POLITECNICO DI TORINO

Corso di Laurea in Ingegneria Informatica

Tesi di Laurea Magistrale

Studio e Sviluppo di un Sistema di Raccomandazione in Ambito Sanitario mediante Tecniche di Data Mining

Relatori:
Prof.ssa Silvia Chiusano
dott.ssa Elena Daraio

Candidato:
Lucia LAROCCA

Anno accademico 2018-2019

Sommario

Il seguente lavoro di tesi consiste nella progettazione e realizzazione di un sistema di raccomandazione di strutture riabilitative.

L'obiettivo è quello di fornire una piattaforma che permetta al paziente di fruire in modo più efficace e personalizzato delle strutture riabilitative presenti sul territorio.

Si è realizzato dunque un sistema che permette di personalizzare la scelta della struttura più idonea, date le caratteristiche dell'utente, le sue preferenze e la patologia da cui è affetto.

Per integrare il lavoro realizzato sul sistema, si è svolta un'analisi delle recensioni degli utenti, tramite tecniche di text mining, al fine di ampliare la conoscenza sulle strutture.

Indice

Elenco delle tabelle

Elenco delle figure

Capitolo 1

Introduzione

Nel nostro paese, il Sistema Sanitario Nazionale ha come obiettivo quello di creare una rete che sia in grado di seguire e supportare il paziente attraverso l'intero iter delle cure, creando una corretta continuità nel passaggio fra diverse figure professionali e fra diversi setting assistenziali.

Di fondamentale importanza è garantire la continuità nel delicato confine tra ospedale e altre strutture sanitarie, che tipicamente sancisce il passaggio del paziente dalla fase acuta della patologia alla fase di cure riabilitative, necessarie al massimo recupero possibile delle funzioni lese. Tale compito è svolto dal NOCC, Nucleo Ospedaliero di Continuità delle Cure, presente all'interno dell'azienda ospedaliera. Il NOCC, definito il piano terapeutico del paziente, collabora con il personale della struttura di degenza, il paziente ed i suoi famigliari, raccoglie le informazioni e definisce i bisogni assistenziali e sociali, individua la struttura di destinazione e guida il paziente durante l'iter fino alla dimissione.

Attualmente la scelta della struttura di destinazione più adatta è effettuata dall'operatore, esaminando la lista delle strutture riabilitative presenti sul territorio e delle relative caratteristiche, individuando quelle adatte alla cura del paziente.

L'obiettivo del lavoro di tesi è la creazione di una piattaforma di supporto al NOCC, in grado di automatizzare la ricerca della struttura di destinazione, ma anche di permettere all'utente di filtrare, dare priorità e accedere facilmente alle informazioni rilevanti.

Il sistema che possiede tale caratteristiche è un sistema di raccomandazione, che è in grado di generare contenuti personalizzati per gli utenti, tenendo conto delle loro necessità e preferenze. Da uno studio sullo stato dell'arte dei sistemi di raccomandazione, valutando aspetti positivi e negativi dei principali modelli ad oggi disponibili, si è individuato il modello *knowledge-based* come il più adeguato per il contesto oggetto di questa tesi. Tale modello genera una raccomandazione

individuando la presenza o meno di compatibilità tra le preferenze dell'utente e le caratteristiche degli articoli.

Il primo passo per la realizzazione del sistema di raccomandazione è stato quello di definire quali potessero essere gli input del sistema. Considerando che l'ambito di applicazione è quello sanitario, è stato necessario dividere le informazioni cliniche da quelle non cliniche, al fine di individuare i parametri su cui l'utente può esprimere una preferenza. Gli input del sistema sono stati divisi in *oggettivi* e *soggettivi*. Gli input oggettivi riguardano le attività riabilitative necessarie, quelli soggettivi invece riguardano le caratteristiche al contorno sulle quali l'utente può esprimere una preferenza. Dopo un'attenta analisi si sono scelti come input soggettivi di riferimento la distanza della struttura dalla posizione dell'utente, la valutazione che la struttura ha ricevuto e la presenza di servizi nelle vicinanze. A questi input è possibile dare un ordine d'importanza.

Le strutture di riabilitazione sono descritte nel sistema di raccomandazione attraverso diversi tipi di informazione. In particolare, la struttura è caratterizzata da: informazioni generali, quali nome e indirizzo; informazioni sulle attività riabilitative e specializzazioni; una valutazione su cinque punti rappresentativa delle recensioni espresse dagli utenti; il numero di servizi di ristorazione e parcheggi presenti nelle vicinanze considerando un raggio di seicento metri.

Il sistema realizzato genera una raccomandazione contenente una classifica delle strutture che rispettano gli input oggettivi e che soddisfano le desiderate dell'utente. La generazione di tale classifica è strutturata in due fasi. In una prima fase, dopo l'inserimento degli input oggettivi, il sistema elabora la lista di strutture idonee presenti sul territorio. Non essendo gli input soggettivi obbligatori, il sistema permette all'utente di esplorare la lista elaborata. In una seconda fase l'utente inserisce gli input soggettivi, ovvero informazioni sulla sua posizione e sugli indici di importanza che vuole attribuire alla distanza, alla valutazione e alla presenza di servizi. Ad ogni struttura viene associato: un punteggio per la sua vicinanza, un punteggio per il livello di valutazione ricevuto ed uno per la presenza o meno di servizi. I tre punteggi sono combinati per determinare la compatibilità della struttura e quindi la classifica finale.

Per arricchire le informazioni sulle strutture presenti nel sistema sviluppato si è eseguita un'analisi, mediante tecniche di *text mining*, sui commenti espressi dagli utenti, al fine di esaminarne la percezione ed individuarne le necessità.

Poiché i commenti testuali sono un tipo di dato non classificato e non etichettato è stato necessario l'utilizzo di tecniche di *unsupervised learning*, ovvero tecniche di *clustering* e *topic modeling*. A tal fine, i commenti sono stati pre-elaborati, eliminando le stop words e la punteggiatura, il testo è stato riportato tutto in minuscolo ed è stata applicata la *lemmatization*, che consiste nel riportare la parola alla sua forma base.

Dopo la fase di pre-elaborazione, si è individuata una rappresentazione vettoriale del testo e si è determinato lo spazio semantico di tale rappresentazione. Lo spazio semantico richiede la definizione di similarità tra i documenti in modo tale che documenti con significati simili siano più vicini tra loro e quelli diversi siano più distanti; la metrica di similarità utilizzata è la distanza euclidea. Dopo aver analizzato alcuni algoritmi non supervisionati, si è scelto di utilizzare il *k-means clustering*, che permette di raggruppare i documenti in cluster, e il *topic modeling*, che mira ad astrarre dai documenti temi fondamentali.

Dai risultati ottenuti con il clustering si è notato che i commenti possono essere raggruppati secondo tre tematiche: rapporto paziente-personale, rapporto del paziente con la malattia e opinioni in generale sulla struttura. I risultati ottenuti tramite tecniche di topic modelling confermano la presenza di temi quali: struttura, pazienti e personale. Data la stretta correlazione tra i temi anche i termini all'interno del topic sono correlati.

Il documento di tesi è stato diviso in due parti: nella prima parte si è progettato e realizzato il sistema di raccomandazione, mentre nella seconda parte si è svolta l'analisi delle recensioni degli utenti utilizzando tecniche di text mining. In particolare, nel capitolo 2 viene descritto il contesto di applicazione studiando il problema della continuità delle cure nel Sistema Sanitario Nazionale. Nel capitolo 3 si è esaminato lo stato dell'arte dei sistemi di raccomandazione ad oggi disponibili. Individuato il modello più adatto per il sistema oggetto di questa tesi, nel capitolo 4 si è definito con maggiore chiarezza la progettazione e la realizzazione dell'architettura di tale sistema, riportando i risultati ottenuti.

Nella seconda fase del lavoro, nel capitolo 5, si è esaminato il contesto di text mining e le principali tecniche disponibili, nel capitolo 6 tali tecniche sono stata applicate alle recensioni degli utenti e si sono analizzati i risultati ottenuti.

In fine, nel capitolo 8 si sono presentate le conclusioni e gli sviluppi futuri del sistema realizzato.

Capitolo 2

Analisi del Problema

Nel Sistema Sanitario Nazionale, garantire l'unitarietà del sistema sanitorio è uno dei principali obiettivi, a tal fine: il Governo e le Regioni concordano e redigono "Il Patto per la Salute" [16], un accordo finanziario e programmatico, di valenza triennale, in cui uno degli intenti principali è quello di assicurare la continuità delle cure.

Integrazione è la parola chiave su cui si basa l'evoluzione del sistema sanitario proposta nel Patto. Integrazione che è necessaria creare fra le diverse figure professionali e anche fra i diversi setting assistenziali, per dare forma ad una rete che sia in grado di seguire e supportare il paziente attraverso l'intero iter delle cure. È fondamentale porre la persona/paziente come centro del sistema e per fare ciò è necessario che le strutture sanitarie dialoghino, condividano informazioni e sia presente una corretta continuità anche nel passaggio fra operatori appartenenti a diverse realtà. La continuità delle cure, quindi, è intesa sia come continuità tra i diversi professionisti integrati in un quadro unitario (lavoro in team, elaborazione e implementazione di percorsi diagnostico-terapeutici condivisi ecc.), sia come continuità tra i diversi livelli di assistenza, soprattutto nel delicato confine tra ospedale e territorio.

