

In questo fase viene effettuato un primo controllo, che misura la differenza percentuale tra le lunghezze delle due liste di ingredienti che descrivono le due ricette in esame. Se questa differenza percentuale supera una certa soglia, si passa direttamente al ciclo successivo, considerando la differenza nel numero di ingredienti troppo ampia. Questa soglia è nei risultati successivi pari al 20%, ovvero la lista di ingredienti maggiore ha un 20% in più di ingredienti rispetto alla lista di lunghezza minore. Questo valore è stato posto al 20% empiricamente, basandosi sulle analisi effettuate sui dati raffinati dopo le varie fasi di preprocessing descritte nei capitoli precedenti.

Lo scopo di questo controllo è quello di limitare i confronti per definizione inutili come ricette con 4 ingredienti e ricette con 9 ingredienti.

Se il controllo effettuato sulla differenza percentuale legata al numero di ingredienti è superato, vengono estratte per ognuna delle ricette della coppia, due liste: una lista dei nomi degli ingredienti e una lista delle quantità percentuale che questi ingredienti ricoprono nella ricetta.

Queste due liste sono successivamente combinate in una unica, andando a formare, per ognuna delle due ricette in esame, una lista che contenga una serie di dizionari ognuno con due chiavi: "name" per il nome della ricetta e "$norm\_on\_100g$" per quanto riguarda invece la quantità in percentuale ricoperta dall'ingrediente.

Su queste due liste ottenute, una per ricetta, viene quindi effettuata l'analisi di similarità precedentemente descritta.

Calcolo Similarità

La similarità è calcolata inizialmente come la percentuale di ingredienti che le ricette hanno in comune, rispetto alla lista di ingredienti più piccola.

Per il calcolo vero e proprio vengono contati gli ingredienti in comune tra le due liste a parità di indice.

Questo numero è successivamente diviso per il massimo tra 1 (per evitare divisioni per 0), e il numero di elementi che compongono la lista di ingredienti di lunghezza minore. Così facendo si ottiene una percentuale di ingredienti in comune rispetto alla lista di ingredienti minore.

Una volta calcolato il valore di similarità tra le due ricette, sulla base di un confronto effettuato tra gli ingredienti che le compongono, se questo valore supera una soglia di sbarramento, le due ricette sono aggiunte, sotto forma di tupla con i loro identificativi, in una lista poi salvata come file pickle.

4.3.4 Evoluzione dell'Algoritmo: Analisi Limiti e Criticità del Primo Approccio

Dopo i primi test effettuati con l'approccio precedentemente descritto, analizzando i primi risultati sono apparsi alcuni limiti dell'approccio stesso.

In particolare questo risultava troppo generico se utilizzato con soglie di similarità basse, poiché associava ricette decisamente diverse come simili, altrimenti risultava esageratamente rigido con soglie di similarità più stringenti, non trovando di fatto nessun pseudo-duplicato potenzialmente valido.

Uno dei principali problemi era legato al confronto ingrediente per ingrediente, molto stringente e sensibile anche a sfasamenti di una sola posizione tra lo stesso ingrediente all'interno delle due liste. Questo portava a non considerare un ingrediente in comune se gli indici non corrispondevano esattamente in entrambe le liste.

Un ulteriore motivo dei risultati poco soddisfacenti ottenuti dall'applicazione di questo primo approccio, si può ritrovare nell'ordinamento alfabetico, che migliora solo le prestazioni dal punto di vista computazionale, poiché considerando due ricette con gli stessi ingredienti, permette una ricerca lineare come il confronto parallelo, senza dover implementare una ricerca ciclica sulla lista maggiore per ogni ingrediente della lista minore, che porterebbe ad una complessità quadratica. Rimane il fatto che oltre a questo, un ordinamento alfabetico non porta con se alcun significato, nessuna informazione utile per il confronto degli ingredienti.

Infine questo ordinamento è particolarmente prono a creare sfasamenti tra gli ingredienti in comune, dato che se due ricette condividono potenzialmente tutti gli ingredienti tranne un, ma questo ingrediente che le differenzia si colloca per primo in ordine alfabetico, le due ricette saranno considerate completamente diverse, poiché tutti gli ingredienti in comune avranno uno sfasamento relativo di una posizione.

4.3.5 Evoluzione dell'Algoritmo: Miglioramenti e Approccio a Finestra Scorrevole

Per superare le limitazioni dell'approccio iniziale, sono state introdotte diverse modifiche significative all'algoritmo. Il cambiamento più rilevante è stato l'adozione di un metodo a finestra scorrevole (sliding window) per il confronto degli ingredienti, abbandonando il confronto parallelo rigido.

Nel nuovo approccio, per ogni elemento della lista più corta, viene controllata la sua presenza all'interno di una finestra di dimensione predefinita nella lista più

lunga. Questo metodo permette una maggiore flessibilità nel rilevamento degli ingredienti comuni, superando il problema degli sfasamenti di posizione.

L'ordinamento degli ingredienti è stato modificato, passando da un ordinamento alfabetico a uno basato sul valore di "norm_on_100g". Questo cambiamento riflette l'importanza relativa degli ingredienti nella ricetta, fornendo un confronto più significativo. Per quanto riguarda il calcolo della similarità, sono stati esplorati diversi approcci. Inizialmente, si è considerato il semplice rapporto tra il numero di elementi in comune e la lunghezza della lista più corta. Questo metodo, nella sua semplicità, offriva una misura diretta della proporzione di ingredienti condivisi.

Tuttavia, ci si è resi conto che questa metrica, da sola, non catturava pienamente la complessità delle relazioni tra le ricette. Si è quindi sperimentato un secondo approccio, che combinava questa proporzione di elementi comuni con la distanza media tra questi elementi nelle rispettive liste. L'idea alla base era di considerare non solo la presenza degli stessi ingredienti, ma anche quanto questi fossero "allineati" nelle due ricette. Spingendosi oltre, si è sviluppato un terzo metodo ancora più sofisticato. Questo approccio non solo teneva conto della proporzione di elementi comuni e della loro distanza, ma considerava anche la posizione relativa degli ingredienti, pesata in modo inversamente proporzionale alla loro importanza nella ricetta. Questa scelta è stata motivata dal fatto che gli ingredienti erano ordinati in base al loro valore "norm_on_100g", riflettendo così il loro impatto relativo nella composizione della ricetta.

Un aspetto emerso durante queste analisi è stato anche il trattamento degli ingredienti troppo comuni, come ad esempio olio, sale e pepe. Per affrontare questa problematica, è stata creata una lista specifica di questi ingredienti. Nel calcolo della similarità, questi elementi venivano esclusi sia dal conteggio degli elementi in comune che dalla lunghezza totale della lista, permettendo così un confronto più mirato sugli ingredienti veramente caratterizzanti di ciascuna ricetta.

4.4 Esperimenti e Analisi dei Risultati della Rimozione di Pseudo-Duplicati

In questa sezione, sono approfonditi i risultati ottenuti dai vari esperimenti condotti per valutare l'efficacia dei diversi approcci descritti nelle sezioni precedenti. In particolare sono confrontati i due criteri di ordinamento degli ingredienti, effettuati prima della fase di confronto (ordinamento alfabetico e per quantità normalizzata). Successivamente sono confrontate, per ognuno dei due ordinamenti, anche le tre funzioni di confronto stesse (confronto parallelo, con SW e con WSW.

4.4.1 Impostazione degli Esperimenti

I diversi esperimenti sono stati effettuati con diversi valori dei parametri principali dello script utilizzato. I parametri in questione sono:

- **Criterio di Ordinamento**:
  Il criterio di ordinamento indica come gli ingredienti sono ordinati all'interno delle liste, prima di essere poi analizzate con l' algoritmo di confronto (SW o WSW) per calcolare la similarità. I due criteri qui analizzati sono Ordinamento Alfabetico e Ordinamento per Quantità Normalizzata (su 100 grammi). Il tipo di ordinamento non è riportato nelle tabelle, ma è indicato nella sezione in cui si trova la tabella.

- **Soglia di Similarità**:
  La soglia di similarità è quel parametro che permette di decidere quali ricette sono potenziali pseudo-duplicati o meno, in base al loro valore di similarità. In questi esperimenti sono stati usati principalmente i valori 0.6, 0.7 e 0.8 per entrambi gli algoritmi di confronto SW e WSW. Per WSW sono stati anche studiati i risultati ottenuti con 0.4 e 0.5. Questo perché la funzione di peso dell'approccio WSW abbassava potenzialmente i valori di similarità, ed è si è ritenuto interessante approfondire il comportamento a soglie di similarità più basse.

- **Dimensione Finestra**:
  La dimensione della finestra è direttamente quella utilizzata negli algoritmi di confronto basati su Sliding Window. In questi esperimenti sono stati testati i valori $ws = 1$ e $ws = 2$. Questa scelta è stata presa empiricamente andando ad analizzare la lunghezza media delle liste di ingredienti.

- **Algoritmo di confronto**:
  L'algoritmo di confronto invece è un parametro che indica quale tipologia di confronto utilizzare tra SW o WSW.

L'approccio con WSW utilizza come funzione di peso [0.8, 1, 0.8], per la dimensione di finestra pari a 1. Per la dimensione di finestra pari a 2 invece utilizza [0.5, 0.7, 1, 0.7, 0.5]

4.4.2 Valutazione dei Risultati

Infine per ogni gruppo di potenziali pseudo-duplicati individuato dagli algoritmi, è stata effettuata una analisi manuale per stabilire quali di queste coppie siano effettivamente ricette estremamente simili se non addirittura identiche, da considerare a tutti gli effetti pseudo-duplicati. Successivamente è poi calcolata la percentuale

di pseudo-duplicati veri trovata all'interno del gruppo di candidati proposto da ogni algoritmo.

Per definire realmente uno pseudo-duplicato, oltre alla forte corrispondenza e somiglianza esistente tra le liste di ingredienti delle due ricette, viene anche fatta una analisi manuale che tiene conto anche delle somiglianze a livello visivo.

È importante tenere conto che l'obiettivo finale di questo lavoro è quello di costruire un dataset di ricette e framework di preprocessing di dati culinari, per aiutare lo sviluppo e l'implementazione di reti neurali dedicate a riconoscere il piatto a partire dall'immagine.

Pertanto se due ricette sono estremamente simili ma differiscono notevolmente a livello visivo, in forma, colore o altro saranno considerate due ricette distinte.

Spiegazione della Tabella dei Risultati

I risultati sono raccolti e riportati in delle tabelle come la seguente.

Soglia Similarità	0.6		0.7		0.8	
Dim. Finestra	-	-	-	-	-	-
Lunghezza lista	-	-	-	-	-	-
Veri Positivi	-	-	-	-	-	-
Richiamo (%)	-	-	-	-	-	-
Precisione (%)	-	-	-	-	-	-

Tabella 4.1: Tabella di esempio per analisi delle ricette simili

La tabella illustra i risultati dell'analisi delle ricette simili mediante i due algoritmi. Le tabelle sono inoltre divise in due sottosezioni, distinte per tipo di ordinamento.

Le prime due righe della tabella rappresentano i diversi parametri dell'analisi effettuata:

"Soglia Similarità" indica il valore soglia impiegato per determinare la similarità tra le ricette.

"Dim. Finestra" si riferisce alla dimensione della finestra utilizzata nell'implementazione dell'algoritmo. "Lunghezza lista" quantifica il numero totale di elementi nella lista risultante, comprendente i potenziali pseudo-duplicati identificati.

Le righe successive invece riportano i dati ottenuti con i parametri sopra riportati. "Veri Positivi" enumera gli pseudo-duplicati potenziali effettivamente individuati.

"Richiamo (%)" esprime la percentuale di pseudo-duplicati individuati rispetto al totale noto di pseudo-duplicati (pari a 63).

"Precisione (%)" invece è calcolata come il rapporto tra "Veri Positivi" e "Lunghezza lista", fornendo una misura della precisione dell'algoritmo, rispetto al numero di candidati individuati.

Risultati del Confronto Parallelo

A prescindere dall'ordinamento, il confronto parallelo è ottenuto ponendo $ws = 0$. Questo primo approccio di confronto però, non riporta nessun risultato, essendo la ricerca di corrispondenze parallele troppo stringente per questo tipo di ricerca. Per questo motivo non sono riportati i risultati nei paragrafi successivi.

4.4.3 Esperimenti con Ordinamento Alfabetico

L'ordinamento alfabetico, nonostante non ordini le liste di ingredienti secondo delle informazioni significative rispetto alla ricetta, si dimostra comunque abbastanza performante per individuare gli pseudo-duplicati. Infatti le differenze con l'ordinamento per percentuale su 100 grammi non sono esageratamente elevate.

Una possibile spiegazione di questa limitata differenza, può essere che i dati e il numero di pseudo-duplicati effettivamente presenti nel dataset di ricette, siano sostanzialmente pochi. Il dataset presenta poco meno di 10k ricette e il numero di pseudo-duplicati identificati è anch'esso ridotto, intorno a 60.

Confronto con Sliding Window

Soglia Similarità	0.6		0.7		0.8	
Dim. Finestra	1	2	1	2	1	2
Lunghezza lista	65	75	56	63	23	30
Veri Positivi	51	57	47	54	16	23
Richiamo (%)	80.95	90.48	74.6	85.71	25.4	36.51
Precisione (%)	78.46	76.0	83.93	85.71	69.57	76.67

Tabella 4.2: Risultati dell'analisi delle ricette simili (Algoritmo: SW)

Confronto con Weighted Sliding Window

Soglia Similarità	0.6		0.7		0.8	
Dim. Finestra	1	2	1	2	1	2
Lunghezza lista	60	37	35	13	10	2
Veri Positivi	50	32	27	11	6	2
Richiamo (%)	79.37	50.79	42.86	17.46	9.52	3.17
Precisione (%)	83.33	86.49	77.14	84.62	60.0	100.0

Tabella 4.3: Risultati dell'analisi delle ricette simili (Algoritmo: WSW)

Soglia Similarità	0.4		0.5		0.6	
Dim. Finestra	1	2	1	2	1	2
Lunghezza lista	136	113	96	71	66	49
Veri Positivi	59	57	58	53	54	46
Richiamo (%)	93.65	90.48	92.06	84.13	85.71	73.02
Precisione (%)	43.38	50.44	60.42	74.65	81.82	93.88

Tabella 4.4: Risultati dell'analisi delle ricette simili (Algoritmo: WSW)

4.4.4 Esperimenti con Ordinamento per Quantità Normalizzata

L'ordinamento per quantità normalizzata tiene conto della quantità percentuale di ogni ingrediente rispetto alla quantità totale della ricetta.