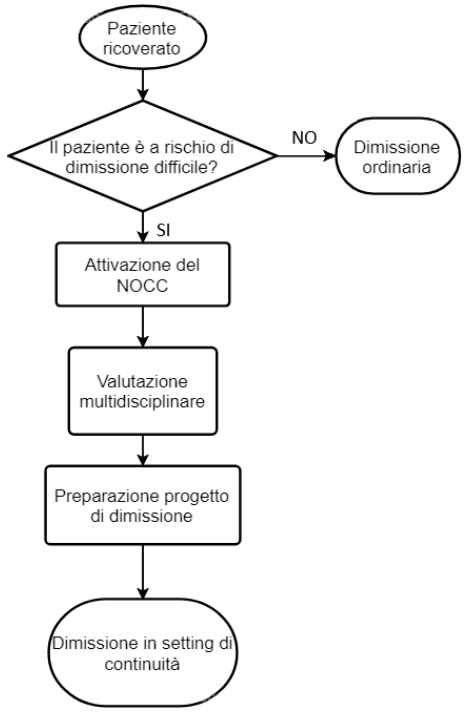
La Regione Piemonte, con la Delibera della Giunta n. 27-3628 del 28 marzo 2012 [23], ha approvato un percorso integrato di continuità delle cure ospedale-territorio all'interno della rete dei servizi, per interventi di tipo sanitario e socio-assistenziale. Ciò ha portato all'istituzione, presso tutti i Presidi Ospedalieri, dei Nuclei Ospedalieri di Continuità delle Cure (NOCC). I NOCC hanno il compito di consolidare le relazioni tra le Aziende Ospedaliere e il Distretto, garantendo il passaggio delle informazioni cliniche, terapeutiche e socio-assistenziali ai vari punti della rete e permettendo la continuità comunicativa e prestazionale in tempo reale tra i vari setting. Successivamente c'è stato l'adeguamento della rete ospedaliera regionale piemontese agli standard del Patto per la Salute 2014/2016.

2.1 Nucleo Ospedaliero di Continuità delle Cure

Gli obiettivi del NOCC sono quelli di proporre il setting di dimissione più appropriato e più sostenibile dal punto di vista economico, sociale, ambientale e familiare e migliorare la *compliance* del paziente al progetto di cure, al fine di aumentarne la soddisfazione e favorirne *l'empowerment*.

Durante il ricovero di un paziente bisogna valutare se alla sua dimissione possono essere presenti dei bisogni sanitari o sociali. Nel caso di bisogni sanitari è presente la necessità di costruire un progetto di cura ed assistenza, di dimissioni protette e di percorsi socio-sanitari per patologie cronico-complesse. Nel caso di bisogni sociali sono presenti indicatori di fragilità della persona (disabilità, patologie dell'età anziana, ecc), del nucleo familiare o della condizione abitativa. Tutti i pazienti che presentano i bisogni sopra descritti, necessitano di continuità delle cure. Tale necessità deve essere segnalata dal personale sanitario al NOCC della struttura di degenza. Uno schema dell'iter per l'attivazione del NOCC è riportato in Figura 2.1.

Figura 2.1: Iter di dimissione



Una prima valutazione, sulla base dei dati clinici, assistenziali e sociali, identifica i bisogni terapeutici, assistenziali e di supporto al paziente e alla famiglia per

assicurare il miglior livello di continuità delle cure nel setting post-ospedaliero che verrà individuato. La valutazione utilizza scale multidimensionali, cliniche ed assistenziali che permettono di costruire indicatori da utilizzare per la proposta, da parte del medico di riferimento, del setting più appropriato di dimissione (CAVS, post-acuzie, cure domiciliari, cure palliative, progetti specifici). I risultati della valutazione multidisciplinare vengono inseriti nel progetto di dimissione.

Il NOCC collabora con il personale della struttura di degenza per l'individuazione di ausili e presidi necessari ed indispensabili per la dimissione. Fino al momento della dimissione il NOCC cura i contatti con il territorio in tutte le componenti coinvolte, monitora l'andamento del ricovero e registra l'avvenuta dimissione disposta dal Medico del reparto, operando affinché siano superate tutte le eventuali criticità.

Nel seguente lavoro di tesi si è analizzato il caso in cui nel setting di dimissione di un paziente sia indicato un trasferimento in post-acuzie, in particolare in un setting di lungodegenza o riabilitazione, che richiede l'elaborazione di un progetto di piano riabilitativo.

2.2 Progetto di Piano Riabilitativo

Un efficace sviluppo del percorso riabilitativo richiede la definizione di modalità organizzative che prevedano la gestione della riabilitazione intra e/o interaziendale. Per le strutture sanitarie è obbligatorio la presa in carico del paziente mediante la predisposizione di un progetto riabilitativo individuale, in cui il medico specialista in medicina fisica e di riabilitazione indica l'attività riabilitativa necessaria al paziente.

Le attività di riabilitazione, così come indicate nella Delibera della Giunta Regionale 2 aprile 2007, n. 10-5605 [22], sono:

- *Attività riabilitativa di primo livello.* Tali attività interessano pazienti con disabilità di entità rilevante, nell'immediata post-acuzie, croniche o in fase di stabilizzazione, che richiedono un intervento riabilitativo non complesso, né intensivo, ma protratto nel tempo; nonché pazienti con disabilità croniche stabilizzate di entità contenuta per le quali possono essere necessari interventi riabilitativi di mantenimento o di prevenzione del degrado motorio-funzionale acquisito.

- *Attività riabilitativa di secondo livello.* Sono quelle dirette al recupero di disabilità importanti, modificabili che richiedono un elevato impegno diagnostico medico specialistico ad indirizzo riabilitativo e terapeutico, in termini di precocità, complessità e/o di durata dell'intervento.

- *Attività riabilitativa di terzo livello.* Sono attività di riabilitazione intensive, ad alta specializzazione, mezzi, attrezzature e personale e sono erogate presso ospedali e sedi di alta specializzazione. Tali attività comprendono Unità

Spinale Unipolare (U.S.U.) e Unità per le Gravi Cerebrolesioni acquisite (U.G.C.).

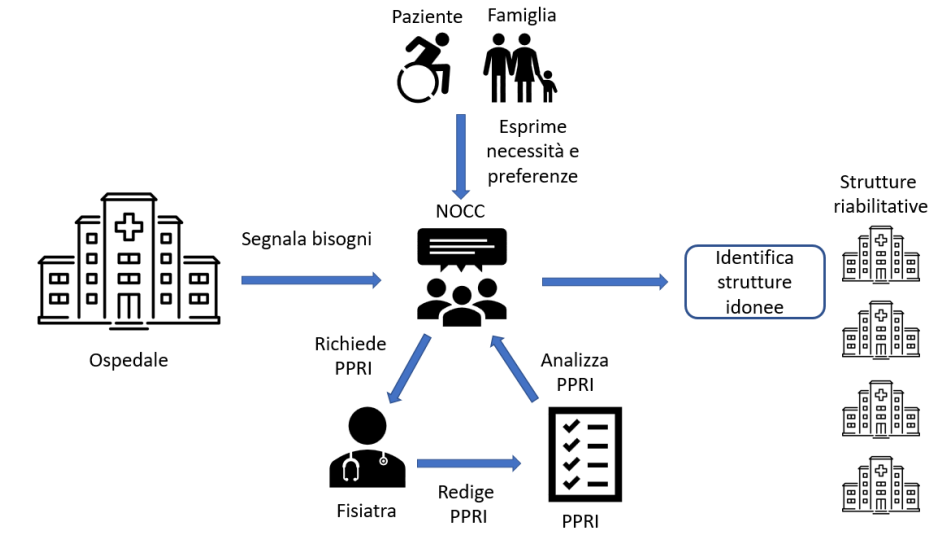
- *Lungodegenza.* Tali attività sono destinati a pazienti di ogni età che, superata la fase acuta della malattia, necessitano ancora, per un periodo protratto di tempo, di cure e trattamenti intensivi appropriati, possibili in ambito ospedaliero, atti a superare o stabilizzare le limitazioni all'autosufficienza derivanti da malattie e/o infortuni.

2.3 Soluzione proposta

Individuati gli interventi riabilitatavi necessari al paziente, indicati nel progetto di piano riabilitativo è necessario individuare la struttura di destinazione più idonea. Il NOCC, collaborando con il personale della struttura, il paziente ed i suoi famigliari, raccoglie le informazioni, stabilisce i bisogni assistenziali e sociali del paziente, individua la struttura riabilitativa di destinazione e guida il paziente durante la fase della dimissione.

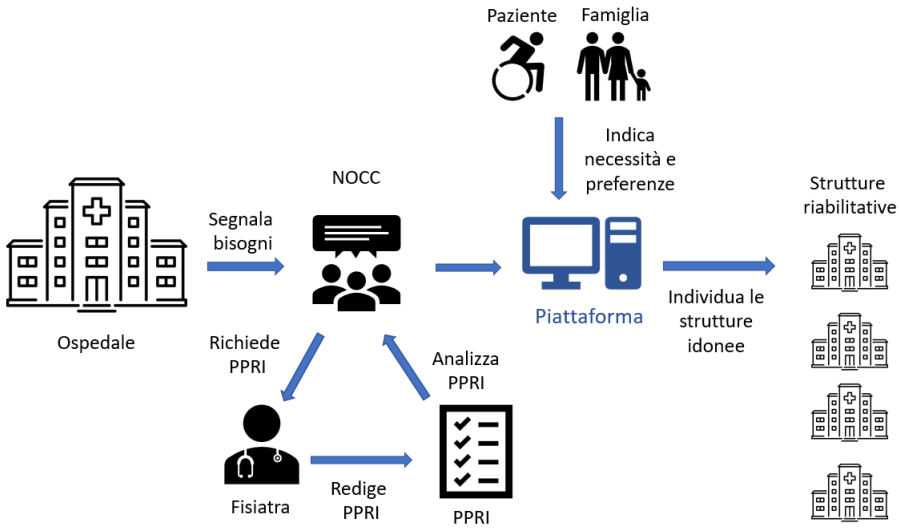
Attualmente la scelta della struttura di destinazione più adatta è effettuata dall'operatore, esaminando la lista delle strutture riabilitative, delle relative caratteristiche ed individuando quelle adatte alla cura del paziente. L'iter eseguito dal NOCC è rappresentato in Figura 2.2.

Figura 2.2: Identificazione strutture idonee



L'obiettivo del seguente lavoro di tesi è la creazione di una piattaforma di supporto al NOCC e al paziente, in grado di ampliare e automatizzare la ricerca della struttura di destinazione, come riportato in Figura 2.3.

Figura 2.3: Identificazione strutture idonee utilizzando la piattaforma



La piattaforma espone le strutture sanitarie disponibili e consente al paziente di fruire in modo più efficace delle strutture riabilitative. In particolare, la piattaforma permette di personalizzare la scelta della struttura in maniera più adeguata alle caratteristiche del paziente, sfruttando un sistema di analisi delle preferenze degli utenti, delle loro esigenze e delle caratteristiche della patologia di cui sono affetti, correlando queste informazioni con le caratteristiche delle strutture riabilitative. La piattaforma è in grado di gestire e correlare diverse sorgenti di informazioni e generare la migliore allocazione possibile dei pazienti nelle strutture.

Capitolo 3

Sistemi di Raccomandazione

3.1 Introduzione ai Sistemi di Raccomandazione

Nel mondo di Internet, dove il numero di informazioni cresce in maniera vertiginosa, è necessario filtrare, dare priorità e rendere facilmente fruibili informazioni rilevanti in modo da guidare l'utente nella navigazione. I sistemi di raccomandazione risolvono questo problema ricercando in un ampio volume di informazioni, generato dinamicamente, contenuti personalizzati per gli utenti, tenendo conto delle loro preferenze, dei loro interessi e comportamenti. I sistemi di raccomandazione possono essere utilizzati per realizzare una predizione oppure per realizzare una classifica. Nel primo caso tipicamente lo scopo è predire la valutazione che l'utente assegnerebbe ad un dato articolo, nel secondo caso lo scopo è realizzare una classifica dei contenuti più d'interesse per l'utente. L'obiettivo finale è comunque quello di portare all'attenzione dell'utente articoli o contenuti rilevanti e di effettivo interesse.

Per valutare come ottima una raccomandazione essa deve rispettare quattro caratteristiche fondamentali [1]:

- rilevanza, i contenuti devono essere effettivamente rilevanti per l'utente;
- novità, i contenuti devono essere qualcosa che l'utente fino a quel momento non ha visto;
- serendipità, i contenuti devono essere inaspettati, diversi da ciò a cui l'utente è normalmente abituato e quindi portare un elemento di scoperta fortunata, di sorpresa;
- diversità delle raccomandazioni, sopratutto nel caso di realizzazione di una classifica; se gli articoli sono tutti simili tra di loro aumenta il rischio che l'utente non scelga nessuno di essi.

I sistemi di raccomandazione sono divisi in tre categorie principali: *collaborative filtering*, *content-based* e *knowledge-based*, di seguito analizzati nel dettaglio. L'evoluzione dei sistemi di raccomandazione ha messo in evidenza come in alcuni casi può essere opportuno considerare una soluzione ibrida che prenda e unisca i vantaggi delle diverse tecniche. Nelle figure 3.1, 3.2 e 3.3 sono rappresentate le tre tipologie e le principali differenze.

Figura 3.1: Tipologie di sistemi di raccomandazione

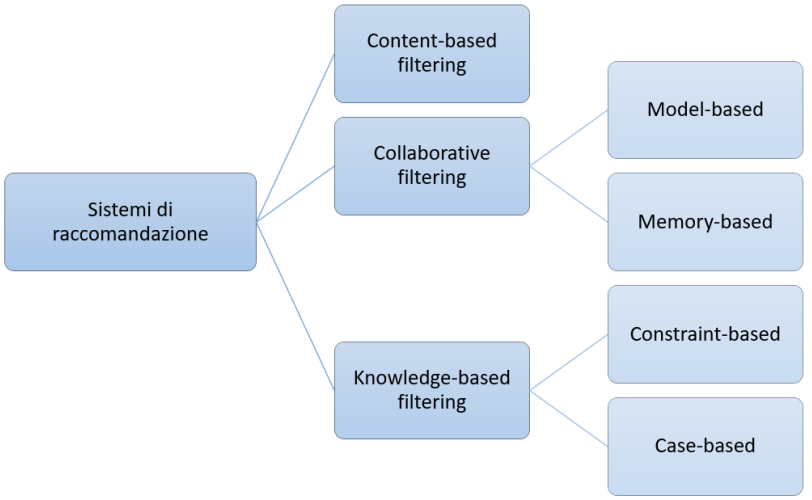


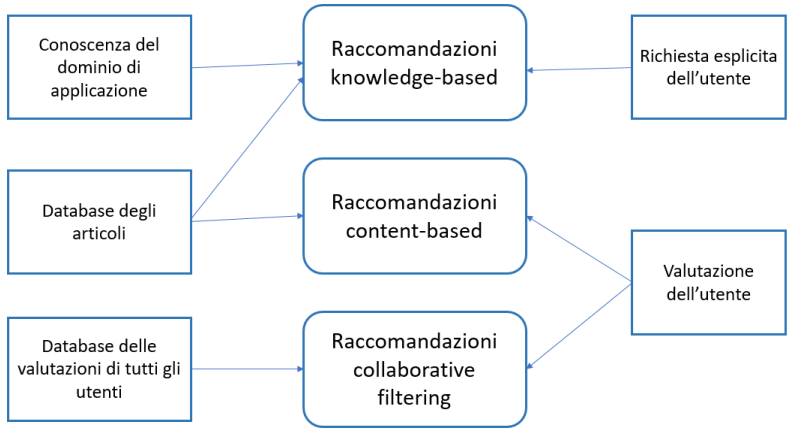
Figura 3.2: Differenze tra le tipologie di sistemi di raccomandazione

Approccio	Obiettivi	Input
Collaborative	La raccomandazione è generata con un approccio collaborativo che utilizza le valutazioni e le azioni di utenti simili.	Valutazioni dell'utente + valutazioni di tutti gli altri utenti
Content-based	La raccomandazione è generata in base al contenuto degli articoli valutati in precedenza dall'utente	Valutazioni dell'utente + contenuto dell'articolo
Knowledge-based	La raccomandazione è generata in base ai vincoli su caratteristiche dell'articolo indicate dall'utente.	Specifiche dell'utente + conoscenza del dominio + contenuto dell'articolo

3.1.1 Raccomandazioni Collaborative

I sistemi di raccomandazione *collaborative* sfruttano il meccanismo con il quale gli esseri umani prendono decisioni, ossia attraverso le proprie esperienze personali ma

Figura 3.3: Similitudini tra le tipologie di sistemi di raccomandazione



anche e soprattutto le esperienze di conoscenti.

La raccomandazione è effettuata individuando utenti con gusti simili all'utente target e utilizzando le valutazioni rilasciate da tali utenti per predire quale sarebbe la valutazione che l'utente target assegnerebbe all'articolo. Esistono due categorie di metodi utilizzati, che si differenziano in base all'algoritmo utilizzato per esplorare le connessioni tra gli utenti [13]: *memory-based* e *model-based*.

I metodi *memory-based* si possono suddividere in due ulteriori categorie: *user-based* e *item-based*. Nel metodo user-based la similarità tra gli utenti è determinata confrontando le valutazioni sugli stessi articoli. La media pesata delle valutazioni dei k utenti più simili, in cui i pesi sono dati dal livello di similarità di questi ultimi rispetto all'utente target, viene usata per predire la valutazione. Nel metodo item-based si calcola la predizione usando la similarità tra gli articoli e non tra gli utenti. Si identifica un set S di articoli che sono simili all'articolo target, di interesse per l'utente, la predizione è calcolata come media pesata delle valutazioni dei k articoli più simili, in cui i pesi sono dati dal livello di similarità di questi ultimi rispetto all'articolo target.

Nei metodi *model-based*, si utilizzano tecniche di machine learning e data mining nel contesto di modelli predittivi.

Il vantaggio dei sistemi di raccomandazione *collaborative-filtering* è che la tecnica di predizione è indipendente dal dominio di applicazione, poiché non si basano sulla descrizione dell'articolo. Per tale motivo questa soluzione è molto utilizzata nel caso di contenuti molto difficili da descrivere oppure difficili da analizzare per un sistema informatico, come ad esempio musica o film. Uno degli svantaggi è la sua poca efficienza nel caso di articoli aggiunti di recenti, per i quali non sono presenti molte valutazioni, ma è molto efficace nel generare raccomandazioni per nuovi

utenti. Inoltre questa tecnica risente del problema del *cold-start* e della scarsità dei dati. Il problema del cold-start si riferisce alla situazione in cui il sistema di raccomandazione non ha abbastanza informazioni sull'utente o sugli articoli per generare predizioni, sopratutto all'inizio [6]. Per scarsità dei dati si intende invece il caso in cui le valutazioni degli utenti siano poche.