Confronto con Sliding Window

Soglia Similarità	0.6		0.7		0.8	
Dim. Finestra	1	2	1	2	1	2
Lunghezza lista	76	81	62	65	31	31
Veri Positivi	58	58	55	55	24	24
Richiamo (%)	92.06	92.06	87.3	87.3	38.1	38.1
Precisione (%)	76.32	71.6	88.71	84.62	77.42	77.42

Tabella 4.5: Risultati dell'analisi delle ricette simili (Algoritmo: SW)

Confronto con Weighted Sliding Window

Soglia Similarità	0.6		0.7		0.8	
Dim. Finestra	1	2	1	2	1	2
Lunghezza lista	66	49	45	21	22	3
Veri Positivi	54	46	38	20	18	2
Richiamo (%)	85.71	73.02	60.32	31.75	28.57	3.17
Precisione (%)	81.82	93.88	84.44	95.24	81.82	66.67

Tabella 4.6: Risultati dell'analisi delle ricette simili (Algoritmo: WSW)

Soglia Similarità	0.4		0.5		0.6	
Dim. Finestra	1	2	1	2	1	2
Lunghezza lista	136	113	96	71	66	49
Veri Positivi	59	57	58	53	54	46
Richiamo (%)	93.65	90.48	92.06	84.13	85.71	73.02
Precisione (%)	43.38	50.44	60.42	74.65	81.82	93.88

Tabella 4.7: Risultati dell'analisi delle ricette simili (Algoritmo: WSW)

4.4.5 Differenze tra Sliding Window e Weighted Sliding Window A parità di Ordinamento

Di seguito sono riportati le differenze tra i dati ottenuti da WSW e SW. In particolare, a parità di ordinamento, le differenze sono calcolate sottraendo al valore ottenuto con WSW quello ottenuto da SW. Se il valore è positivo, significa quindi che WSW riporta un valore maggiore di SW, se è negativo allora WSW riporta un valore minore.

Con Ordinamento Alfabetico

Soglia Similarità	0.6		0.7		0.8	
Dim. Finestra	1	2	1	2	1	2
Lunghezza lista	-5	-38	-21	-50	-13	-28
Veri Positivi	-1	-25	-20	-43	-10	-21
Richiamo (%)	-1.58	-39.69	-31.74	-68.25	-15.88	-33.34
Precisione (%)	4.87	10.49	-6.79	-1.09	-9.57	23.33

Tabella 4.8: Differenze nei risultati dell'analisi delle ricette simili (WSW - SW)

Con Ordinamento per Quantità Normalizzata

Soglia Similarità	0.6		0.7		0.8	
Dim. Finestra	1	2	1	2	1	2
Lunghezza lista	-10	-32	-17	-44	-9	-28
Veri Positivi	-4	-12	-17	-35	-6	-22
Richiamo (%)	-6.35	-19.04	-26.98	-55.55	-9.53	-34.93
Precisione (%)	5.50	22.28	-4.27	10.62	-4.67	-10.75

Tabella 4.9: Differenze nei risultati dell'analisi delle ricette simili (WSW - SW)

4.4.6 Analisi Comparativa tra Sliding Window e Weighted Sliding Window a parità di Ordinamento

Di seguito è presentata un'analisi schematica dei risultati ottenuti confrontando gli algoritmi Sliding Window e Weighted Sliding Window. Una particolare attenzione è stata rivolta all'impatto della funzione di peso utilizzata in WSW.

Funzione di Peso in WSW

WSW utilizza una funzione di peso che distribuisce l'importanza degli elementi all'interno della finestra di analisi. Per finestre di con $ws = 1$, i pesi sono [0.8, 1, 0.8], mentre per $ws = 2$ sono [0.5, 0.7, 1, 0.7, 0.5]. Questa distribuzione dei pesi influenza significativamente e negativamente le prestazioni di WSW rispetto a SW, specialmente a parità di soglie di similarità.

Impatto Generale

WSW ha costantemente generato liste più corte rispetto a SW, indipendentemente dal metodo di ordinamento utilizzato. Questa maggiore selettività può essere direttamente attribuita alla funzione di peso, che dà meno importanza alle corrispondenze individuate ai margini della finestra.

Efficacia nell'Identificazione degli Pseudo-duplicati

WSW ha identificato un numero inferiore di "Veri Positivi" in tutti i casi esaminati. La riduzione è risultata più marcata con dimensione della finestra pari a 2. Una possibile causa può essere probabilmente la distribuzione più ampia dei pesi (da 0.5 a 1).

Analisi Richiamo e Precisione

Richiamo WSW ha mostrato un richiamo inferiore in tutti gli scenari analizzati, con differenze particolarmente evidenti con soglie alte come 0.7 e dimensione della finestra 2: -68.25% per l'ordinamento alfabetico e -55.55% per l'ordinamento per quantità normalizzata. Questa riduzione potrebbe essere una conseguenza diretta della maggiore selettività indotta dalla funzione di peso.

Precisione WSW ha mostrato risultati misti, con miglioramenti in alcuni casi. Con ordinamento alfabetico, WSW si è distinto come approccio (+23.33%) con soglia 0.8 e finestra 2, nonostante questa soglia sia molto alta e principalmente utilizzata per analizzare gli algoritmi di confronto, non per essere effettivamente utilizzata. Con ordinamento per quantità normalizzata, i miglioramenti sono stati più frequenti, specialmente con soglia 0.6. Questi miglioramenti potrebbero essere attribuiti alla capacità della funzione di peso di abbassare il valore di alcune corrispondenze più marginali, e "preferire" non considerare alcune ricette come pseudo-duplicati.

Impatto dell'Ordinamento

L'ordinamento per quantità normalizzata sembra ridurre le differenze tra SW e WSW. Questo potrebbe suggerire una sinergia tra questo tipo di ordinamento e la funzione di peso utilizzata in WSW. Infatti la distanza tra le corrispondenze è legata alla quantità che gli ingredienti occupano all'interno delle rispettive ricette.

Effetto della Dimensione della Finestra

Le differenze tra SW e WSW sono generalmente più pronunciate con dimensione della finestra 2. Questo effetto potrebbe essere dovuto alla distribuzione più ampia dei pesi nella finestra più grande, che amplifica l'impatto della pesatura. Inoltre la funzione di peso per finestre di dimensione 2 è più penalizzante rispetto alla funzione di peso per finestre di dimensione 1. Infatti se calcoliamo la media aritmetica come la somma dei valori della funzione peso, fratto il numero degli stessi otteniamo $3.4/5 = 0.68$ per $ws = 2$, mentre otteniamo $2.6/3 = 0.87$ per $ws = 1$. Questo indica, ipotizzando una distribuzione uniforme delle corrispondenze trovate durante il confronto, che la funzione legata a $ws = 2$ penalizza statisticamente di più il valore della similarità finale tra due ricette.

Considerazioni Finali

La funzione di peso in WSW gioca un ruolo cruciale nelle differenze osservate rispetto a SW, influenzando sia la selettività che la precisione dell'algoritmo. La maggiore

selettività di WSW è probabilmente dovuta alla diminuzione dell'importanza degli elementi ai margini della finestra. Mentre WSW trova meno pseudo-duplicati in termini assoluti, in alcuni casi mostra una migliore precisione, suggerendo una potenziale accuratezza migliore nell'identificazione delle corrispondenze più rilevanti.

È molto probabile che l'efficacia di WSW vari significativamente con diverse funzioni di peso, e i risultati osservati sono specifici per le funzioni di peso utilizzate in questo studio. L'ordinamento per quantità normalizzata sembra favorire le prestazioni di WSW, indicando una possibile interazione positiva con la funzione di peso, in linea con le aspettative e i motivi che hanno ispirato la scelta di questo ordinamento. La scelta tra SW e WSW dovrebbe considerare non solo i risultati numerici, ma anche l'impatto della funzione di peso per il dominio specifico dei dati analizzati.

Future ricerche potrebbero concentrarsi sull'ottimizzazione della funzione di peso in WSW, sperimentando con diverse distribuzioni per migliorarne ulteriormente le prestazioni in scenari specifici.

4.4.7 Differenze tra Ordinamento Alfabetico e Ordinamento per Quantità Normalizzata a Parità di Algoritmo

Di seguito sono riportati le differenze tra i dati ottenuti da WSW e SW rispetto però all'ordinamento. In particolare, sono confrontati i valori ottenuti, a parità di algoritmo di confronto, utilizzando i due criteri di ordinamento. Come riferimento è preso l'ordinamento per quantità normalizzata.

I dati riportati nelle tabelle sono quindi frutto della differenza tra il valore ottenuto ad esempio con l'algoritmo SW, ma su ordinamento per quantità normalizzata, meno il valore ottenuto sempre con SW ma con ordinamento alfabetico. Se il valore della differenza è positivo, allora il valore ottenuto con ordinamento per quantità normalizzata sarà maggiore rispetto a quello ottenuto con ordinamento alfabetico.

Soglia Similarità	0.6		0.7		0.8	
Dim. Finestra	1	2	1	2	1	2
Lunghezza lista	+11	+6	+6	+2	+8	+1
Veri Positivi	+7	+1	+8	+1	+8	+1
Richiamo (%)	+11.11%	+1.58%	+12.7%	+1.59%	+12.7%	+1.59%
Precisione (%)	-2.14%	-4.4%	+4.78%	-1.09%	+7.85%	+0.75%

Tabella 4.10: Differenze nei risultati dell'analisi delle ricette simili (Algoritmo: SW, Quan.Norm - Alfabetico)

Soglia Similarità	0.6		0.7		0.8	
Dim. Finestra	1	2	1	2	1	2
Lunghezza lista	+6	+12	+10	+8	+12	+1
Veri Positivi	+4	+14	+11	+9	+12	0
Richiamo (%)	+6.34	+22.23	+17.46	+14.29	+19.05	0.00
Precisione (%)	-1.51	+7.39	+7.30	+10.62	+21.82	-33.33

Tabella 4.11: Differenze nei risultati dell'analisi delle ricette simili (Algoritmo: WSW, Quan.Norm - Alfabetico, 0.6-0.8)

Soglia Similarità	0.4		0.5		0.6	
Dim. Finestra	1	2	1	2	1	2
Lunghezza lista	+20	-5	+25	+8	+6	+12
Veri Positivi	+7	0	+7	0	+4	+14
Richiamo (%)	+11.11	0.00	+11.11	0.00	+6.34	+22.23
Precisione (%)	-1.45	+2.13	-11.41	-9.48	-1.51	+7.39

Tabella 4.12: Differenze nei risultati dell'analisi delle ricette simili (Algoritmo: WSW, Quan.Norm - Alfabetico, 0.4-0.6)

4.4.8 Analisi Comparativa tra Ordinamento per Quantità Normalizzata e Ordinamento Alfabetico

L'analisi dei risultati ottenuti confrontando l'ordinamento per Quantità Normalizzata con l'Ordinamento Alfabetico per i diversi algoritmi e soglie di similarità ha rivelato diverse osservazioni interessanti, con particolare attenzione all'impatto del metodo di ordinamento sulle prestazioni degli algoritmi SW e WSW.

Impatto sull'Algoritmo Sliding Window

L'ordinamento per Quantità Normalizzata ha mostrato una tendenza a produrre liste di candidati più lunghe e a identificare un maggior numero di "Veri Positivi" rispetto all'ordinamento Alfabetico, in particolare nell'algoritmo SW.

Il richiamo ha beneficiato di questo approccio, con incrementi fino al 12.7%. La precisione, mostra invece risultati più variabili, con miglioramenti più evidenti per soglie di similarità più alte (0.7 e 0.8). È interessante notare che l'effetto dell'ordinamento per Quantità Normalizzata è risultato più pronunciato con dimensione della finestra 1 rispetto a 2.

Effetti su Weighted Sliding Window con Soglie Alte

Per WSW con soglie di similarità alte (0.6-0.8), l'ordinamento per Quantità Normalizzata ha confermato la tendenza a generare liste più lunghe e a identificare un maggior numero di "Veri Positivi". Il richiamo ha mostrato miglioramenti notevoli, con incrementi fino al 22.23%.

La precisione evidenzia miglioramenti nella maggior parte dei casi, con un'eccezione per la soglia 0.8 con finestra 2. L'effetto dell'ordinamento per Quantità Normalizzata sembra essere più consistente per WSW rispetto a SW, specialmente per le soglie 0.7 e 0.8.

Impatto su Weighted Sliding Window con Soglie Basse

Nel caso di WSW con soglie di similarità basse (0.4-0.5), il pattern di liste più lunghe e più "Veri Positivi" identificati è mantenuto, seppur con alcune eccezioni.

Il richiamo mostra miglioramenti, ma in misura minore rispetto alle soglie più alte. La precisione presenta risultati più variabili, con diversi casi di peggioramento. L'effetto dell'ordinamento per Quantità Normalizzata appare quindi meno consistente rispetto agli scenari con soglie più alte. Questo può essere legato al fatto che una soglia bassa rende il processo di individuazione di pseudo-duplicati meno selettivo.

Considerazioni Generali

L'analisi complessiva suggerisce che l'ordinamento per Quantità Normalizzata tende a migliorare la capacità di identificare "Veri Positivi" in quasi tutti gli scenari esaminati, indipendentemente quindi dall'algoritmo di confronto. Il richiamo beneficia di questo ordinamento. Anche l'impatto sulla precisione, sebbene più variabile, è generalmente positivo, specialmente per soglie di similarità più alte. WSW sembra trarre maggior vantaggio dall'ordinamento per Quantità Normalizzata rispetto a SW.

È interessante notare che l'effetto di questo ordinamento è generalmente più pronunciato con dimensione della finestra 1 rispetto a 2. Una possibile spiegazione potrebbe essere che con finestre di analisi più ampie, gli algoritmi di confronto tendano ad identificare più candidati, e portando alla luce una lista di candidati più lunga, questi abbiano più possibilità di identificare i veri pseudo-duplicati. Con finestre di dimensione minore invece, identificano meno candidati e quindi l'impatto dell'ordinamento sulla qualità delle selezioni fatte dagli algoritmi sia più netto.