3.1.2 Raccomandazioni Content-based

I sistemi di raccomandazione *content-based* generano una predizione tenendo conto delle informazioni e delle caratteristiche dell'utente ed analizzando gli attributi descrittivi dell'articolo. Quindi, il contributo di altri utenti è ignorato completamente.

Innanzitutto si individuano le caratteristiche dell'utente basandosi sugli articoli da lui valutati nel passato. Le valutazioni e il comportamento di acquisto dell'utente, combinati con le informazioni sui contenuti degli articoli, permettono di ottenere così la predizione della valutazione che l'utente darebbe al dato articolo. Le descrizioni dell'articolo, alle quali sono associate delle valutazioni, sono usate come *training data* per creare raccomandazioni. Questa tecnica è una generalizzazione di un problema di classificazione o regressione [11].

Il più grande vantaggio di questa tecnica di raccomandazione è la sua capacità di raccomandare nuovi articoli anche se ad essi non sono associate delle valutazioni. Tra gli svantaggi abbiamo invece che questo tipo di approccio dipende molto dal dominio di applicazione e dalla capicità di descrizione degli articoli, poiché si basa sull'analisi degli attributi degli articoli.

Questa soluzione è adatta per raccomandazioni di pagine web, pubblicazioni o notizie, in cui tipicamente il contenuto degli articoli è descritto tramite l'utilizzo di parole chiave. Questo introduce un altro svantaggio ovvero l'ovvietà delle raccomandazioni.

Nonostante questi sistemi di raccomandazione siano molto efficaci nel raccomandare nuovi articoli, non lo sono nel generare raccomandazioni per nuovi utenti. I sistemi di raccomandazione content-based soffrono del problema del *cold-start*.

3.1.3 Raccomandazioni Knowledge-based

Il processo di raccomandazione viene eseguito sulla base di similitudini tra le descrizioni degli articoli e le esigenze dell'utente, oppure tramite vincoli che specificano tali esigenze. Non si tiene conto della storia passata dell'utente o delle valutazioni degli altri utenti. Tale processo si basa sulle esigenze dell'utente e sull'utilizzo della conoscenza del dominio degli articoli, di regole e funzioni di similarià tra gli articoli.

Un particolare articolo può avere attributi che rappresentano delle sue proprietà, un utente può essere interessato solo ad articoli con proprietà specifiche. La

raccomandazione consiste nel suggerire all'utente articoli che rispecchiano le sue preferenze.

I sistemi di raccomandazione knowledge-based possono essere classificati in due sottocategorie, *constraint-based* e *case-based*, che si differenziano per le modalità di interazione con l'utente [1]. Nel modello constraint-based l'utente specifica requisiti o vincoli che l'articolo deve rispettare; la conoscenza del dominio degli articoli e delle regole da esso imposto è molto importante per determinare la corrispondenza tra i vincoli dell'utente e gli attributi dell'articolo. Nel modello case-based l'utente identifica un caso ideale come punto di partenza, che viene utilizzato come punto di riferimento per guidare l'utente nella ricerca di articoli simili, permettendo di cambiare iterativamente uno o più dei suoi attributi.

Il vantaggio di questo approccio è che può essere utilizzato in quasi totale assenza di valutazioni da parte degli utenti, non risente dunque del problema del cold-start. Uno degli svantaggi è che richiede una conoscenza dettagliata del dominio dell'articolo che tipicamente tende ad essere complesso in termini di descrizione delle proprietà. È necessaria una dettagliata descrizione degli attributi degli articoli ed una profilazione dell'utente ben organizzata.

3.2 Valutazioni

La valutazione legata ad un articolo svolge un ruolo fondamentale nei sistemi di raccomandazione. In alcuni casi l'algoritmo per la definizione della raccomandazione è influenzato dalla valutazione dei contenuti, che rappresenta il livello di apprezzamento dell'utente.

Esistono diversi tipi di valutazioni. Le più usate sono quelle basate su intervalli discretizzati di numeri ordinati in cui il valore minimo indica il non apprezzamento e il valore massimo l'apprezzamento. Ad esempio una scala di valutazione a cinque punti è rappresentata dal seguente set {1, 2, 3, 4, 5}, la valutazione consiste nell'associare un numero intero, presente nel set, all'articolo. Un caso particolare di questo tipo di valutazione è quella a cinque stelle, illustrata in Figura 3.4.

Figura 3.4: Esempio valutazione a 5 stelle

Fonte: Bengfort et al. [4]

Tipicamente l'insieme dei valori che la valutazione può assumere è un numero dispari, di modo che si abbia il valore centrale corrispondente al significato "indifferente" e che i valori negativi e positivi siano presenti in egual numero. Si parla in questo caso di scala di valutazione bilanciata. Nel caso in cui l'insieme dei valori che la valutazione può assumere dovesse essere un numero pari, allora il valore neutro potrebbe non essere presente. Questo approccio è indicato come sistema di valutazione a scelta forzata, ovvero l'articolo è stato apprezzato oppure non è stato apprezzato.

Si possono anche utilizzare valori categorici ordinati come ad esempio: {Per niente d'accordo, Poco d'accordo, Neutro, Abbastanza d'accordo, Molto d'accordo}, nota anche come scala Likert e riportata in Figura 3.5. In questo caso si tratta di sistemi di valutazione ordinale poiché si utilizza il concetto di variabili categoriali ordinali.

Figura 3.5: Esempio valutazione secondo la scala Likert

Indichi il suo grado di soddisfazione/insoddisfazione per i seguenti aspetti	Per niente soddisfatto	Poco soddisfatto	Indeciso	Abbastanza soddisfatto	Molto soddisfatto
Tempo di attesa	☐	☐	☐	☐	☐
Gentilezza e cortesia del personale	☐	☐	☐	☐	☐
Chiarezza delle informazioni ricevute	☐	☐	☐	☐	☐
Comfort dei locali	☐	☐	☐	☐	☐

Nel caso di valutazioni binarie l'utente può esprimere solo "mi piace" o "non mi piace" e nient'altro.

Un caso speciale sono le valutazioni unarie, nelle quali si ha un meccanismo per l'utente di mostrare il suo apprezzamento ma non c'è nessuno meccanismo per rappresentare l'avversione. Le valutazioni unarie sono utilizzate nei casi di feedback impliciti, in cui le preferenze dell'utente derivano dalle sue attività piuttosto che dalle sue valutazioni esplicitamente specificate. Ad esempio, quando un utente acquista un articolo, o visualizza un articolo, può essere visto come una preferenza per quell'articolo. Le valutazioni unarie sono particolarmente diffuse nei social networks.

3.3 Sistemi di Raccomandazione in ambito sanitario

Al giorno d'oggi i sistemi di raccomandazione vengono utilizzati principalmente per aiutare gli utenti a fare scelte relative all'intrattenimento e all'e-commerce: tutti si

affidano ad algoritmi di raccomandazione per aumentare le loro vendite.

Quando si tratta di settori che riguardano la nostra salute, come l'educazione, l'alimentazione, l'esercizio fisico, i sistemi di raccomandazione dovrebbero prestare particolare attenzione a concetti quali attendibilità e affidabilità [27]. Nonostante il potenziale di sviluppo dei sistemi di raccomandazione in campo sanitario sia alto, sono altrettanto alti i potenziali rischi di un uso non consapevole di tali sistemi. Dal punto di vista del paziente, il sistema deve fornire una semplice interazione, il potenziamento attraverso la spiegazione delle raccomandazioni proposte e la sicurezza contro le raccomandazioni dannose. Ciò consentirà ai pazienti di avere fiducia nel sistema. Per i medici e gli esperti, invece, ciò che conta è una rappresentazione precisa e corretta delle loro conoscenze e dei processi nel settore.

Di seguito si analizzano gli aspetti da affrontare nella realizzazione di un sistema di raccomandazione in ambito sanitario [26]:

- *Profilazione dell'utente e personalizzazione.* Per eseguire la profilazione dell'utente è necessario combinare dati provenienti da fonti eterogenee come informazioni demografiche, diagnosi, risultati di test di laboratorio ecc. Nei sistemi di raccomandazione tradizionali le preferenze degli utenti vengono estratte da suoi comportamenti oppure da sue valutazioni. Nel caso di sistemi di raccomandazione in ambito sanitario le informazioni necessarie sono le necessità dell'utente, le sue malattie ecc. Poiché questi fattori potrebbero contrastare le preferenze dell'utente, è importante prestare attenzione alla sua profilazione e al concetto di personalizzazione. Nei contesti di intrattenimento e e-commerce le preferenze dell'utente vengono utilizzate per rendere l'articolo più appetibile ai suoi occhi, in un contesto sanitario il concetto di personalizzazione assume diverse sfaccettature che vanno dalla personalizzazione vera e propria al modo in cui si mostrano i contenuti.

- *Persuasione, responsabilizzazione e fiducia.* È importante prestare attenzione a come vengono presentate le raccomandazioni e su come l'utente può interagire con esse. È fondamentale che l'utente comprenda a pieno le informazioni che gli si stanno fornendo. Le raccomandazioni personalizzate, se adeguatamente dettagliate, migliorano la comprensione da parte degli utenti della loro condizione medica. È opportuno che l'utente capisca tramite quali passaggi il sistema di raccomandazione abbia prodotto una certa raccomandazione, in questo modo si avrà una trasparenza di ciò che è stato fatto e l'utente potrà fidarsi del risultato ottenuto.