Inoltre è utile ricordare che la lunghezza media delle liste di ingredienti è compresa tra 5 e 15, con una netta media su 10 ingredienti per lista e ricordare che una finestra di dimensione 1, o meglio con $ws = 1$ è una finestra di dimensione

$2ws+1$ centrata sui vari ingredienti in analisi della lista più lunga. Infatti con queste due informazioni si può capire come una finestra complessivamente larga 5, sia meno selettiva nel pesare l'informazione della quantità integrata nella lista tramite questo ordinamento, specialmente se confrontata con una finestra di dimensione complessiva pari a 3.

Questi risultati indicano che l'ordinamento per Quantità Normalizzata appare come una strategia efficace per migliorare le prestazioni sia di SW che di WSW, soprattutto riguardo alla capacità di identificare un maggior numero di pseudo-duplicati veri. Tuttavia, l'ottimizzazione della precisione potrebbe richiedere una calibrazione attenta dei parametri, specialmente per WSW con soglie di similarità più basse.

Queste conclusioni non stupiscono particolarmente, poiché un tipo di ordinamento per Quantità Normalizzata prende in considerazione delle informazioni che hanno valenza semantica all'interno della ricetta. Infatti le ricette possono contenere gli stessi ingredienti, ma in quantità nettamente diverse, e ciò può già portare a considerare le ricette come diverse.

Capitolo 5

Database Ingredienti Estratti e Categorie Personalizzate

5.1 Costruzione Database Ingredienti Estratti

Durante la fase di estrazione, all'interno della pipeline di preprocessing, è stata menzionata una tabella di riferimento utilizzata come base per effettuare delle ricerche di corrispondenze tra il contenuto del campo testuale degli ingredienti, e delle stringhe rappresentati ingredienti veri e propri selezionati manualmente.

Proprio questi ingredienti, sono stati ottenuti andando a analizzare in maniera iterativa tutti quegli output generati dalla fase di estrazione. Ad ogni iterazione si analizzavano quegli ingredienti che non superavano la soglia ma erano corretti, oppure che venivano associati ad ingredienti della tabella sbagliati.

In questi casi veniva analizzato brevemente l'ingrediente e se ritenuto necessario, veniva aggiunto alla tabella di riferimento.

Così facendo iterazione dopo iterazione, le corrispondenze individuate tramite LD durante questa fase crescevano, diversamente dal numero di errori o ingredienti ignorati che invece diminuiva.

La tabella di riferimento conta approssimativamente 550 ingredienti, inferiore ai circa 700-800 ingredienti (effettivamente distinti, quindi senza contare differenze come "sott'olio", "sotto sale", "al naturale") che sembrano esistere, guardando anche altri database costruiti con più rigore come quelli analizzati nel capitolo 6.

La tabella stessa è presente all'interno della code base con il nome di "extraction_ingredient_list" in formato JSON. Riporta per l'appunto una lista contenente

tutti i 500+ ingredienti, ordinati in ordine alfabetico.

Questa tabella è utilizzata anche in fase di creazione del file di collegamento nonché anche nella creazione del file di categorizzazione degli stessi ingredienti estratti.

5.2 Categorizzazione Ingredienti secondo Categorie Personalizzate

Il cuore di questa fase consiste nel classificare gli ingredienti estratti, presenti nel dataset, in categorie ben definite.

5.2.1 Processo di Categorizzazione

Per effettuare la classificazione degli ingredienti, dopo aver deciso le categorie più utili per gli scopi del progetto, ad ogni categoria è associato un indice numerico di riferimento chiamato *cat_idx* (category index, ad esempio 0 per vegetali, 1 per frutta, 2 per la frutta secca etc.).

Per attuare effettivamente la categorizzazione in maniera manuale, è stata dapprima costruita una tabella (sotto forma di dizionario Python) contenente come chiavi i nomi degli ingredienti, già puliti ed estratti durante i precedenti step di preprocessing, e come valore l'indice della categoria *cat_idx*.

Inizialmente, a ciascun ingrediente è stato assegnato in modo casuale un valore numerico estratto uniformemente dall'intervallo definito tra 0 e il numero di categorie meno uno. Così facendo l'intero *cat_idx* identifica in modo provvisorio la categoria di appartenenza. Questo dizionario preliminare è stato quindi salvato in un file JSON.

Successivamente, è stata effettuata una lettura manuale di questo stesso file JSON.

Per ogni ingrediente, viene effettuata una classificazione scegliendo una tra le diverse categorie predefinite definite nelle sezioni precedenti.

Questo processo di classificazione manuale entrata per entrata ha portato alla creazione di un nuovo file, denominato "$original - ing - cat\_table$" un file JSON che mantiene la tabella che permette di associare ad ogni ingrediente un indice di categoria. Questo file di riferimento contiene quindi tutti gli ingredienti correttamente classificati secondo gli indici cat_idx definitivi.

Il file "$original - ing - cat\_table$" rappresenta ora la versione di riferimento per la classificazione degli ingredienti. Tuttavia, è possibile effettuare ulteriori

aggiornamenti creando nuovi file e incorporando in essi le voci già presenti nell' "$original - ing - cat\_table$". Ogni nuova versione aggiornata può essere quindi salvata come la versione corrente della tabella contenente tutti gli ingredienti classifcati.

Durante l'analisi manuale, alcuni ingredienti sono stati etichettati come "Miscellanea" assegnando loro un indice -1. Questi casi particolari, come "bicarbonato", "stevia" o alcuni elementi decorativi, potranno essere rivalutati successivamente per decidere se rimuoverli in quanto non rappresentano reali ingredienti, o se categorizzarli adeguatamente, eventualmente consultando anche il relatore.

È importante notare che questo processo è costruito per essere iterativo e migliorare iterazione dopo iterazione, andando a migliorare sia la classificazione sia il numero di ingredienti non classificati, che tende a diminuire.

Inoltre questo approccio permette alla tabella di rimanere flessibile nei confronti di modifiche alle categorie.

Questa fase di classificazione a mano degli ingredienti ha permesso inoltre di migliorare e integrare le tabelle contenenti gli ingredienti stessi e le parole da rimuovere durante il preprocessing, aggiungendo nuove entrate di ingredienti e parole inutili individuate durante l'analisi. Ciò consente di affinare ulteriormente la pulizia e l'estrazione degli ingredienti, considerando anche elementi che potrebbero essere stati precedentemente scartati.

5.2.2 Categorie Personalizzate Scelte

Di seguito sono elencate tutte le categorie scelte per classificare a più alto livello gli ingredienti del dataset.

- **Vegetali**: Questa categoria contiene verdure come zucchini, pomodori o patate ma anche verdure a foglie come insalata, spinaci etc.

- **Frutta**: Questa categoria contiene tipi di frutta come mela, banane, pesche etc.

- **Frutta Secca**: Questa categoria contiene la frutta a guscio come noci, mandorle anacardi o pistacchi.

- **Carne bianca**: Questa categoria contiene carni bianche come pollame o coniglio.

- **Carne rossa**: Questa categoria contiene carni rosse come maiale, manzo e pecora ma anche la carne rossa macinata.

- **Pesce**: Questa categoria contiene tutti i pesci come merluzzo, salmone orate o sgombro.

- **Molluschi e crostacei**: Questa categoria contiene molluschi come cozze o vongole ma anche crostacei come aragoste, gamberi etc.

- **Salumi e insaccati**: Questa categoria contiene salumi come salame, prosciutti, ma anche salamini o salsicce.

- **Latticini**: Questa categoria contiene i latticini come formaggi, yogurt o latti animali.

- **Uova**: Questa categoria contiene le uova, sia intere che solo tuorli o albumi.

- **Legumi**: Questa categoria contiene tutti i legumi come fagioli, piselli, fave etc.

- **Carboidrati semplici**: Questa categoria contiene sia le farine semplici come la 0 o 00, sia le paste ma anche gli zuccheri e i dolci.

- **Carboidrati complessi**: Questa categoria contiene invece le farine meno raffinate, le farine integrali, il riso integrale o anche prodotti di farina integrale come pani integrali o paste.

- **Grassi vegetali**: Questa categoria contiene oli vegetali sani, come olio di oliva o di girasole, ma anche la margarina.

- **Grassi animali**: Questa categoria contiene sia grassi animali come strutto burro, ma anche alcuni oli meno sani come olio di palma etc.

- **Spezie e condimenti**: Questa categoria contiene le spezie come curry, pepe, zenzero ma anche condimenti come aceto, succo di limone, o erbe aromatiche come rosmarino, timo etc.

- **Alcol**: Questa categoria contiene tutte le bevande contenenti alcol, come vini, birre, liquori vari.

- **Te tisane e caffe**: Questa categoria contiene bevande ad infusione come te caffe e tisane.

- **Miscellanea**: Questa categoria raccoglie tutti quegli ingredienti non categorizzabili nelle altre categorie o addirittura quegli ingredienti su cui esiste un dubbio che siano effettivamente da considerarsi tali. Con questa categoria si aggiunge flessibilità rispetto agli ingredienti e si permette in futuro di analizzare eventualmente in maniera diversa gli elementi così caratterizzati.

5.3 Categorie BDA

Il database di ingredienti BDA, riporta oltre ai numerosi dettagli relativi agli ingredienti, anche una loro categorizzazione effettuate utilizzando delle categorie diverse.

Le categorie riportate sono le seguenti:

- TUBERI, PATATE, FECOLA (1)
- VERDURE, FUNGHI (2)
- LEGUMI E PRODOTTI DELLA SOIA (3)
- FRUTTA FRESCA, SECCA, FARINE, SUCCHI (4)

- LATTE E YOGURT (6)*
- FORMAGGI (7)
- CEREALI, FARINE, PASTA, PANE, CRACKERS (8)
- INSACCATI E SALUMI (9)
- CARNI E FRATTAGLIE (10)
- PESCI, CROSTACEI, MOLLUSCHI (11)
- UOVA (12)
- OLI, MARGARINE, BURRO, PANNA (13)
- DOLCIUMI, ZUCCHERO, MARMELLATE, GELATI (14)
- BRIOCHES, MERENDINE, BISCOTTI, BUDINI, TORTE (15)
- BEVANDE ANALCOLICHE E ALCOLICHE, CACAO, CAFFE, TE, INFUSI (16)
- ERBE AROMATICHE, SPEZIE (17)
- MISCELLANEA (18)

(*La categoria numero 5 non è riportata, non per errore ma perché effettivamente assente nel file)

Importante sottolineare che le categorie sono integrate tramite l'utilizzo di alcuni schemi integrati all'interno dei codici identificativi.

Ogni categoria infatti, ha un prefisso di 2 cifre seguito da altre 3 cifre che identificano una sotto categoria più precisa e caratterizzata. Queste cifre identificano a due livelli la categoria in cui è inserito l'ingrediente e vanno a comporre, insieme, il prefisso del codice dell'ingrediente specifico nel database.

È necessario anche evidenziare che il livello di dettaglio che riporta attraverso la seconda parte del prefisso è molto elevato. D'altronde questo database si presenta come estremamente rigoroso e con scopi medici.

Infatti un esempio potrebbe essere il seguente:

DOLCIUMI, ZUCCHERO, MARMELLATE, GELATI

- torrone

- cioccolatini, tavolette e creme spalmabili
- frutta candita
- zucchero e miele
- caramelle, liquirizia, confetti
- marmellate
- gelati, ghiaccioli
- dolcificanti
- sciroppi

Per il livello di dettaglio attualmente desiderato nel calcolare i valori nutrizionali delle ricette, quello riportato da questa classificazione è a tratti esagerato. Per questo motivo è stata implementata una categorizzazione personalizzata, che attualmente rispecchia meglio il livello di dettaglio ideale per il progetto.

Capitolo 6

Analisi dei Database di Ingredienti

6.1 Introduzione alle Analisi dei Database di Ingredienti

Questo capitolo si propone di confrontare quattro diversi database che catalogano ingredienti culinari (a titolo di esempio alcuni ingredienti come uova, pomodori e altri alimenti comuni). L'obiettivo principale di questa analisi è capire quale tra questi database sia il miglior candidato per estrarre successivamente informazioni dettagliate riguardanti il valore nutrizionale degli ingredienti, come zuccheri, fibre, grassi e altri valori. Il fine ultimo rimane l'integrazione di questi dati nell'applicativo del progetto per inserire la stima dei valori nutrizionali per ogni ricetta.

Analizzando questi database, si tenta di identificare ed esplorare le somiglianze e le differenze che esistono tra i dati riportati dai database in esame, riguardo gli ingredienti. Ad esempio, saranno esaminati alimenti comuni come il latte, le uova o i pomodori per illustrare il livello generale di dettaglio fornito da ciascun database e le informazioni che questi riportano.

Per ogni database sarà fornita una prima introduzione, seguita poi da una visione di insieme di come si presenta il database in risposta a ricerche di alcuni ingredienti comuni. Qui si riportano i risultati riguardanti gli ingredienti campione (vedi tabella 6.10).

Successivamente viene analizzato il database rispetto alle metriche definite successivamente nella sotto sezione 6.2.2.

Per alcuni database è presente anche una breve presentazione di come sia stato integrato a livello di codice, così da agevolare eventuali lavori futuri.

Alimenti
Latte di vacca
Uovo di Gallina
Pomodori
Mele
Arance
Salmone
Patate

Tabella 6.1: Campionario Utilizzato

In conclusione, questo capitolo presenta un'analisi comparativa di quattro database di ingredienti culinari, concentrandosi sulle analisi delle somiglianze tra gli ingredienti rappresentati, sul livello di dettaglio fornito, sui legami o bias culturali che questi database hanno etc. Lo scopo ultimo rimane comunque quello di determinare quale database sia il più consono per l'estrazione di informazioni dettagliate sugli ingredienti, in particolare per ingredienti diffusi nella tradizione culinaria italiana.