- *Identificare le necessità dell'utente.* È importante che l'utente abbia chiara la sua situazione medica e quale sia effettivamente la cura/riabilitazione di cui ha bisogno. Molto spesso i pazienti non sono istruiti circa le loro malattie e le opzioni di trattamento fino alle fasi successive, quindi ciò di cui il paziente crede possa avere bisogno potrebbe non essere ciò di cui ha effettivamente bisogno.

- *Soddisfazione dell'utente.* Poiché le preferenze dell'utente e ciò di cui l'utente ha effettivamente bisogno potrebbero andare in direzioni diverse, misurare la soddisfazione dell'utente può essere una sfida.
- *Privacy.* Quando i sistemi di raccomandazione operano in campo sanitario, è necessario tenere conto della questione privacy, non solo per quanto riguarda i dati forniti dall'utente ma anche per quanto riguarda le valutazioni, che possono contenere involontariamente informazioni personali. Questi sistemi devono, inoltre, considerare i potenziali danni, le contraddizioni dei servizi di pubblica utilità e le questioni etiche.
- *Conseguenze di una raccomandazione errata.* Bisogna sempre tenere a mente che le conseguenze di una falsa o errata raccomandazione in ambito sanitario possono essere disastrose per la salute e la vita dei pazienti.

Capitolo 4

Soluzione Proposta

Nel seguente capitolo si analizza nel dettaglio la creazione del sistema di raccomandazione. In particolare nei successivi paragrafi si descrivono la progettazione di tale sistema, le fonti dato utilizzate, le strutture dei vari dataset e l'architettura realizzata.

4.1 Progettazione del sistema

Dalla breve introduzione sui sistemi di raccomandazione, si evince che il sistema che più rispecchia le specifiche di progettazione sia il sistema knowledge-based, poiché permette di individuare la struttura riabilitativa più adeguata specificando i vincoli clinici e le preferenze dell'utente.

4.1.1 Definizione degli input e delle fasi del sistema

Il primo passo per la progettazione del sistema è stato quello di definire quali fossero gli input del sistema di raccomandazione. Di tali input è stato necessario stabilire quali potessero essere espressi come preferenze da parte dell'utente e quali invece dovessero essere considerati come non opinabili. Bisogna tenere sempre presente, come evidenziato da Suksom et al. [28], che bilanciare le cure di cui il paziente ha bisogno e le sue preferenze, tipicamente poco salutari, non sia cosa semplice.

Si è deciso di individuare due categorie di input: input oggettivi e input soggettivi. Gli input oggettivi, considerati parametri su cui l'utente non può esperire preferenze, riguardano le cure cliniche di cui ha bisogno. Gli input soggettivi, invece, riguardano delle caratteristiche al contorno sulle quali l'utente può esprimere una preferenza. Queste preferenze possono essere indicate come vincoli da parte dell'utente: sulle

caratteristiche secondarie delle struttura, sulle modalità di raggiungimento della strutture e sul livello di valutazione che la struttura ha ricevuto.

Analizzando alcuni esempi di sistemi di raccomandazione si è osservato come il miglior modo di procedere per generare una raccomandazione sia quello di dividere la generazione della classifica, nel nostro caso di strutture riabilitative, in fasi tematiche, migliorando il risultato del filtraggio ad ogni fase [8, 30].

Colombo-Mendoza et al. [8], ad esempio, hanno realizzato un sistema in cui la raccomandazione di una proiezione di un film è determinata tenendo conto della posizione dell'utente, dell'istante di tempo in cui la richiesta di raccomandazione è stata fatta, dei livelli di affollamento dei vari cinema e delle preferenze dell'utente in merito al genere di film che vorrebbe vedere. La caratteristica interessante di questa soluzione è la divisione dell'architettura di raccomandazione in diversi livelli di filtraggio. Dopo avere ottenuto informazioni sulla posizione dell'utente e sull'istante di tempo, si esegue un pre-filtraggio sulla posizione, eliminando i cinema troppo lontani, e sul tempo, eliminando i film non al momento trasmessi oppure il cui orario di inizio non è adeguato, considerando la distanza dell'utente dal cinema. Una volta ottenuta questa prima lista di possibili proiezioni, si procede nel determinare quali possano essere i film di interesse dell'utente analizzando le preferenze espresse tramite il suo profilo. A questo punto è possibile determinare la raccomandazione composta da: nome del cinema, nome del film e ora della proiezione.

Con una logica simile, Yuan et al. [30] hanno proposto un sistema di raccomandazione per la vendita e l'affitto di immobili. In questo caso come punto di partenza si considera la zona geografica di gradimento dell'utente, indicata su una mappa, e si genera una prima classifica riportante tutti gli immobili presenti in quella zona. In seguito l'utente indica altre caratteristiche, relative ad esempio al numero di camere, la fascia di prezzo ecc. e ad ogni vincolo impostato la classifica si aggiorna, gli immobili che non rispettano più certi vincoli non sono più presenti.

Partendo dall'analisi di questi lavori, nel sistema sviluppato è stato opportuno dividere la generazione della classifica in due fasi: nella prima si considerano solo i vincoli oggettivi e nella seconda i vincoli soggettivi.

Essendo l'ambito di applicazione del sistema quello sanitario, si sono poi analizzati dei casi di sistemi di raccomandazione realizzati in questo ambito, per individuarne le caratteristiche e le peculiarità.

Salunke e Kasar [25], per esempio, hanno proposto un sistema per la ricerca di un medico o di una struttura ospedaliera. La caratteristica interessante di questa soluzione è l'introduzione dell'utilizzo delle valutazioni degli utenti come attributo della ricerca. La raccomandazione non è eseguita tenendo conto solo delle specifiche dell'utente, ma anche delle valutazioni che il medico o la struttura ospedaliera hanno ricevuto. Per questo si è deciso di introdurre nella soluzione sviluppata un ulteriore fase di filtraggio e di riordinamento della classifica, che tenesse conto delle valutazioni dell'utente.

L'architettura realizzata è stata suddivisa in tre fasi: inserimento dei vincoli sul quadro clinico del paziente, inserimento dei vincoli di posizione dell'utente e della distanza massima che è disposto a percorrere e inserimento dei vincoli sulla valutazione che la struttura ha ricevuto e la presenza di servizi nelle vicinanze.

4.1.2 Bilanciare le preferenze dell'utente

Una volta definiti i parametri sui quali il paziente può esprimere la propria preferenza, si è preso in considerazione la possibilità che l'utente associasse ad ogni parametro un indice di importanza. A tal fine, diverse sono le soluzioni proposte nella letteratura scientifica.

Chen et al. [7] hanno sviluppato un motore decisionale per supportare il paziente nella scelta di strutture ospedaliere. In questa soluzione, fissati i cinque parametri su cui l'utente può esprimere le preferenze, si crea una matrice di confronto accoppiato, su una riga è disposto un parametro, che chiameremo "a", sulle colonne sono riportati i valori che rappresentano l'importanza che l'utente da al parametro a rispetto a tutti gli altri. L'importanza è espressa su una scala a cinque punti, che va da "altrettanto importante" a "estremamente importante". Il vantaggio di questa soluzione è la matrice di confronto accoppiato che può essere dinamicamente cambiata dall'utente.

Omotosho et al. [17] nell'affrontare un problema simile al precedente, al contrario suggeriscono l'utilizzo di pesi fissi non più acquisiti dagli utenti, ma intrinseci del sistema.

Nella soluzione sviluppata si è deciso di utilizzare uno scala tra zero e uno per rappresentare l'importanza che l'utente può attribuire ad un parametro. Il valore zero indica "per nulla importante", uno indica "estremamente importante". La somma dei pesi associati agli attributi deve essere pari ad uno.

4.2 Fonti dato

Nel seguente paragrafo si descrivono brevemente le fonti dati utilizzate nella realizzazione del sistema di raccomandazione.

4.2.1 Strutture di riabilitazione

Le informazioni sulle strutture provengono da fonti molti diverse, contenenti attributi di diversa natura. Tipicamente sono presenti: informazioni di natura generali (quali nome, indirizzo ecc), informazioni sulle caratteristiche della struttura e informazioni relative alle tipologie di patologie trattate. Le fonti dato utilizzate sono fonti regionali, con le quali si è creato un dataset contenente 46 strutture.

4.2.2 Geolocalizzazione, distanza e servizi al contorno: OpenstreetMap

OpenStreetMap (OSM) [18] è un progetto mondiale libero e collaborativo per la raccolta di dati geografici da cui si possono derivare innumerevoli lavori e servizi. OSM è libero poiché i dati al suo interno possiedono una licenza libera (Open Database License). È collaborativo poiché tutti possono contribuire arricchendo o correggendo i dati, è la comunità che inserisce i dati, arricchisce il progetto e ne controlla anche la qualità. OpenStreetMap raccoglie e distribuisce dati geografici, non limitandosi a mostrare mappe e consentendo l'accesso alla totalità dei propri dati.

OSM è utilizzato, dunque, come fonte per ottenere informazioni sulla posizione geografica delle strutture riabilitative di interesse. Oltre ad esser una fonte dato, all'interno del sistema sviluppato, OpenSteetMap è utilizzato come tool per risolvere due problemi: calcolo delle distanze e conoscenza dei servizi presenti in prossimità di una struttura riabilitativa.