6.2 Situazione Database presenti online

Attualmente, esistono diversi database online che catalogano gli ingredienti culinari, ognuno con le proprie peculiarità e specializzazioni. Tuttavia, non esiste un unico database onnicomprensivo che si presenti come "tabella periodica degli ingredienti", contenente informazioni dettagliate su valori nutrizionali, proprietà chimiche e applicazioni culinarie. Alcuni database globali si concentrano principalmente sugli aspetti nutrizionali, come il USDA FoodData Central, mentre altri sono più orientati verso ricette e combinazioni di sapori, come Spoonacular o Edamam. Questa frammentazione rende difficile per gli utenti trovare un'unica fonte esaustiva e coerente di informazioni sugli ingredienti. La mancanza di standardizzazione e l'eterogeneità dei dati disponibili rappresentano una sfida per chi cerca informazioni complete e accurate per scopi sia professionali che amatoriali.

6.2.1 Database Candidati

Nel panorama italiano sono stati identificati 4 database principali che potessero essere dei buoni candidati per essere poi utilizzati nell'applicativo di questo progetto:

- CREA (Centro Alimenti e Nutrizione): Offre dati dettagliati sulla composizione degli alimenti maggiormente consumati in Italia, raccolti seguendo gli standard europei del progetto EuroFIR.
- FatSecret: Piattaforma globale che fornisce un database alimentare con informazioni su milioni di alimenti, supportata da tecnologie di riconoscimento degli alimenti e codici a barre.
- INRAN: Associato alla sezione scuola di Zanichelli e al libro "Scienza dell'alimentazione", offre una tabella PDF con dati sulla composizione degli alimenti, nonostante alcune limitazioni del sito.
- BDA (Banca Dati di Composizione degli Alimenti): Fornisce dati sulla composizione energetica e nutrizionale di 788 alimenti, fondamentale per studi epidemiologici sulla dieta e aggiornato regolarmente.

Questi database rappresentano le fonti trovate online, che risultano più complete, affidabili e pratiche per la raccolta di dati sugli ingredienti culinari, ciascuno con caratteristiche specifiche per diverse applicazioni.

6.2.2 Metriche di Analisi dei Database

Nel presente lavoro, i database culinari candidati saranno confrontati sulla base di specifiche metriche per valutare la loro qualità e la loro utilità rispetto alle nostre esigenze applicative. Le analisi dettagliate saranno esposte nei paragrafi e nelle sezioni a seguire.

I criteri di analisi includono:

- **Completezza rispetto agli ingredienti italiani**:
  questa metrica valuta la copertura e la rappresentazione degli ingredienti tipici della cucina italiana presenti nel database. Un database completo dovrebbe contenere una vasta gamma di ingredienti comunemente utilizzati in Italia.

  Per misurare ciò viene cercato un certo insieme di alimenti tipicamente italiani in ogni database e si riporta il numero di corrispondenze trovate.

  Il campionario utilizzato per questa analisi è il seguente:
  "Parmigiano Reggiano", "Gorgonzola", "Pecorino", "Ricotta", "Provolone", "Fontina", "Mascarpone", "Grana Padano", "Taleggio", "Pecorino Romano", "Burrata", "Stracchino", "Tiramisu", "Cannoli Siciliani", "Panettone", "Pandoro", "Savoiardi", "Taralli", "Cantucci"

  Questo campionario è sbilanciato sui formaggi e sui dolci e biscotti, poiché sono quegli ingredienti che più rispecchiano la cultura italiana, non essendo globali come altri ingredienti come i pomodori il cacao etc.

- **Livello di dettaglio per ingrediente**:
  questa metrica misura quanto dettagliatamente sono descritti gli ingredienti. Un elevato livello di dettaglio potrebbe includere informazioni su varietà, qualità, origine e utilizzi specifici degli ingredienti, nonché dettagli precisi e particolari riguardanti le proprietà nutritive degli ingredienti.

  Per misurare questa metrica si calcola il numero di elementi che riporta un singolo ingrediente, ovvero il numero di colonne che l'entrata della tabella possiede.

- **Accessibilità, licenza**:
  questa metrica analizza le condizioni d'uso del database, valutando se è liberamente accessibile, quali sono le restrizioni di utilizzo e quale sia il tipo di licenza sotto cui è distribuito.

- **Accessibilità tecnica**: questa metrica valuta invece quanto sia facile accedere e utilizzare il database da un punto di vista tecnico ed applicativo. Ciò viene fatto considerando il formato dei dati, la disponibilità di API, e la documentazione fornita.

- **Autorevolezza**:
  questa metrica esamina la fonte del database e la sua credibilità. database provenienti da fonti riconosciute e autorevoli sono considerati più affidabili. Inoltre è anche considerato il rigore metodologico con cui i dati sono stati misurati.

- **Prezzo**:
  questa metrica confronta il costo dei vari database, considerando se sono gratuiti o a pagamento, e valutando il rapporto qualità-prezzo offerto.

- **Dimensione del database**:
  questa metrica valuta la quantità di dati contenuti nel database, misurata in termini di numero di record, e quindi varietà di ingredienti. Un database di grandi dimensioni può offrire una maggiore varietà di informazioni, ma potrebbe richiedere più risorse per essere gestito e analizzato.

- **Aggiornamento**:
  questa metrica valuta la frequenza e quanto recentemente è avvenuto l'ultimo aggiornamento del database. Un database aggiornato di frequente è più probabile che contenga informazioni attuali e rilevanti, migliorando la sua utilità per analisi e applicazioni.

Queste analisi comparative ci permetteranno di identificare i punti di forza e le eventuali lacune di ciascun database, fornendo una guida chiara su quale database potrebbe essere più adatto per le esigenze di questo progetto.

6.3 Crea

6.3.1 Introduzione al Database

Il Database Italiano di Composizione degli Alimenti raccoglie i dati analitici e compilativi prodotti dal CREA, Centro Alimenti e Nutrizione. L'obiettivo è quello di fornire uno strumento completo e di alta qualità che permetta di accedere alla informazioni riguardanti la composizione degli alimenti maggiormente consumati in Italia.

Questo database nasce dal costante impegno del CREA nel raccogliere e pubblicare dati italiani aggiornati e affidabili, nonostante le notevoli risorse richieste da tale attività. Essendo pochi gli enti pubblici in grado di realizzare simili compiti, il CREA mira a rappresentare l'evoluzione continua dei prodotti alimentari presenti sul mercato nazionale.

I dati contenuti nel database provengono da un decennio di ricerche svolte presso il Centro Alimenti e Nutrizione, con una revisione totale delle informazioni precedentemente pubblicate. Le procedure di campionamento, analisi, documentazione e compilazione seguono gli standard elaborati con partner europei nell'ambito del progetto EuroFIR, di cui il Centro è membro associato.

L'attuale versione 2019 raccoglie i principali dati di composizione prodotti negli ultimi dieci anni, al fine di offrire uno strumento quanto più completo possibile. Per un utilizzo consapevole delle informazioni, si consiglia di consultare attentamente la sezione "Presentazione Dati" che guida gli utenti nell'impiego di questo database.

6.3.2 Un Campione del Database

Il database Crea ritorna una pagina web dedicata all'ingrediente ricercato. Questa pagina riporta numerose informazioni dettagliate.

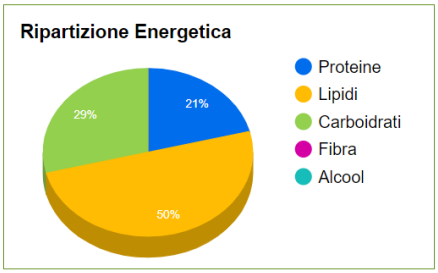
Per motivi di rappresentazione sono mostrati solo alcuni ingredienti del campionario, e parzialmente.

TABELLE DI COMPOSIZIONE DEGLI ALIMENTI

LATTE DI VACCA, PASTORIZZATO, INTERO	
Categoria	Latte e yogurt
Codice Alimento	135010
English Name	Milk, cow, whole
Informazioni	Un litro di latte pesa mediamente 1030 g
Parte Edibile	100 %
Porzione	125 g

Codice Langual

A0148 | A0719 | A0724 | A0740 | A0780 | B1201 | C0235 | E0123 | F0001 | G0003 | H0001 | J0001 | K0003 | M0001 | N0001 | P0024 | R0001



(a) Informazioni iniziali

Vedi tutti i campi

MACRO NUTRIENTI					
Descrizione Nutriente	Valore per 100 g	Valore per Porzione 125 g	Origine Dato	Metodiche	Referenze
Acqua (g)	87.0	108.8	A	Drying	i
Energia (kcal)	64	80	C	Energy calculated according to Southgate (kcal)	i
Energia (kJ)	268	334	C	Energy calculated according to Southgate (kJ)	i
Proteine (g)	3.3	4.1	A	Protein calculated from total nitrogen	i
Lipidi (g)	3.6	4.5	A	Solvent extraction	i
Colesterolo (mg)	11	14	A	Gas-liquid chromatography	i
Carboidrati disponibili (g)	4.9	6.1	C	Carbohydrate, available calculated from sugar and starch	i
Amido (g)	0	0	S	Imputation	
Zuccheri solubili (g)	4.9	6.1	A	Colorimetric method	i
Alcool (g)	0	0	ZL	Imputation	
Fibra totale (g)	0	0	ZL	Imputation	

(b) Informazioni sui Macronutrienti

Inoltre la pagina riporta una grande quantità di informazioni dettagliate riguardanti minerali, vitamine, acidi grassi e aminoacidi, in un formato equivalente a quello usato per i macronutrienti in figura 6.1b.

E' interessante notare che sia riportato anche il metodo di acquisizione dei dati, nonché le referenze per lo stesso.

6.3.3 Analisi rispetto alle metriche

- **Completezza rispetto agli ingredienti italiani**:
  Il database non riporta solo 5 ingredienti su 19 del campionario. Per tanto

Codice Alimento	**Nome Alimento ITA**
V	PARMIGIANO
V	GORGONZOLA
V	PECORINO
V	RICOTTA INTERA, TIPO TEDESCO
V	PROVOLONE
V	FONTINA
V	MASCARPONE
V	GRANA
V	TALEGGIO
V	PECORINO ROMANO
	BURRATA
V	STRACCHINO
	TIRAMISU'
V	CANNOLI SICILIANI
V	PANETTONE
	PANDORO
V	BISCOTTI SAVOIARDI
	TARALLI
	CANTUCCI

Tabella 6.2: Lista di Alimenti

  offre una buona rappresentazione anche degli ingredienti italiani.

- **Livello di dettaglio per ingrediente**:
  il livello di dettaglio riportato da Crea è molto elevato. Vengono riportate le informazioni base, i macronutrienti, le vitamine, i micronutrienti etc con grande precisione. Il tutto arriva a 75 dati per ingrediente.

Tabella 6.3: Informazioni Generali, Macronutrienti, Minerali e Vitamine riportati nella scheda

Informazioni Gen.	**Macronutrienti**	**Minerali**	**Vitamine**
Categoria	Acqua	Sodio	Tiamina
Codice Alimento	Energia (kcal)	Potassio	Riboflavina
English Name	Energia (kJ)	Calcio	Niacina
Informazioni	Proteine	Magnesio	Vitamina C
Parte Edibile	Lipidi	Fosforo	Folati
Porzione	Colesterolo	Ferro	Vitamina A ret. eq.
	Carboidrati disp.	Rame	Vitamina E
	Amido	Zinco	
	Zuccheri solubili	Selenio	
	Alcool		
	Fibra totale		

Tabella 6.4: Acidi Grassi, Aminoacidi, e Altri Composti riportati nella scheda

Acidi Grassi	**Aminoacidi**	**Altri Composti**
Acidi grassi Saturi	Lisina	Acido fitico
C4:0-C10:0	Istidina	
C12:0 acido laurico	Arginina	
C14:0 acido miristico	Acido aspartico	
C16:0 acido palmitico	Treonina	
C18:0 acido stearico	Serina	
C20:0 acido arachitico	Acido glutammico	
C22:0 acido beenico	Prolina	
Acidi grassi Monoinsaturi	Glicina	
C14:1 acido miristoleico	Alanina	
C16:1 acido palmitoleico	Cistina	
C18:1 acido oleico	Valina	
C20:1 acido eicosenoico	Metionina	
C22:1 acido erucico	Isoleucina	
Acidi grassi Polinsaturi	Leucina	
C18:2 acido linoleico	Tirosina	
C18:3 acido linolenico	Fenilalanina	
C20:4 acido arachidonico	Triptofano	
C20:5 acido eicosapentaenoico EPA	Indice Chimico	
C22:6 acido docosaesaenoico DHA	Aminoacido limitante	
Polinsaturi/Saturi		

- **Accessibilità, licenza**:
  Il database risulta consultabile principalmente online senza precise indicazioni di come ottenere una versione utilizzabile localmente del database stesso. Per quanto riguarda invece l'utilizzabilità, viene semplicemente richiesto di citare la fonte, rispettando i diritti di proprietà intellettuale.

- **Accessibilità tecnica**:
  Il database risulta scomodo da utilizzare tramite un applicativo poiché si pone come un database online. Sarebbe quindi necessario interfacciarsi con questo database tramite un API personalizzato o tramite degli strumenti web dedicati.

- **Autorevolezza**:
  Il Crea Centro di Ricerca Alimenti è una fonte autorevole per le informazioni alimentari in Italia, supportata da collaborazioni con il Ministero delle Politiche Agricole Alimentari e Forestali (MIPAAF), l'EFSA, la FAO, e l'Istituto Superiore di Sanità (ISS). SI sottolinea inoltre il rispetto degli standard del EuroFIR.

- **Prezzo**:
  Online è consultabile gratuitamente.

- **Dimensione del database**:
  il database contiene circa 900 alimenti suddivisi in 20 categorie alimentari.

- **Aggiornamento**:
  L'ultimo aggiornamento effettuato a Dicembre 2019.

6.4 FatSecret

6.4.1 Introduzione al Database

FatSecret è una piattaforma globale di monitoraggio nutrizionale e gestione della dieta che fornisce agli utenti strumenti per tenere traccia della loro alimentazione, attività fisica e progressi nel raggiungimento degli obiettivi di salute. Fondata nel 2007, FatSecret offre un database alimentare completo che include informazioni dettagliate su milioni di alimenti, marchi e ristoranti. La piattaforma permette agli utenti di registrare i loro pasti, esercizi fisici e dati biometrici, offrendo inoltre ricette, piani pasto e una comunità online per il supporto e la condivisione delle esperienze.