Per il calcolo delle distanze è necessario conoscere le coordinate geografiche della posizione, si utilizza a tale scopo il componente *nominatim* di OSM, utilizzato in combinazione con la libreria *geopy*.[1] In questo modo data una stringa contenente l'indirizzo di una struttura è possibile determinare le coordinate geografiche. Oltre all'indirizzo è possibile anche indicare semplicemente il comune, il CAP o il nome di un momento o di un ente, ad esempio specificare "Politecnico di Torino" oppure "Corso Duca degli Abruzzi, 24 Torino" non farà alcuna differenza, le coordinate geografiche indicate saranno le stesse.

Per il calcolo della distanza è utilizzata la libreria *pyroutelib3*.[2] Note le coordinate geografiche del punto di partenza, del punto di arrivo e la modalità con la quale si vuole percorre tale distanza, il risultato è un percorso indicato come un lista di punti da attraversare. I punti riportati corrispondono a cambiamenti di direzione, ad esempio il passaggio per un incrocio oppure per una rotonda. Calcolando la distanza tra i punti intermedi è possibile determinare la distanza totale tra il punto di partenza e quello di arrivo.

Per la determinazione dei servizi al contorno invece si è utilizzato *Overpass turbo* [21]. Si tratta di un'applicazione web, realizzata da Martin Raifer, che permette di effettuare delle *query*, basate sulle Overpass API, ed esportare i risultati ottenuti in diversi formati. Le Overpass API sono API di sola lettura, che permettono di personalizzare la selezione dei dati su una mappa OSM. Overpass turbo Agisce come un database sul web: il client invia una query all'API e recupera il set di dati che corrisponde alla query. Tramite le query è possibile indicare la zona di

[1] Geopy - https://github.com/geopy/geopy

[2] Pyroutelib - https://github.com/MKuranowski/pyroutelib3

interesse sulla mappa OSM (che può essere un'area irregolare, una città, una regione ecc) e specificare le informazioni che si vogliono ottenere. Tra i vari servizi presenti in OSM, all'interno del sistema di raccomandazione sviluppato, si sono presi in considerazione: i parcheggi ed i servizi di ristorazione. Tramite un'opportuna query è stato possibile ottenere informazioni sui parcheggi e sui servizi di ristorazione presenti in tutta la regione Piemonte.

4.2.3 Valutazioni: QSalute

Le valutazioni sulle strutture riabilitative provengono dal sito di QSalute [20]. QSalute, nato nel 2008, è il portale di riferimento sul mondo della salute e della sanità italiana.

Sul portale è possibile leggere ed inserire recensioni su ospedali, case di cura, medici, in base alle testimonianze degli altri pazienti. Per ogni struttura è presente un riquadro contenente una valutazione generale divisa in ambiti. La caratteristica di questo portale è infatti quella di aver già definito gli aspetti fondamentali su cui l'utente è tenuto ad esprimere una valutazione: competenza, assistenza, pulizia e servizi. Oltre al riquadro generale, sono presenti le recensioni lasciati dagli utenti. Le recensioni contengono: il commento testuale dell'utente, informazioni relative ai trattamenti effettuati e le valutazioni a cinque stelle espresse sulle quattro caratteristiche.

4.3 Preproccessing

Affinché il sistema possa determinare la raccomandazione è necessario definire quali siano gli input e il formato in cui questi sono ricevuti, è necessaria dunque una fase di proprocessing dei dati.

4.3.1 Struttura di riabilitazione

Poiché le informazioni sulla struttura provengono da fonti diverse, esse presentano caratteristiche diverse, è necessario scartare le informazioni non di interesse e perfezionare quelle di interesse.

Ad ogni struttura, innanzitutto, si è associato un identificativo univoco, rappresentato da un numero intero positivo. Si sono mantenute poi informazioni generali quali: nome, l'indirizzo della struttura, comune e relativo CAP in cui è situata la struttura e l'indicazione se si tratti di una struttura pubblica o privata. La struttura deve contenere informazioni sulla tipologia di riabilitazione che offre, indicato come setting, e l'ambito in cui è specializzata. In particolare i setting considerati sono quattro: primo livello (RRF1), secondo livello (RRF2), terzo livello (RRF3) e

lungodegenza. Poiché ogni struttura può fornire uno, alcuni o tutti i servizi indicati si è deciso di aggiungere al dato struttura quattro attributi booleani, uno per ogni setting, in cui il valore zero indica che la struttura non fornisce tale servizio, 1 che la struttura fornisce il servizio. Gli ambiti di specializzazione possono essere: neuro-psichiatria, cardio-respiratoria, respiratoria. Uno schema del dato struttura è riportato in Figura 4.1.

Figura 4.1: Informazioni sulla struttura

4.3.2 Posizione

Conoscendo le informazioni sull'indirizzo della struttura, tramite il tool OpenStreetMap, si determina in maniera automatica la sua latitudine e longitudine, utilizzati nel calcolo delle distanze. Determinare le coordinate delle strutture durante la fase di preprocessing dei dati permette di evitarne il calcolo durante la fase di esecuzione e diminuisce notevolmente i tempi di latenza del programma. Il dato struttura viene dunque aggiornato come riportato in Figura 4.2.

4.3.3 Valutazione

Le valutazioni sulle strutture sono ottenute dal sito di QSalute. Da tale portale è possibile estrapolare informazioni di diversa granularità come: una valutazione sintetica della struttura, che tiene conto delle recensioni di tutti gli utenti ed una specifica recensione sulla struttura espressa da un utente.

Per la generazione della raccomandazione, è d'interesse d'interesse solo la valutazione generale legata ad una struttura ; in seguito si analizzeranno nel dettaglio le recensioni testuali. La valutazione generale legata ad una struttura è riportata in un riquadro (si veda la Figura 4.3) contenente: il numero di recensioni che la

Figura 4.2: Informazioni sulla struttura con latitudine e longitudine

struttura ha ricevuto, una valutazione espressa su cinque punti per ogni ambito e una valutazione generale espressa su cinque stelle.

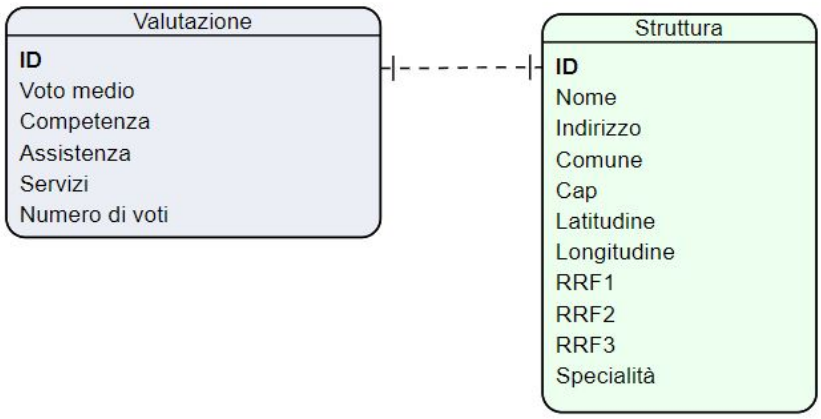
Figura 4.3: Informazioni presenti nella valutazione (QSalute)

Fonte: QSalute [20]

La caratteristica di QSalute è quella di aver già definito gli ambiti fondamentali su cui l'utente è tenuto ad esprimere una valutazione: competenza, assistenza, pulizia e servizi.

Si è deciso di tenere tutti le informazioni riportate nella valutazione, aggiungendo al dato struttura ulteriori attributi come riportato in Figura 4.4.

Figura 4.4: Informazioni sulla struttura con valutazione

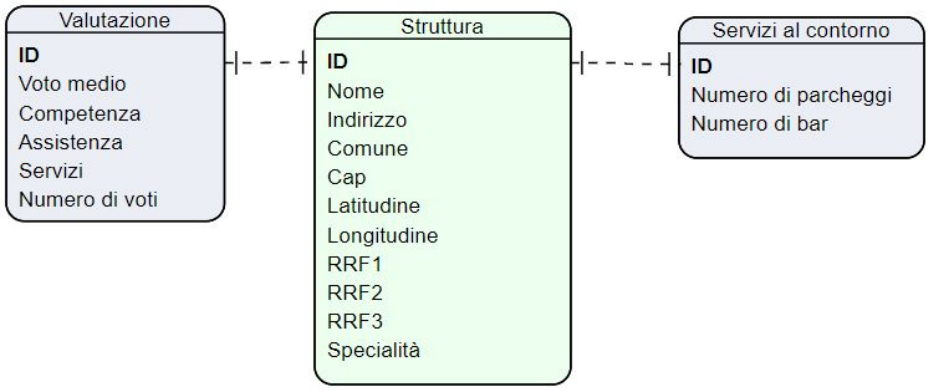


4.3.4 Servizi al contorno

Tramite OpenStreetMap è possibile determinare informazioni sui servizi presenti nelle vicinanze di una struttura.

All'interno del sistema di raccomandazione sviluppato si sono presi in considerazione i parcheggi ed i servizi di ristorazione. Si sono definiti come vicini i servizi presenti nel raggio di seicento metri dalla struttura. Ad ogni struttura si sono associati, dunque, l'informazione sul numero di parcheggi e sul numero di servizi di ristorazione presenti, come riportato in Figura 4.5.

Figura 4.5: Informazioni sulla struttura con servizi presenti nelle vicinanze



4.4 Implementazione della piattaforma

La raccomandazione generata dal sistema proposto avviene tramite la creazione di una classifica di strutture con in cima la struttura che rispecchia maggiormente le desiderate dell'utente ed infondo la struttura che rispecchia meno le desiderate.

Si sono individuati input oggettivi e input soggettivi. Gli input oggettivi rappresentano il quadro clinico del paziente, contenente l'attività di riabilitazione, indicato come setting e la specializzazione. Gli input soggettivi rappresentano gli attributi sui quali l'utente può esprimere una preferenza, che sono: le caratteristiche secondarie delle struttura, la modalità di raggiungimento della struttura e il livello di valutazione che la struttura ha ricevuto.

La soluzione proposta si articola nelle seguenti fasi:

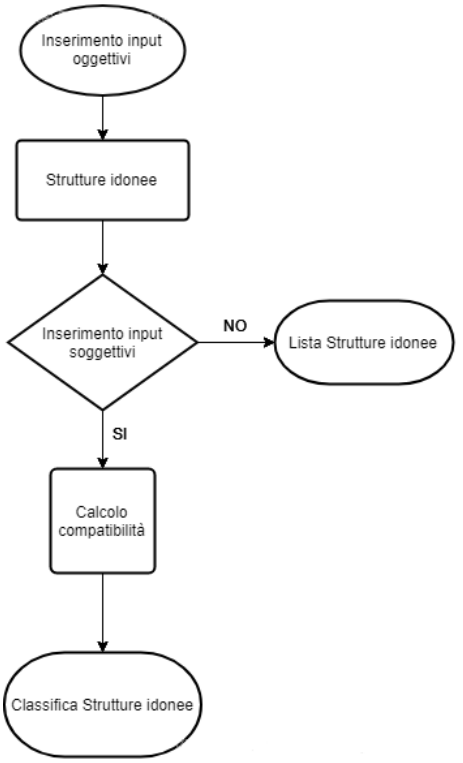
- inserimento dei vincoli sul quadro clinico del paziente;
- creazione di una lista iniziale contenente solo le strutture che soddisfano tali requisiti;
- inserimento dei vincoli di posizione dell'utente e della distanza massima che è disposto a percorrere;
- generazione, per ogni struttura, di un punteggio che rappresenta la sua vicinanza alla posizione dell'utente;
- calcolo dei punteggi rappresentanti il livello di valutazione che la struttura ha ricevuto e la presenza o meno di servizi nelle vicinanze;
- definizione da parte dell'utente dell'ordine di importanza che vuole dare alla distanza, alla valutazione ed alla presenza di servizi;
- calcolo, per ogni struttura, di un punteggio complessivo, calcolato a partire dai punteggi di distanza, valutazione ricevuta e presenza di servizi;
- ordinamento della lista delle strutture in base al punteggio complessivo

L'inserimento degli input soggettivi non è obbligatorio: in questo caso il sistema permetterà all'utente di esplorare la lista delle strutture che rispettano gli input oggettivi indicati; uno schema è riportato in Figura 4.6. In Figura 4.7 è riportata l'architettura del sistema realizzato. Nei paragrafi successivi sono analizzati in dettaglio le fasi qui brevemente descritte.

4.4.1 Selezione di strutture che possiedono i requisiti medicali

L'utente inserisce le informazioni relative al proprio quadro clinico, che come abbiamo visto in precedenza è composto dall'attività di riabilitazione e dalla specializzazione:

Figura 4.6: Schema del sistema realizzato



queste informazioni andranno a determinare il vincolo oggettivo, cioè le caratteristiche che la struttura deve possedere obbligatoriamente. L'attività di riabilitazione deve essere sempre indicata, mentre la specializzazione può essere omessa, di default saranno considerate tutte le specializzazioni. Il sistema provvede a determinare la lista delle strutture che offrono il servizio richiesto e posseggono la specializzazione indicata, se indicata.

4.4.2 Informazioni sulla posizione e calcolo del punteggio relativo alla distanza

Inserite le informazioni relative al proprio quadro clinico, l'utente fornisce informazioni sulla sua posizione, specificando indirizzo, comune ed una distanza massima che è disposto a precorrere. Le strutture che distano dalla posizione dell'utente oltre la distanza indicata non sono di interesse.

Noto l'indirizzo si determinano le coordinate geografiche della posizione e si calcola la distanza per ogni coppia: posizione del paziente, struttura presente in lista.

A questo punto ad ogni struttura è associata la distanza della stessa dalla posizione del paziente, è necessario però stabilire un punteggio da associare all'attributo distanza. La distanza può assumere un valore tra zero e infinito in un dominio continuo, si definisco allora degli intervalli di distanza ed ad ogni intervallo viene associato il corrispettivo valore del punteggio. Nel caso del sistema realizzato, per le strutture che distano entro i cinque chilometri dalla posizione dell'utente si assegna un punteggio pari ad uno, per quelle che distano oltre i dieci chilometri si assegna un punteggio pari a 0,8 e così via come riportato nella tabella Tabella 4.1.

Figura 4.7: Architettura del sistema realizzato

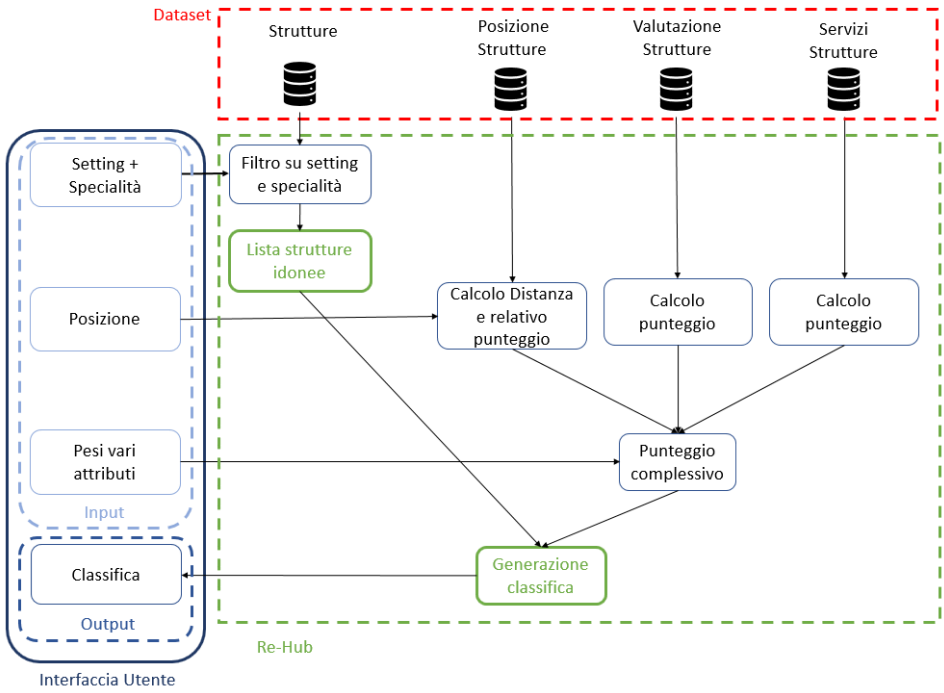


Tabella 4.1: Associazione distanza punteggio

Distanza	Valore (D)
La struttura dista al più 5 km	1
La struttura dista al più 10 km	0,8
La struttura dista al più 30 km	0,6
La struttura dista il valore massima indicato dall'utente	0,4
La struttura dista oltre il valore massimo indicato dall'utente	0

4.4.3 Calcolo del punteggio relativo alla valutazione

Dall'analisi dei dati presenti su Qsalute si è notato che la valutazione di una struttura è stata divisa in cinque punti: voto medio, competenza, assistenza, pulizia e servizi. È inoltre noto il numero di voti che la struttura ha ricevuto.

Per classificare le strutture in base alle valutazione che hanno ricevuto è necessario determinare un punteggio che tenga conto dei valori associati ai cinque punti. Una possibile soluzione potrebbe essere quella di normalizzare la valutazione della struttura, espressa su una scala a cinque punti, in un valore compreso tra zero e uno. In questo modo la struttura con la valutazione più alta si trova in cima alla classifica mentre quella con la valutazione più bassa si trova in fondo alla lista. Questa soluzione però è troppo semplice e non tiene conto del fatto che il numero delle valutazione che una struttura riceve è molto vario. Considerare in egual modo una struttura con una valutazione complessiva di cinque punti ed un numero di valutazioni alto ed una con la stessa valutazione ma un numero di valutazioni pari ad uno, è troppo semplicistico. Si è deciso così di tenere conto del numero di valutazioni che la struttura ha ricevuto oltre alla valutazione stessa, utilizzando un *shrinkage estimator* [29].

Un shrinkage estimator prende una stima grezza e la migliora combinandola con altre informazioni, in questo caso la stima grezza è la valutazione mentre l'informazione aggiuntiva è il numero di valutazioni ricevute. L'idea di base è che più valutazioni siano state espresse più la valutazione complessiva della struttura è rappresentativa. La formula utilizza è la seguente

$$E = \frac{n}{n+m} * (v + \frac{m}{m+n}) * C \tag{4.1}$$

in cui:

- v è la valutazione che la struttura ha ricevuto;
- n è il numero di valutazioni;
- m è il numero minimo di valutazioni richiesto
- C è il valore medio delle valutazioni complessive di tutte le strutture.