FatSecret utilizza la tecnologia di riconoscimento degli alimenti e codici a barre per semplificare l'inserimento dei dati alimentari, garantendo una facile e precisa gestione della dieta. Accessibile tramite sito web e app mobile, FatSecret è uno

strumento versatile per chiunque desideri migliorare la propria salute e forma fisica attraverso un monitoraggio consapevole e informato della propria alimentazione.

6.4.2 Un Campione del Database

Il database disponibile anche gratuitamente su richiesta tramite API, si presenta come un database sintetico, di facile uso, e ricco di ingredienti. Infatti per ogni ricerca effettuata, il database ritorna una lista di ingredienti che possono soddisfare la richiesta stessa. Nella tabella 6.5 sono riportati i primi elementi della lista. Questo permette più flessibilità rispetto ad errori o mancanza di specificità nella fase di ricerca, visto che per esempio, anche solo specificando "Milk" si possono ottenere anche informazioni su diversi tipi di latte.

Come si può notare dalla tabella stessa il database funziona esclusivamente in lingua inglese, sia negli input che negli output.

Food ID	**Food Name**	**Food Type**	**Food Description**	**Food URL**
794	Whole Milk	Generic	Per 100g - Calories: 60kcal | Fat: 3.25g | Carbs: 4.52g | Protein: 3.22g	fatsecret.com/.../milk-cows-fluid-whole
3092	Egg	Generic	Per 100g - Calories: 147kcal | Fat: 9.94g | Carbs: 0.77g | Protein: 12.58g	fatsecret.com/.../egg-whole-raw
6138	Tomatoes	Generic	Per 100g - Calories: 18kcal | Fat: 0.20g | Carbs: 3.92g | Protein: 0.88g	fatsecret.com/.../tomatoes-raw
35718	Apples	Generic	Per 100g - Calories: 52kcal | Fat: 0.17g | Carbs: 13.81g | Protein: 0.26g	fatsecret.com/.../apples
1350	Beef	Generic	Per 101g - Calories: 290kcal | Fat: 19.70g | Carbs: 0.00g | Protein: 26.55g	fatsecret.com/.../beef-cooked-ns-as-to-fat-eaten
4501	White Rice	Generic	Per 160g - Calories: 206kcal | Fat: 0.45g | Carbs: 44.50g | Protein: 4.24g	fatsecret.com/.../rice-white-cooked-regular
35878	Oranges	Generic	Per 100g - Calories: 47kcal | Fat: 0.12g | Carbs: 11.75g | Protein: 0.94g	fatsecret.com/.../oranges
2057	Salmon	Generic	Per 100g - Calories: 146kcal | Fat: 5.93g | Carbs: 0.00g | Protein: 21.62g	fatsecret.com/.../salmon-raw
5719	Potato (with or Without Peel)	Generic	Per 100g - Calories: 77kcal | Fat: 0.09g | Carbs: 17.47g | Protein: 2.02g	fatsecret.com/.../white-potato-raw-with-or-without-peel

Tabella 6.5: Informazioni nutrizionali di FatSecret per il campionario

E' anche disponibile un database più ricco e dettagliato, consultabile online e gratuitamente. I link ottenuti dal database con API rimandano alle rispettive pagine del database online, del quale è disponibile un campione in figura 6.2.

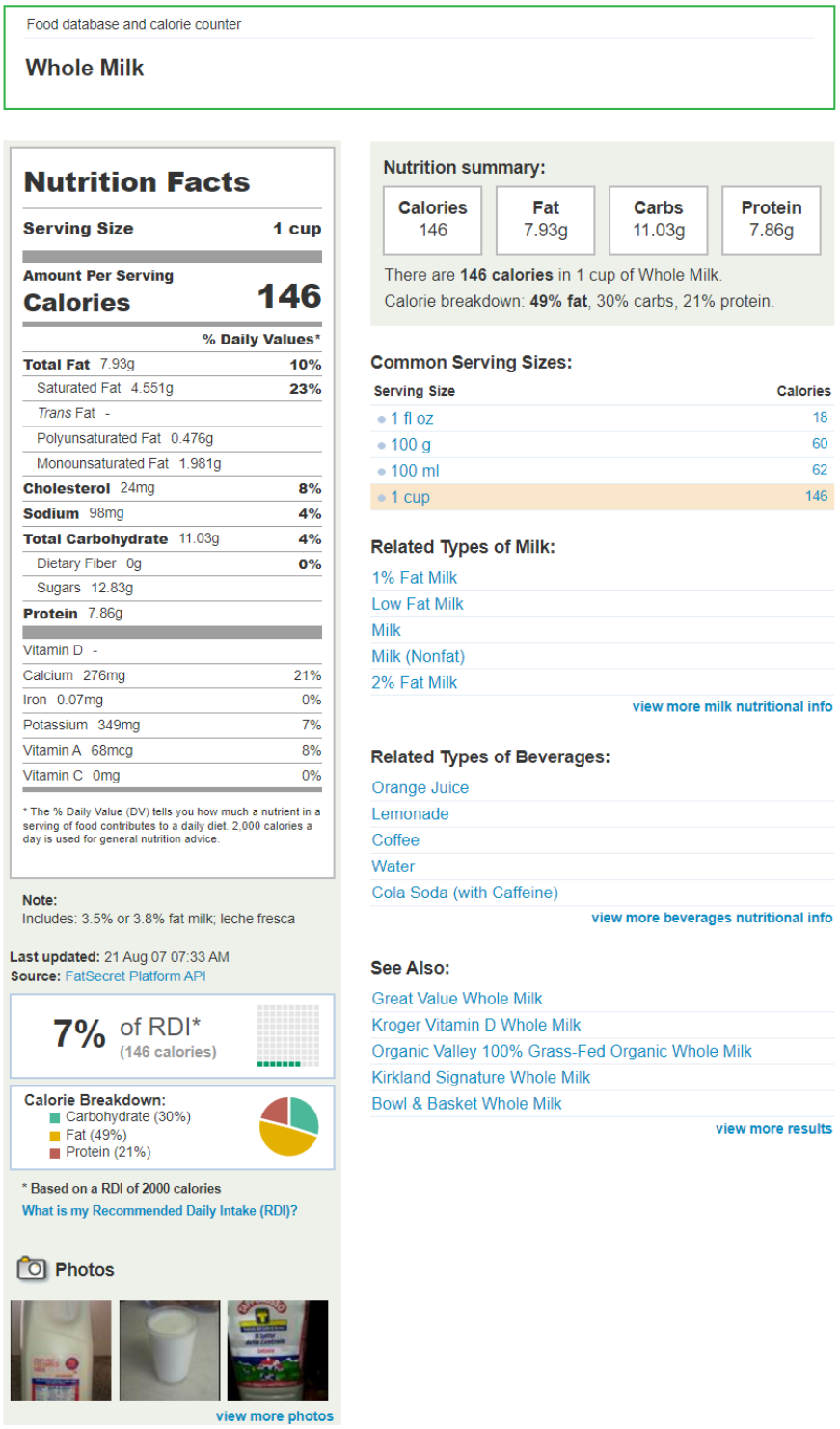


Figura 6.2: Vista online del database di Fatsecret

6.4.3 Analisi rispetto alle metriche

- **Completezza rispetto agli ingredienti italiani**:
  Il database è molto fornito anche rispetto ad ingredienti italiani, nonostante bisogna cercare l'ingrediente manualmente tra le alternative proposte. Non sono presenti solo 2 ingredienti su 19 del campionario. E' necessario sottolineare che questi ingredienti sono stati cercati senza la traduzione inglese, non essendo questa disponibile per questi nomi.

Presente	**Alimento**
V	PARMIGIANO
V	GORGONZOLA
V	PECORINO
V	RICOTTA
V	PROVOLONE
V	FONTINA
V	MASCARPONE
	GRANA
V	TALEGGIO
V	PECORINO ROMANO
V	BURRATA
V	STRACCHINO
V	TIRAMISU'
V	CANNOLI SICILIANI
V	PANETTONE
	PANDORO
V	BISCOTTI SAVOIARDI
V	TARALLI
V	CANTUCCI

Tabella 6.6: Presenza degli alimenti

- **Livello di dettaglio per ingrediente**:
  Ogni ingrediente viene rappresentato da poche informazioni basilari, molto sintetiche. Questi sono i dati se l'accesso al database viene fatto attraverso API gratuito. Utilizzando il sito di fatsecret (grazie anche al link disponibile tramite questo API) è invece possibile ottenere più informazioni riguardo all'ingrediente.

  Le infomazioni ottenute dall'API direttamente sono: Food ID, FoodName, Food Type, Food Description, e Food URL. Nella food Description sono riportate la porzione le calorie i grassi i carboidrati e le proteine.

Accedendo invece alla pagina online le informazioni sono in quantità maggiore. La seguente tabella 6.7 riporta le informazioni presenti online:

Categoria	**Informazione**
Porzioni	Dimensione della porzione, Dimensioni comuni delle porzioni
Informazioni caloriche	Calorie, Ripartizione delle calorie
Contenuto di grassi	Grassi totali, Grassi saturi, Grassi trans, Grassi polinsaturi, Grassi monoinsaturi
Colesterolo	Colesterolo
Sodio	Sodio
Carboidrati	Carboidrati totali, Fibra alimentare, Zuccheri
Proteine	Proteine
Vitamine e minerali	Vitamina D, Calcio, Ferro, Potassio, Vitamina A, Vitamina C
Informazioni aggiuntive	Nota, Ultimo aggiornamento, Fonte
Assunzione di riferimento	% Valori giornalieri AR (o AGR)
Immagini Alimento	Alcune foto

Tabella 6.7: Tipi di dati offerti per il latte intero

E' inoltre possibile accedere a più informazioni anche tramite API, utilizzando però piani a pagamento.

- **Accessibilità, licenza**:
  L'accessibilità al database Fatsecret è gratuita seppur limitata. Per ottenere le credenziali per accedere ed utilizzare l'API bisogna iscriversi e informare i gestori dei motivi che giusificano la richiesta.

- **Accessibilità tecnica**:
  L'utilizzo dell'API è facile ed intuitivo, nonché ben documentato. Un primo problema è la risposta del database in fase di ricerca. Essa è una lista di risultati e identificare quale sia quello corretto o più attinente all'input ricercato, richiede un intervento manuale per una resa ottimale. Inoltre si ricorda che il database è in lingua inglese e tutti gli ingredienti devono essere prima tradotti in fase di ricerca per ottenere risultati soddisfacenti.

- **Autorevolezza**:
  Fatsecret si pone come un compagnia multinazionale con un database utilizzato da diverse aziende come Amazon, Samsung, Fitbit o Medtronic. Inoltre l'applicazione e il suo successo indica che genera molta fiducia negli utenti. I dati sono anche verificati da un team di esperti del settore.

- **Prezzo**:
  Fatsecret, avendo come cuore della propria azienda queste informazioni nutrizionali, permette l'accesso gratuito al suo database con limitazioni. Sostanzialmente queste limitazioni riguardano il numero di chiamate effettuabili tramite API in certi intervalli temporali e le informazioni disponibili da consultare. Ciò che è mostrato in questa tesi riporta quanto è ottenibile gratuitamente. Con altri piani di acquisto Fatsecret rilassa questi limiti e permette l'accesso a dati alimentari più completi.
- **Dimensione del database**:
  Le dimensioni del database non sono rese pubbliche con precisione. Risulta però tra i database più ampi costruiti.
- **Aggiornamento**:
  I dati sono in continuo aggiornamento.

6.4.4 Implementazione del Database

L'obiettivo di questa prima implementazione è stato creare uno script che creasse automaticamente una tabella di link tra gli ingredienti del database di ricette e gli id di Fatsecret.

Tutti gli ingredienti distinti sono raccolti in una tabella, e tradotti utilizzando l'api di google traduttore per ottenere il corrispettivo nome inglese. Questo passaggio è necessario per interrogare il più propriamente possibile il database di Fatsecret.

L'interrogazione avviene con l'invio di un nome in inglese e la ricezione di una lista di ingredienti più rilevanti. Di questa lista viene preso il primo elemento, da cui si estrae il fat secret id ($fsid$) e le relative informazioni nutrizionali (nut_vals). Prendere il primo elemento della lista non è ottimale ma la lista di ingredienti sono ordinati per pertinenza, pertanto, volendo automatizzare il processo di collegamento, è necessaria una implementazione simile o comunque definita.

Una alternativa sarebbe progettare una funzione di similarità che stimi la similarità tra tutti gli elementi della lista ottenuta da FatSecret e la stringa utilizzata in fase di ricerca. Si potrebbe utilizzare una distanza di Levensthein ma considerando l'alta variabilità delle stringhe della lista di Fatsecret e anche il passaggio da italiano a inglese e viceversa, è ritenuta una funzione con scarsa resa.

Con queste nuove informazioni si crea quindi una nuova tabella che associ il nome italiano ai relativi valori nutrizionali (raccolti nella chiave nut_vals), da utilizzare in fase di recupero delle informazioni nutrizionali per ogni ricetta.

Il codice utilizza la libreria FatSecret per Python per interagire con l'API, autenticandosi tramite le chiavi cliente fornite.

Il cuore dell'integrazione è la funzione search, che cerca di mappare gli ingredienti del database locale con quelli presenti nel database di FatSecret. Questa funzione prima controlla se l'ingrediente è già presente in un dizionario locale (sdb), che funge da cache per evitare chiamate API inutili. Se l'ingrediente non è presente localmente, viene effettuata una ricerca tramite l'API di FatSecret. I risultati della ricerca vengono poi memorizzati in due strutture dati locali: una che associa il nome dell'ingrediente all'ID FatSecret (sdb search database) e un'altra che associa l'ID ai valori nutrizionali (ldb, local database). Questo approccio di caching migliora significativamente le prestazioni riducendo il numero di chiamate API necessarie.

Il codice include anche un sistema di gestione degli errori per gestire varie eccezioni che potrebbero verificarsi durante l'interazione con l'API, come timeout o errori di chiave. Queste funzioni di gestione degli errori registrano i problemi e implementano strategie di retry quando appropriato, garantendo la robustezza del processo di integrazione.

Inoltre, il codice implementa un sistema di checkpoint, salvando periodicamente lo stato del processo di integrazione. Ciò permette di riprendere il processo da dove si era interrotto in caso di interruzioni, rendendo l'integrazione resiliente a potenziali problemi di connettività o altri errori.

6.5 InRan

6.5.1 Introduzione al Database

Il database Inran risulta essere associato alla sezione scuola di Zanichelli e facente parte del libro scritto da Patrizia Cappelli e Vanna Vannucchi, "Scienza dell'alimentazione".

Sul sito si trova solamente un link per scaricare direttamente la tabella in formato pdf e un secondo link che rimanda al sito INRAN. Il sito sembra non essere particolarmente affidabile a primo impatto, poiché riporta principalmente diverse notizie che possono rientrare facilmente nella definizione di clickbait, nonostante i contenuti siano a prima vista di discreto valore.

Nonostante la fonte, non è stato possibile ottenere altre informazioni dal sito, pertanto non è stato ritenuto affidabile, nonostante la tabella si presenti abbastanza esplicativa e ben costruita.

6.5.2 Un Campione del Database

Il database riporta gli alimenti divisi per categorie. Ogni ingrediente viene riportato per nome e seguono, come specificazioni, altri dettagli che identificano più precisamente l'ingrediente.

132010	**Latte**, bufala	100	81,2	4,5	8,5	5,1	0	5,1	0	114	475				198	121						
130010	capra	100	86,3	3,9	4,8	4,7	0	4,7	0	76	320	40	180	0,1	141	106	0,05	0,11	0,30	86	1	
131010	pecora	100	82,	5,3	6,9	5,2	0	5,2	0	103	430	30	182	0,1	180	96	0,06	0,25	0,50	53		
135610	vacca, in polvere, intero	100	3,2	25,7	24,9	42,0	0	42,0	0	484	2027	440	1270	0,7	1050	728	0,24	1,31	0,70	295•	4	
135620	parzial. scremato	100	4,2	28,8	12,7	50,2	0	50,2	0	418	1748			0,7	1124	879	0,32	1,80		201	6	
135600	scremato	100	5,0	33,1	0,9	56,2	0	56,2	0	351	1470	550	1650	0,9	1323	1030	0,35	1,80	0,80	12	6	tr
135010	pastorizzato, intero[1]	100	87,0	3,3	3,6	4,9	0	4,9	0	64	268	50	150	0,1	119-	93	0,04	0,18	0,10	37•	1	0,07
135020	parzial. scremato	100	88,5	3,5	1,5	5,0	0	5,0	0	46	194			0,1	120	94	0,04	0,17	0,09	19	1	0,04
135030	scremato	100	90,5	3,6	0,2	5,3	0	5,3	0	36	151	52	150	0,1	125	97	0,04	0,17	0,09	tr	1	tr
135810	UHT, intero	100	87,6	3,3	3,6	4,7	0	4,7	0	63	265		150	0,2	120	95	0,04			50	1	0,07
135820	parzial. scremato	100	89,5	3,2	1,6	5,1	0	5,1	0	46	194			0,1	120	94					1	0,04
135510	condensato zuccherato	100	26,5	8,7	9,0	56,5	0	56,5	0	328	1371	130	390	0,2	293	235	0,09	0,43	0,20	99	3	
135520	evaporato non zuccherato	100	74,0	7,0	8,2	8,6	0	8,6	0	134	561	150	323	0,1	240	200	0,06	0,35	0,20	96	1	

(a) Latte

008000	**Arance**	80	87,2	0,7	0,2	7,8	0	7,8	1,6	34	142	3	200	0,2	49	22	0,06	0,05	0,20	71	50	
008010	**Arance**, succo	100	89,3	0,5	tr	8,2	0	8,2	0	33	137			0,2	15	17	0,05	0,03	0,40	38	44	tr

(b) Arance

007120	**Mele**, fresche con buccia	85	85,2	0,2	tr	10,0	0	10,0	2,6	38	161										
007110	senza buccia	79	82,5	0,3	0,1	13,7	tr	13,7	2,0	53	224	2	125	0,3	7	12	0,02	0,02	0,30	8	6	
007130	Annurca	84	84,3	0,3	0,1	10,2	tr	10,2	2,2	40	169										6	
007140	Deliziosa	79	86,8	0,2	0,1	11,1	tr	11,1	1,5	43	180										7	
007150	Golden	73	86,9	0,4	0,1	10,7	tr	10,7	1,7	43	180	tr	132	0,2	5						5	
007160	Granny Smith	81	87,8	0,5	0,1	9,4	tr	9,4	1,9	38	160										2	
007170	Imperatore	75	86,2	0,2	0,1	11,3	tr	11,3	1,9	44	184										7	
007180	Renetta	71	87,6	0,3	0,1	10,0	tr	10,0	1,7	40	166	tr	105	0,2		5					8	
007190	disidratate	100	2,7	1,3	2,0	92,2	tr	92,2	12,5	369	1544	7	730	2,0	40	66	tr	0,06	0,60		10	
007220	**Mele cotogne**	79	84,3	0,3	0,1	6,3	tr	6,3	5,9	26	108			0,1	4	14	0,02	0,03	0,70	0	14	

(c) Mele

006500	**Patate**, crude	83	78,5	2,1	1,0	17,9	15,9	0,4	1,6	85	354	7	570	0,6	10	54	0,10	0,04	2,50	3	15	
006511	arrosto	100	65,0	2,9	4,5	25,7	22,8	0,6	1,8	148	621	9	570	0,7	8	55	0,23	0,02	0,70	tr	8	
006505	bollite, con buccia	100	78,5	2,1	1,0	17,9	15,9	0,4	1,6	85	354	7	570	0,6	10	54	0,08	0,03	2,40	3	8	
006515	bollite senza buccia	100	80,1	1,8	0,1	16,9	15,0	0,4	1,3	71	299	7	280	0,4	5	31	0,08	0,01	0,50	tr	6	
006514	cotte al microonde	100	78,5	2,1	1,0	17,9	15,9	0,4	1,6	85	354	7	570	0,6	10	54	0,10	0,04	2,50	3	8	

(d) Patate

006600	**Pomodori**, da insalata	100	94,2	1,0	0,2	2,8	0	2,8	1,0	17	72	3	290	0,4	11	26	0,03	0,03	0,70	42	21	
006610	maturi	100	94,0	1,0	0,2	3,5	0	3,5	2,0	19	79	6	297	0,3	9	25	0,02		0,80	610	25	
006620	San Marzano	100	94,1	1,1	0,2	3,0	0	3,0		17	73	3	259	0,3	4					610	24	
006660	conserva	100	70,0	3,9	0,4	20,4	0	20,4	2,0	96	400		250	2,2	27	85	0,20	0,10	3,10		43	
006670	passata	100	90,8	1,3	0,2	3,0	0	3,0	1,5	18	76	160			16	35				530	8	
006680	pelati in scatola, frutti e succo	100	94,7	1,2	0,5	3,0	0	3,0	0,9	21	86	9	230	0,2	9	24	tr	tr	0,80	400	18	
006690	succo	100	93,8	0,8	tr	3,0	tr	3,0	0,6			230	230	0,4	10	19	0,02	0,02	0,70	200	2	

(e) Pomodori

122400	**Salmone**, fresco	65	68,0	18,4	12,0	1,0	0	1,0	0	185	776	98	310	0,7	27	280	0,20	0,15	7,00	13	tr	
122410	affumicato	100	64,9	25,4	4,5	1,2	0	1,2	0	147	613	1880	420	0,6	19	250	0,16	0,17	8,80	13	tr	
122430	in salamoia	98	64,3	21,1	11,5	1,0	0	1,0	0	192	802	540	300	0,8	66	286	0,05	0,19	7,00	60	0	

(f) Salmone

180010	**Uova**, anatra, intero	88	70,9	12,2	15,4	0,7	0	0,7	0	190	795	122	129	2,88	57	195	0,15	0,30	0,10	360•	0	
181100	gallina, intera[1]	87	77,1	12,4	8,7	tr	0	tr	0	128	535	137	133	1,5	48	210	0,09	0,30	0,10	225•	0	
181500	intero, congelato	100	77,1	12,4	8,7	tr	0	tr	0	128	535										0	
181105	cotto alla coque o sodo	100	77,1	12,4	8,7	tr	0	tr	0	128	535										0	
181600	in polvere	100	4,1	51,9	36,4	0,4	0	0,4	0	537	2246	573	556	6,3	201	879	0,38	1,26	0,42	942	0	
182010	albume	100	87,7	10,7	tr	tr	0	tr	0	43	180	179	135	0,1	7	15	0,02	0,27	0,10	0	0	0
183010	tuorlo	100	53,5	15,8	29,1	tr	0	tr	0	325	1360	43	90	4,9	116	586	0,27	0,35	0,10	640•	0	
183800	tuorlo congelato	100		15,8	29,1	tr	0	tr	0	325	1360										0	
185010	oca, intero	87	69,7	13,8	14,4	1,0	0	1,0	0	189	789			2,6	52	200	0,10	0,30	0,10	330•	0	
189010	tacchina, intero	87	72,6	12,8	10,2	1,0	0	1,0	0	147	614			2,5	50	200	0,15	0,30	0,10	225•	0	

(g) Uova

Figura 6.3: Ingredienti del campionario nel Database INRAN. Le colonne sono spiegate successivamente in Livello di Dettaglio.

6.5.3 Analisi rispetto alle metriche

- **Completezza rispetto agli ingredienti italiani**:

Presente	Alimento
V	PARMIGIANO
V	GORGONZOLA
V	PECORINO
V	RICOTTA
V	PROVOLONE
V	FONTINA
V	MASCARPONE
V	GRANA
V	TALEGGIO
V	PECORINO ROMANO
	BURRATA
V	STRACCHINO
	TIRAMISU'
V	CANNOLI SICILIANI
V	PANETTONE
	PANDORO
V	BISCOTTI SAVOIARDI
	TARALLI
	CANTUCCI

Questo database non riporta 5 ingredienti su 19 del campionario.

- **Livello di dettaglio per ingrediente**:
La tabella mostra la composizione chimica e il valore energetico degli alimenti per 100g di parte edibile. Include informazioni base per la gestione dei dati, la percentuale di parte edibile, contenuto di acqua, e anche i macronutrienti. Fornisce anche il valore energetico in kcal e kJ, e il contenuto di vari micronutrienti come minerali e vitamine. In totale riporta 23 dati per ingrediente

COMPOSIZIONE ALIMENTI (100 g)		
Informazioni Base	**Macronutrienti**	**Micronutrienti**
Numero codice	Acqua (g)	Sodio (mg)
Alimenti	Proteine (g)	Potassio (mg)
Parte Edibile (%)	Lipidi (g)	Ferro (mg)
Energia (kcal)	Carboidrati (g)	Calcio (mg)
Energia (kJ)	Amido (g)	Fosforo (mg)
	Zuccheri solubili (g)	Tiamina (mg)
	Fibra totale (g)	Riboflavina (mg)
		Niacina (mg)
		Vit. A ret. eq. (ug)
		Vit. C (mg)
		Vit. E (mg)

Tabella 6.9: Composizione chimica e valore energetico degli alimenti

- **Accessibilità, licenza**:
  Il database si presenta sottoforma di file PDF, contenente diverse pagine che descrivono i diversi alimenti. Non è indicato nessun limite esplicito di utilizzo di questo database, sebbene ciò non escluda che ne esistano.

- **Accessibilità tecnica**:
  Il formato PDF non è particolarmente maneggevole come file.

- **Autorevolezza**:
  Il sito che viene presentato con la tabella come sito INRAN risulta sospetto di essere un sito clickbait , privo di informazioni concrete riguardanti la chimica e i nutrienti di ingredienti o in generale privo di un approccio scientifico all'alimentazione. Inoltre non sono rinvenute fonti autorevoli o associazioni coinvolte nella verifica di questa tabella.

- **Prezzo**:
  La tabella PDF risulta gratuita online.

- **Dimensione del database**:
  Il database riporta circa 750 ingredienti diversi, suddivisi in 16 categorie.

- **Aggiornamento**:
  Non è espressamente indicata la data di creazione del database.

6.6 BDA

6.6.1 Introduzione al Database

Il progetto Banca Dati di Composizione degli Alimenti per Studi Epidemiologici in Italia (BDA) è nato nel 1998 grazie a un finanziamento dell'Associazione Italiana per la Ricerca sul Cancro, con la collaborazione di diversi gruppi di ricercatori italiani.

La prima pubblicazione, "Banca Dati di Composizione degli Alimenti per Studi Epidemiologici in Italia" ("Food Composition Database for Epidemiological Studies in Italy" by Gnagnarella P, Salvini S, Parpinel M. Version 1.2022 Website http://bda.ieo.it/), curata da Simonetta Salvini, Maria Parpinel, Patrizia Gnagnarella, Patrick Maisonneuve e Aida Turrini, ha risposto all'esigenza degli epidemiologi italiani di avere accesso a una banca dati ampia e dettagliata sui principali alimenti consumati in Italia. La BDA fornisce dati sulla composizione energetica e nutrizionale di 788 alimenti, includendo acqua, macronutrienti, minerali e vitamine, ed è basata su tabelle di composizione degli alimenti italiane e internazionali.

Questo strumento è cruciale per trasformare le informazioni dietetiche in nutrienti e per studiare il ruolo della dieta nella prevenzione e nella causa di malattie croniche. La BDA ha subito successivi aggiornamenti nel 2008, 2015 e 2022, ampliando e migliorando continuamente le informazioni fornite, e aderisce alle linee guida di EuroFIR AISBL, un network europeo di eccellenza per l'informazione sugli alimenti.

6.6.2 Un Campione del Database

All'interrogazione del database con la lista di ingredienti presi dal seguente campionario di test (opportunamente modificato per attenersi al meglio al database in fase di ricerca), lo stesso risponde fornendo come risultato le seguenti entrate riportate sinteticamente nella tabella successiva.

Dati Alimentari	
Simbolo	Nome Alimento ITA
1	Latte di vacca
2	Uovo di Gallina
3	Pomodori
4	Mele
5	Arance
6	Salmone
7	Patate

Tabella 6.10: Campionario Utilizzato per BDA

Dati Alimentari - Parte 1		
Simbolo	Codice Alimento	Nome Alimento ITA
2	805018_2	LATTE DI VACCA, INTERO, UHT
2	1800_2	UOVO DI GALLINA, INTERO
1	390_1	POMODORI, MATURI
3	420_2	MELE COTOGNE
3	404_2	ARANCE
2	700405_2	SALMONE
1	381_1	PATATE

Dati Alimentari - Parte 2			
Simbolo	Nome Scientifico	Fruttosio	Galattosio
2		0	0
2		0	0
1	Solanum lycopersicum	1.7	0
3	Pyrus cydonia	3.7	0
3	Citrus aurantium	2.2	0
2	Salmo salar	0	0
1	Solanum tuberosum	-3	-3

Dati Alimentari - Parte 3			
Simbolo	Saccarosio (MSE)	Maltosio (MSE)	Lattosio (MSE)
2	0	0	4.7
2	0	0	0
1	0.1	0	0
3	0.3	0	0
3	3.6	0	0
2	0	0	0
1	-3	-3	-3

Nella presente tabella molti dati sono stati omessi, per motivi di rappresentabilità dei dati, essendo 95 le colonne effettivamente presenti nel database per ogni ingrediente. Qui è riportato solamente un piccolo campione a titolo di esempio del database.

6.6.3 Analisi rispetto alle metriche

- **Completezza rispetto agli ingredienti italiani**:
  Il database BDA riporta i seguenti risultati se interrogato sul campionari di alimenti italiani:

Presente	**Nome Alimento ITA**
V	PARMIGIANO
V	GORGONZOLA
V	PECORINO
V	RICOTTA
V	PROVOLONE
V	FONTINA
V	MASCARPONE
V	GRANA
V	TALEGGIO
V	PECORINO ROMANO
V	BURRATA
V	STRACCHINO
V	TIRAMISU'
V	CANNOLI SICILIANI
V	PANETTONE
V	PANDORO
V	BISCOTTI SAVOIARDI
V	TARALLI
V	CANTUCCI

  Questa tabella riporta solo un piccolo spaccato della risposta completa fornita invece dal database BDA. Infatti tutte le informazioni nutrizionali sono state omesse per evidenziare il fatto che gli alimenti italiani fossero correttamente presenti nel database.

  Il database BDA contiene tutti gli ingredienti presenti nel campionario di ingredienti italiani.

- **Livello di dettaglio per ingrediente**:
  Il database BDA riporta ben 95 elementi per ogni ingrediente. Queste informazioni sono per la maggior parte informazioni chimiche e nutrizionali, unitamente al nome dell'ingrediente anche in inglese e un codice identificativo dell'alimento che permette anche di categorizzarlo in alcune categorie merceologiche identificate dallo stesso gruppo di ricerca.

Tabella 6.12: Dettaglio completo delle informazioni nutrizionali

Categoria	**Dati**
Informazioni base	Simbolo, Codice Alimento, Nome Alimento ITA, Nome Alimento ENG, Nome Scientifico, Categoria Merceologica, Parte edibile
Energia	Energia Ric con fibra (kcal), Energia Ric con fibra (kJ)
Macronutrienti	Proteine totali, Proteine animali, Proteine vegetali, Lipidi totali, Lipidi animali, Lipidi vegetali, Carboidrati disponibili (MSE), Amido (MSE), Carboidrati solubili (MSE), Fibra alimentare totale, Alcol, Acqua
Altri componenti	Colesterolo
Minerali	Ferro, Calcio, Sodio, Potassio, Fosforo, Zinco, Magnesio, Rame, Selenio, Cloro, Iodio, Manganese, Zolfo
Vitamine	Vitamina B1 (Tiamina), Vitamina B2 (Riboflavina), Vitamina C, Niacina, Vitamina B6, Folati totali, Acido pantotenico, Biotina, Vitamina B12, Retinolo equivalenti (RE), Retinolo, ß-carotene equivalenti, Vitamina E (ATE), Vitamina D, Vitamina K
Acidi grassi	Acidi grassi saturi totali, Acidi grassi monoinsaturi totali, Acidi grassi polinsaturi totali, Somma degli acidi butirrico (C4:0), caproico (C6:0), caprilico (C8:0) e caprico (C10:0), Acido laurico (C12:0), Acido miristico (C14:0), Acido palmitico (C16:0), Acido stearico (C18:0), Acido arachidico (C20:0), Acido beenico (C22:0), Acido miristoleico (C14:1), Acido palmitoleico (C16:1w-7), Acido oleico (C18:1w-9cis), Acidi eicosenoico (C20:1w-9), Acido erucico (C22:1w-9), Acido linoleico (C18:2w-6), Acido linolenico (C18:3w-3), Acido arachidonico (C20:4w-6), Acido eicosapentaenoico (EPA) (C20:5w-3), Acido docosaesaenoico (DHA) (C22:6w-3), Altri acidi grassi polinsaturi
Amminoacidi	Triptofano, Treonina, Isoleucina, Leucina, Lisina, Metionina, Cistina, Fenilalanina, Tirosina, Valina, Arginina, Istidina, Alanina, Acido aspartico, Acido glutammico, Glicina, Prolina, Serina
Zuccheri	Glucosio, Fruttosio, Galattosio, Saccarosio (MSE), Maltosio (MSE), Lattosio (MSE)

Come si può notare dalla tabella **??**riportata sopra, la maggior parte delle informazioni riportate dal database BDA quando interrogato riguardo un

ingrediente, sono di carattere chimico. Le informazioni sono molto tecniche e specifiche ed espolarano in profondità le proprietà del singolo ingrediente. Questa meticolosità nella analisi non stupisce se considerato il motivo per cui è nato questo database, ovvero per agevolare le analisi legate alla dieta per monitorare l'evoluzione di problematiche oncologiche ed epidemiologiche. Avere un tale livello di dettaglio è fondamentale per analisi e ricerche accurate in questi campi.

- **Accessibilità, licenza**:
  Il database BDA è consultabile online tramite un portale dedicato che permette la ricerca singola degli ingredienti in diverse modalità come per Nome, per codice, per categoria alimentatare, per nutriente o per ricerca alfabetica. Inoltre dopo la ricerca è possibile scaricare una scheda in formato pdf.

  Per consulatare liberamente il database privatamente o in locale è necessario effettuare una donazione alla associazione di ricerca tramite i canali e le modalità indicate sul sito stesso.

  A seguito di una donazione opportuna in funzione delle motivazioni per le quali si richiede il database è possibile ottenere un file xlsx, utilizzabile rispettando le norme imposte e definite dal gruppo in fase di donazione.

- **Accessibilità tecnica**:
  Una volta ottenuto il database in formato xlsx implementarlo nel proprio ambiente è facile. Nell'ambiente utilizzato in questo lavoro è utilizzata la libreria pandas unitamente alla libreria openpyxl per la lettura effettiva del file.

  Nel file è possibile accedere a diversi fogli. Uno è il database vero e proprio mentre un secondo riporta anche la categorizzazione merceologica effettuata per ogni elemento, attraverso la costruzione di un prefisso numerico presente nel codice del singolo ingrediente.

- **Autorevolezza**:
  La compilazione di questo database si basa su una metodologia rigorosa. Essendo una banca dati compilativa, i suoi dati di composizione sono stati raccolti da fonti preesistenti come tabelle, banche dati e articoli scientifici.

  Una particolare attenzione è stata posta nella selezione di alimenti rappresentativi della dieta nazionale italiana, con dati analitici riferiti a tali alimenti dove possibile.

  Nei casi in cui mancavano dati italiani, sono state utilizzate fonti straniere per completare le schede e fornire un profilo nutrizionale più esaustivo.

  Nel 2012, il processo di compilazione e gestione della BDA è stato sottoposto a revisione e certificazione da parte di EuroFIR (European Food Information

Resource Network), un ente di riferimento internazionale, allo scopo di verificare l'adesione a standard e procedure operative che assicurano la qualità e la validità dei dati generati a livello nazionale e internazionale.

Le metodologie utilizzate sono descritte dettagliatamente in due pubblicazioni, rispettivamente del 2022 e del 1998, garantendo trasparenza e autorevolezza alle fonti.

- **Prezzo**:
  Il prezzo non è fisso ma essendo stabilito tramite un sistema di donazioni minime in funzione dell'utilizzo previsto che si farà del database si possono porre dei minimi.

  Una donazione per utilizzo privato e non commerciale parte da 50 euro. Un utilizzo per altri scopi anche economici può richiedere donazioni a partire da 200 euro.

- **Dimensione del database**:
  Il database riporta 1110 entrate. Di queste alcune sono riferite allo stesso alimento, ma in forme diverse. Alcuni esempi potrebbero essere tonno sott'olio, o tonno al naturale o ancora tonno in scatola. Questi ingredienti ricoprono tre distinte entrate del database, nonostante sia aggregabili se si considera un livello di dettaglio minore.

- **Aggiornamento**:
  Il database BDA viene costantemente aggiornato per riflettere le evoluzioni delle scelte alimentari italiane e i cambiamenti nella composizione degli alimenti. Gli aggiornamenti avvengono regolarmente, con l'inclusione di nuovi dati o la correzione di errori riscontrati in alcuni valori.

  Gli aggiornamenti principali includono l'edizione del 2022, che ha ampliato le categorie di cereali, dolciumi, brioches e biscotti, aggiungendo nutrienti come aminoacidi, acidi grassi omega-3, minerali e vitamine. L'aggiornamento del 2015 ha integrato nuove categorie di frutta e aggiunto alimenti come gorgonzola con le noci, mentre quello del 2008 si è focalizzato su latte, carni, pesci, bevande, uova, e condimenti, migliorando la qualità dei dati grazie a potenziati controlli informatici.

  Lo stato attuale del database è quello dell'ultimo aggiornamento di Maggio 2023 che mira a correggere alcuni valori errati presenti in alcuni ingredienti.

6.6.4 Implementazione del Database

Per implementare il database è stata costruita una classe python apposita definita come IngredientDatabase. Questa classe mira a creare una classe di interfaccia,

personalizzabile per ogni database da implementare anche in futuro. La classe deve implementare alcune funzioni base per l'analisi e l'utilizzo del database. I metodi principali sono:

- **load:**
  la funzione di load permette di passare un generico file e caricare opportunamente i dati in esso contenuto a disposizione della classe.

- **search:**
  la funzione search permette di cercare una corrispondenza tra una stringa passata come argomento e una entrata nel database.

  Ritorna la corrispondenza ma non l'intera entrata. Questa funzione è utile se la ricerca diretta nel database deve essere mediata in qualche modo.

- **get_ingr_entry:**
  la funzione get_ingr_entry permette di ottenere una entrata del database a partire da una stringa di testo passata come parametro. A differenza della search, oltre a ritornare l'entrata del database, questa funzione necessita di più precisione nella ricerca per ottenere precisamente la'entrata desiderata.

  L'utilizzo ideale di questa funzione è insieme alla funzione search. Con la funzione search si ottiene il nome effettivo nel database dell'ingrediente cercato. Utilizzando questo nome effettivo come stringa di ingresso della funzione get_ingr_entry che permette di ottenere l'entrata desiderata in maniera più sicura.

- **get_ingr_list e get_info_list:**
  queste due funzione riportano rispettivamente una lista di stringhe che rappresenta la lista di ingredienti presenti nel database e una lista di stringhe che rappresenta i campi associati ad ogni ingrediente, ovvero i campi descrittivi.

- **link:**
  la funzione link permette di passare come parametro una tabella di ricette e collegare ogni ingrediente di ogni ricetta al rispettivo ID dell'ingrediente del database, così da ottenere una corrispondenza corretta tra ingredienti delle ricette ed entrate del database.

  In questo caso però non è stato creato un algoritmo di collegamento automatico ma si è utilizzata una tabella di collegamento creata manualmente.

  Il motivo risiede nella complessità di progettare un algoritmo che riesca a collegare coerentemente gli ingredienti corretti dei due database, utilizzando una analisi testuale standard.

6.7 Considerazioni Finali e Analisi Risultati

In questo capitolo abbiamo sviluppato una analisi comparativa di 4 database di ingredienti culinari che riportavano principalmente le loro caratteristiche nutrizionali.

Per effettuare questa analisi abbiamo utilizzato diverse metriche per stabilire l'idoneità dei database rispetto alle necessità del progetto.

In particolare abbiamo considerato l'attinenza agli ingredienti italiani, il numero di informazioni riportate per ingredienti, l'accessibilità effettiva ai dati e l'accessibilità tecnica, l'autorevolezza e l'affidabilità dei dati, nonché il prezzo, la dimensione in termine di ingredienti e la gestione degli aggiornamenti.

I risultati di questa analisi hanno rivelato per ciascun database i rispettivi punti di forza e di debolezza. CREA offre un alto livello di dettaglio una buona completezza e autorevolezza, ma un'accessibilità tecnica limitata. FATSECRET si pone come uno dei migliori database di ingredienti esistenti, vantando un'ottima completezza, discreto livello di dettaglio e aggiornamenti continui, ma con accesso gratuito limitato. INRAN, pur essendo facilmente accessibile, propone un database in formato PDF che manca però di autorevolezza e aggiornamenti recenti. BDA si è distinto per la sua copertura completa degli ingredienti italiani, i dettagli nutrizionali notevolmente estesi e il rigore scientifico, nonché una forte affidabilità sostenuta da diversi enti anche internazionali noti nel settore.

Metrica	CREA	FATSECRET	INRAN	BDA
Completezza ingredienti italiani	14/19	17/19	14/19	19/19
Livello di dettaglio (info per ingr.)	75	25	23	95
Accessibilità/Licenza	MEDIO	MEDIO	ALTO	MEDIO
Accessibilità tecnica	Online	API	File PDF	Online/File Excel
Autorevolezza	ALTO	ALTO	BASSO	ALTO
Prezzo	Gratuito	Gratuito*	Gratuito	Donazione 50€+
Dimensione database (n. ingr.)	900	N/D	750	1110
Aggiornamento	2019	Continuo	N/D	2023

Tabella 6.13: Valutazione comparativa dei database di ingredienti culinari. *FATSECRET: gratuito con limitazioni, piani a pagamento disponibili. N/D: Non Disponibile.

La tabella 6.13 riepiloga brevemente le valutazioni ottenute dai singoli database secondo le metriche definite.

Sulla base di questa analisi, è stato scelto il database BDA come l'opzione più adatta. Le ragioni principali di questa selezione includono la sua copertura completa degli ingredienti italiani, le informazioni nutrizionali eccezionalmente dettagliate (95 elementi per ingrediente), l'autorevolezza scientifica supportata da una metodologia rigorosa e da alcune certificazioni, come quella EuroFIR, e gli aggiornamenti regolari. Sebbene richieda una donazione per ottenere l'accesso completo, la profondità e la qualità dei dati giustificano questo investimento per analisi e ricerche nutrizionali serie. Inoltre la possibilità di utilizzare il database in formato xlsx anche in locale rende l'implementazione facile ed efficente se paragonata con l'uso di API.

In conclusione, questa analisi ha fornito una discreta panoramica di quello che è il contesto dei database di ingredienti culinari, in particolare per quel che riguarda la tradizione italiana. Sono inoltre evidenziati i punti di forza e i limiti di alcuni database culinari che hanno permesso di scegliere opportunamente il database più consono alle esigenze del progetto generale.

Capitolo 7

Conclusioni

L'obiettivo principale è stato quello di raffinare e migliorare notevolmente il dataset di ricette culinarie, con lo scopo di renderlo utilizzabile all'interno di un applicativo in grado di identificare una ricetta a partire da un'immagine, e stimare i relativi valori nutrizionali.

Per raggiungere questo obiettivo e rendere il dataset utile a livello pratico, è stato affrontato un primo problema legato alla qualità e all'organizzazione dei dati. Successivamente è stato selezionato ed integrato un database di ingredienti culinari, che riportasse informazioni nutrizionali sugli ingredienti stessi, e che fosse contemporaneamente affidabile e autorevole. Il tutto è da inserire all'interno di un contesto concentrato sulle ricette italiane.

A livello operazionale sono state effettuate le seguenti azioni:

- Implementazione e valutazione di una pipeline di preprocessing per i dati delle ricette
- Implementazioni di uno script di ricerca di Pseudo-duplicati tramite analisi degli ingredienti nelle ricette, unitamente ad alcune analisi su alcuni algoritmi di confronto usati e ordinamenti delle liste di ingredienti delle ricette
- Analisi comparativa di 4 database di ingredienti culinari rispetto ad alcune metriche, per stabilire quello più idoneo ad essere integrato nel progetto

7.1 Considerazioni Finali sui Risultati

In questa sezione sono raccolte le varie considerazioni fatte all'interno dei tre capitoli principali, ovvero l'effetto della pipeline di preprocessing, le prestazioni dello script di ricerca degli pseudo-duplicati e infine i risultati dell'analisi comparativa effettuata per identificare il database migliore per integrare le informazioni nutrizionali degli ingredienti.

7.1.1 Effetto della Pipeline di Preprocessing

La pipeline di preprocessing nasce con lo scopo di standardizzare, pulire e aggiungere nuove informazioni (come quantità o percentuale nella ricetta) ad ogni ingrediente presente all'interno delle ricette raccolte nel dataset di ricette culinarie.

Per valutare concretamente l'effetto della pipeline sono state effettuate due analisi. Una prima analisi sui dati grezzi e successivamente una seconda sui dati raffinati in uscita dalla pipeline. Queste analisi riguardavano in particolare il numero di parole presenti nei campi testuali degli ingredienti. Dai risultati ottenuti confrontando le due analisi effettuate si può affermare che:

- La pulizia e l'estrazione dei campi testuali degli ingredienti migliorano la qualità dei dati, andando a ridurre drasticamente la lunghezza dei campi testuali degli ingredienti (il numero di parole che descrivono l'ingrediente), generando quindi ingredienti con un campo testuale principalmente occupato dal nome dell'ingrediente stesso. Numericamente la media del numero di parole per ingrediente passa da 3.97 a 1.33.

- Le statistiche dei campi testuali, misurate in uscita dalla pipeline, sono dipendenti dai dati della tabella utilizzata per l'estrazione. In sostanza questa tabella è una struttura dati molto importante per la qualità dei risultati ottenuti in fase di estrazione del nome dell'ingrediente.

7.1.2 Ricerca di Pseudo-duplicati

Per migliorare ulteriormente la qualità del dataset, oltre a rimuovere i duplicati semplici come le ricette con lo stesso identico nome, è necessario rimuovere anche i cosiddetti pseudo-duplicati, ricette fondamentalmente identiche che però sono riportate con nomi differenti, o addirittura pochi ingredienti differenti.

Per rimuovere questi elementi superflui, si è deciso di confrontare tra di loro le liste degli ingredienti, per ogni coppia di ricetta possibile esistente nel database. Per ogni coppia si calcola un valore di similarità direttamente proporzionale al numero di ingredienti che le due ricette condividono, numero identificato tramite gli algoritmi di confronto. Per questa analisi sono stati utilizzati due tipi di algoritmi di confronto (SW e WSW), e due tipi di ordinamento delle liste degli ingredienti (Alfabetico e per Quantità Normalizzata).

Alla luce dei risultati si può affermare che:

- L'Ordinamento per Quantità Normalizzata risulta più efficace per migliorare le prestazioni della ricerca di pseudo-duplicati. Un motivo (lo stesso per cui questo ordinamento è stato testato) è probabilmente che questo ordinamento integra delle caratteristiche della ricetta nell'ordine degli ingredienti nella lista.

Infatti la percentuale che un ingrediente ricopre in una ricetta è un'informazione utile per identificarla. Due ricette possono condividere alcuni ingredienti, ma se questi sono in quantità nettamente diverse, difficilmente le ricette saranno la stessa.

- L'algoritmo WSW si dimostra generalmente migliore per quanto riguarda la precisione con soglie di similarità alte, ma generalmente cede il passo all'algortimo SW in termini di richiamo. Da questi risultati possiamo supporre che il WSW sia più preciso nel classificare gli pseudo-duplicati correttamente grazie alla funzione di peso, ma che questa lo renda troppo selettivo e di conseguenza ne infici la classificazione. Infatti la precisione alta è bilanciata dal basso numero di candidati individuati, che abbassano notevolmente anche il valore del richiamo.

7.1.3 Analisi Comparativa dei Database di Ingredienti

Una fase cruciale è stata la selezione del database di ingredienti culinari. Questo database ha un ruolo centrale all'interno dell'applicativo perché sarà la fonte di dati per il calcolo dei valori nutrizionali delle ricette.

Per scegliere quale database fosse più idoneo tra i candidati, è stata stipulata una lista di metriche, con le quali sono stati analizzati i database e successivamente confrontate.

Le metriche in questione sono

- Completezza per ingredienti italiani: utilizzano un campionario contenente ingredienti che rappresentassero al più la tradizione culinaria italiana, sono stati contati quanti tra gli ingredienti del campionario fossero presenti nel database
- Livello di dettaglio: corrisponde al numero di colonne per ogni ingrediente, in cui sono riportate delle informazioni ad esso associate.
- Accessibilità/Licenza: metodo e difficoltà di accesso al database vero e proprio, nonché limiti di utilizzo legali.
- Accessibilità tecnica: indica la difficoltà nell' integrare praticamente il database nel proprio progetto.
- Autorevolezza: indica quanto sia validato il database per utilizzi simili. Considera quindi certificazioni di enti, o grandi aziende che lo utilizzano.
- Prezzo: se è gratuito o a pagamento o se esistono diversi piani di acquisto.
- Dimensione: riporta il numero, se disponibile, di ingredienti contenuti

- Aggiornamento: come sono gestiti gli aggiornamenti, e in particolare qual è stato il più recente.

I database di ingredienti selezionati come candidati sono: CREA, FATSECRET, INRAN e BDA. Alla fine delle analisi effettuate secondo le metriche appena descritte i risultati ottenuti sono:

- **CREA**: alto livello di dettaglio, buona completezza, autorevolezza elevata (con certificazioni), ma accessibilità tecnica limitata.
- **FATSECRET**: ottima completezza, discreto dettaglio, aggiornamenti continui, ma accesso gratuito limitato (numero chiamate API e informazioni ridotte sugli ingredienti).
- **INRAN**: facilmente accessibile, in formato PDF, manca di autorevolezza nonché di aggiornamenti recenti.
- **BDA**: copertura completa degli ingredienti italiani, dettagli nutrizionali estesi, rigore scientifico, affidabilità, aggiornamenti regolari.

Al termine di questa analisi la scelta del database è stata in favore di BDA. I motivi principali che giustificano questa decisione sono la sua copertura completa sugli ingredienti italiani, l'elevatissimo livello di dettagli nutrizionali e infine una forte autorevolezza scientifica dimostrata da certificazioni europee.

7.2 Contributi Apportati

7.2.1 Codice Sviluppato

- Pipeline di preprocessing per la pulizia dei dati delle ricette: questa pipeline permette di migliorare notevolmente la rappresentazione di un ingrediente, trasformando la precedente descrizione testuale in diversi campi facilmente utilizzabili da altri script di codice. Questo è ottenuto con specifiche analisi testuali appositamente sviluppate per il contesto delle ricette culinarie italiane.
- Script per la ricerca di pseudo-duplicati nel dataset: la ricerca di pseudo duplicati tra le ricette si basa sull'analisi di ingredienti in comune, non nel senso assoluto, ma considerando, oltre al nome, anche la relativa differenza tra le percentuali di peso ricoperte dai due ingredienti confrontati.

7.2.2 Dati e File utili prodotti

- risultati dell'analisi comparativa dei principali database di ingredienti del panorama italiano

- Un file JSON contenente tutti i nomi dei nostri ingredienti, utilizzati per l'estrazione dell'ingrediente dal campo testuale, opportunamente categorizzati. Questa tabella può essere usata per ottenere la lista di ingredienti in codice, da utilizzare in fase di estrazione del nome dell'ingrediente vera e propria
- Un file JSON contenente le parole superflue utilizzate per la fase di pulizia precedente a quella d'estrazione del nome.
- Dataset di ricette pulite e raffinate: si passa da un dataset di circa 10800 ricette con ingredienti rappresentati come [name: "500 gr di farina 00 per la base"] a un dataset di circa 9900 ricette con ingredienti rappresentati come [name: "farina", quantity: 500, unit: 'gr', norm_on_100g: 39.7 ...].
- Database di ingredienti con relativi nutrienti (BDA).
- Un file Excel che raccoglie i dati in un formato pratico e consultabile. È diviso in tre file.
  - Il primo riporta ogni ricetta con la relativa lista di nostri ingredienti estratti dai campi testuali.
  - Il secondo riporta un file di collegamento e categorizzazione, costruito manualmente, che associa i nomi dei nostri ingredienti con i codici BDA corretti, e mantenga i nomi dei nostri ingredienti categorizzati secondo la nostra scelta delle categorie. Eventualmente è possibile accedere anche alla categorizzazione effettuata da parte del database BDA, utilizzando il codice BDA.
  - Il terzo file infine riporta invece ancora le ricette ma con il loro codice BDA ottenuto tramite il file di collegamento.

7.3 Migliorie e Sviluppi Futuri

Per quanto riguarda gli algoritmi di confronto, utilizzati nella ricerca di pseudo-duplicati ed in particolare il WSW, un primo punto interessante da analizzare in futuro potrebbe essere studiare un metodo di ottimizzazione della funzione di peso.

Infatti potrebbe aiutare sperimentare con diverse distribuzioni per ottenere più dati e poter così migliorare le prestazioni in scenari specifici.

Una prima idea potrebbe essere implementare una rete neurale ridotta che calibri in maniera ottimale i pesi per i casi specifici, ipoteticamente in maniera dipendente dalla singola coppia di lista.

Sarebbe interessante anche considerare altri campi applicativi e miglioramenti come ad esempio un ulteriore peso che diminuisce man mano che il centro della

finestra analizza elementi verso la fine della lista, implementazione particolarmente interessante con liste di dimensioni più elevate.

Infine in questo studio gli aspetti computazionali di questi algoritmi non sono stati considerati con particolare attenzione, e probabilmente esistono delle modifiche che permetterebbero di aumentarne l'efficienza.

Considerando invece la pipeline di preprocessing ed in particolare quelle fasi di estrazione di informazioni dai campi testuali, sarebbe interessante implementare l'estrazione delle informazioni principali mediante l'utilizzo di un LLM avanzato in grado di identificare correttamente l'ingrediente, la sua quantità e altri dettagli necessari.

Inoltre potrebbe sostituire, tramite prompt appropriati, molti processi qui effettuati manualmente o con processi di analisi testuali non "intelligenti". Ciò permetterebbe di analizzare quantità di dati enormemente maggiori in poco tempo.

Oltre alla mera estrazione delle informazioni base, un altro esempio potrebbe essere la gestione delle quantità particolari come 'qb' e 'gen'. In questo lavoro, le quantità 'qb' e 'gen' sono state parzialmente gestite tramite l'utilizzo di una tabella che ricopre almeno i 20 ingredienti più diffusi nel dataset di ricette. Per ogni ingrediente è stata effettuata una ricerca per stabilire un peso medio (nel caso di quantità 'gen') e un peso minimo (da utilizzare nel caso 'qb'). Questo processo, se correttamente implementato potrebbe essere fatto da un modello LLM su tutti gli ingredienti con quantità particolari, aumentando quindi la qualità dei dati delle ricette nel dataset.

Infine considerando l'intero lavoro svolto, alcune ottimizzazioni possono essere fatte in generale andando ad aumentare la precisione e il livello di dettaglio dedicato ad ogni fase del progetto. Come già citato, molte fasi hanno richiesto un intervento manuale, per costruire strutture di supporto (vedi tabella di estrazione nome ingrediente o tabella per la gestione delle quantità particolari, scelta di soglie) che oltre a richiedere molto tempo e attenzione, su moli di dati maggiori non sono applicabili. Rimane comunque spazio per ulteriori esperimenti e per una gestione ancora più meticolosa dei vari parametri sia all'interno della pipeline di preprocessing sia nello script per la ricerca degli pseudo-duplicati, ma anche nelle fasi di analisi testuale e dei nomi degli ingredienti.