Il numero minimo di valutazioni richiesto nel sistema implementato è pari a tre. In questo modo se si considera il caso di due strutture con la stessa valutazione complessiva ma con differente numero di valutazioni, quella col numero maggiore di valutazioni si trova più in alto nella classifica rispetto a quella con numero di valutazioni inferiore.

Poiché il risultato della formula è un valore maggiore di uno, si normalizza tale valore con la tecnica di *min-max normalizzation*, con la seguente formula

$$v' = \frac{v - min}{max - min} \tag{4.2}$$

in cui:

- v è il valore precedentemente calcolato;
- min è il valore minimo assunto dal shrinkage estimator calcolato per tutte le strutture;
- max è il valore massimo assunto dal shrinkage estimator calcolato per tutte le strutture.

Questo metodo è ripetuto per ogni campo della valutazione: voto medio, competenza, assistenza, pulizia e servizi, ottenendo così una rappresentazione della valutazione per ogni suo aspetto. Il flusso di esecuzione dunque è il seguente, ripetuto per ogni campo della valutazione: si calcola il valore medio delle valutazioni di tutte le strutture, si calcola il valore E applicando la formula (4.1), si individuano i valori di minimo e massimo assunti dall'shrinkage estimator e il valore E calcolato in precedenza viene normalizzato utilizzando la formula (4.2).

Determinati i punteggi associati a voto medio, competenza, assistenza, pulizia e servizi, si vuole calcolare un punteggio complessivo utilizzando la media pesata. Poiché si vuole dare maggiore importanza al voto medio rispetto agli attributi, il peso associato ad esso è maggiore, mentre i pesi associati agli altri campi è inferiore ed è uguale per tutti. I pesi utilizzati nel sistema proposto sono riportati nella seguente Tabella 4.2.

Tabella 4.2: Peso associato ad ogni campo della valutazione

Campo della valutazione	Peso
Voto medio	0,6
Competenza	0,1
Assistenza	0,1
Pulizia	0,1
Servizi	0,1

4.4.4 Calcolo del punteggio relativo ai servizi al contorno

Per quanto riguardo la presenza di servizi nelle vicinanze della struttura si sono presi in considerazione il numero di parcheggi e di servizi di ristorazione e ad entrambi questi attributi si sono associati dei punteggi.

Dato il numero di parcheggi presenti nelle vicinanze, tale numero è normalizzato tra zero e uno utilizzando la formula (4.2). L'idea è che più parcheggi siano presenti nelle vicinanze della struttura più sarà facile per l'utente trovare parcheggio e questo è considerato un valore in più. Nel caso dei servizi di ristorazione si è

valutata semplicemente la presenza o meno di questi servizi nelle vicinanze, quindi nel caso una struttura abbia un servizio di ristorazione avrà un punteggio pari ad uno altrimenti zero. L'obiettivo finale è il calcolo di un punteggio complessivo che rappresenti queste due caratteristiche, per ottenerlo si utilizza la media pesata in cui si dà maggiore importanza al parcheggio, con un peso di 0,7, e minor importanza alla presenta di servizi di ristorazione con un peso di 0,3.

4.4.5 Realizzazione della classifica

Per realizzare la classifica si è associato ad ogni struttura un punteggio, inteso come valore compreso tra zero e uno, in cui il valore uno indica che la struttura rispecchia tutti i requisiti espressi dell'utente. Tale punteggio è calcolato come media pesata dei punteggi associati ai tre attributi analizzati in precedenza: distanza della struttura dalla posizione del paziente, valutazione che la struttura ha ricevuto e presenza di servizi nelle vicinanze. La formula utilizzata per il calcolo della media pesata è la seguente:

$$p = \mathcal{D} * p_{\mathcal{D}} + \mathcal{V} * p_{\mathcal{V}} + \mathcal{S} * p_{\mathcal{S}}$$

in cui:

- D rappresenta il punteggio associato alla distanza della struttura dalla posizione del paziente;
- V rappresenta il punteggio associato alla valutazione che la struttura ha ricevuto;
- S rappresenta il punteggio associato alla presenza di servizi nelle vicinanze della struttura;
- p_D rappresenta l'importanza che l'utente attribuisce alla distanza della struttura dalla posizione del paziente;
- p_V rappresenta l'importanza che l'utente attribuisce alle valutazioni che la struttura ha ricevuto;
- p_S rappresenta l'importanza che l'utente attribuisce alla presenza di servizi nelle vicinanze della struttura.

I punteggi ed i pesi associati agli attributi sono entrambi espressi come valori tra zero e uno. Mentre i punteggi sono calcolati dal sistema, i pesi sono espressi dall'utente che fissa i valori di ogni peso considerando che la somma dei pesi deve essere pari ad uno. Nella Tabella 4.3 è riportato un esempio di pesi che l'utente potrebbe attribuire, in cui da uguale importanza alla distanza e alle valutazioni e minor importanza ai servizi.

Tabella 4.3: Esempio di pesi stabiliti dall'utente

Attributo	Peso
Distanza	0,4
Valutazione	0,4
Presenza di servizi	0,2

4.5 Casi di studio

Per illustrare i risultati ottenuti con il sistema realizzato si sono riportati alcuni casi studio.

In Tabella 4.4 è riportata la lista delle strutture, ottenuta specificando come setting secondo livello e specializzazione cardio-respiratoria. Nella tabella sono riportati: l'identificativo della struttura, il numero di parcheggi e servizi di ristorazione presenti nelle vicinanze e il voto medio che la struttura ha ricevuto. Si tratta di una lista e non di una classifica poiché non sono stati specificati input soggetivi.

Tabella 4.4: Lista strutture con setting secondo livello e specializzazione cardio-respiratoria

Struttura (Ident.)	Parcheggi (N.)	Ristorazione (N.)	Valutazione (V. Medio)
2	0	0	4.1
10	0	0	2.8
19	8	5	3
22	0	0	3.9
26	0	0	5
30	0	0	4.8
34	0	0	3.8
37	1	3	4
46	5	0	0

In Tabella 4.5 è riportata la classifica delle strutture, ottenuta specificando come setting secondo livello, indirizzo di partenza "Corso duca degli Abruzzi, 24 Torino" e dando ugual importanza alla distanza della struttura dalla posizione del paziente, alla presenza di servizi e alla valutazione che la struttura ha ricevuto. Nella tabella oltre alle informazioni indicate in precedenza si è aggiunto il valore corrispondente alla compatibilità, ovvero il punteggio complessivo, espresso in percentuale, calcolato dal sistema.

In Tabella 4.6 è riportata la classifica delle strutture, ottenuta specificando come setting secondo livello, indirizzo di partenza "Corso duca degli Abruzzi, 24 Torino"

e dando maggior importanza alla valutazione che la struttura ha ricevuto e minor importanza alla distanza e alla presenza di servizi.

Tabella 4.5: Classifica strutture con setting secondo livello e con indicazione di uguale importanza su distanza, valutazione e servizi

Struttura (Ident.)	Compatibilità (%)	Distanza (Km)	Parcheggi (N.)	Valutazione (V. Medio)	Ristorazione (N.)
18	91,71	2,33	14	4,3	5
21	78,68	3,53	7	4,2	2
1	73,21	2,41	2	4,4	1
19	66,36	3,06	8	3	5
7	56,98	27,29	0	4,4	1
17	55,38	3,51	5	3,3	0
20	52,31	3,01	0	0	0
32	47,43	24,37	0	4,6	0
31	47,38	33,57	0	4,2	1
5	46,99	20,33	0	4,4	0
22	46,72	6,48	0	3,9	0
2	42,30	18,04	0	4,1	0
33	40,70	45,34	0	4,7	0
28	40,67	38,07	1	5	0
27	39,28	10,35	5	3,1	0
23	39,16	16,86	0	3,8	0
25	38,99	10,16	0	0	0
12	38,53	57,24	0	5	1
24	35,74	16,65	1	2,9	0
11	35,61	44,97	0	4,1	0
35	33,17	90,56	0	5	4
26	32,75	81,13	0	5	0
36	32,45	62,12	0	4,8	1
37	32,41	57,39	1	4	3
46	27,33	97,73	5	0	0
14	26,17	74,16	0	4,4	0
39	24,12	62,03	1	4,8	0
29	24,09	83,29	0	4,3	0
30	22,46	129,50	0	4,8	0
34	19,12	76,58	0	3,8	0
38	19,12	81,75	0	3,8	0
40	18,22	86,46	0	3,7	0
10	8,03	122,07	0	2,8	0

In Tabella 4.7 è riportata la classifica delle strutture, ottenuta specificando come setting secondo livello, indirizzo di partenza "Via Cernaia, 4 Novara" e dando maggior

importanza alla distanza e minor importanza alla valutazione e alla presenza di servizi nelle vicinanze.

Tabella 4.6: Classifica strutture con setting secondo livello e con indicazione di maggiore importanza su valutazione e minore su distanza e servizi

Struttura (Ident.)	Compatibilità (%)	Distanza (Km)	Parcheggi (N.)	Valutazione (V. Medio)	Ristorazione (N.)
18	80.34	2.33	14	4.3	5
26	78.68	81.13	0	5	0
1	77.87	2.41	2	4.4	1
7	73.89	27.29	0	4.4	1
21	73.53	3.53	7	4.2	2
32	71.94	24.37	0	4.6	0
12	71.57	57.24	0	5	1
5	70.89	20.33	0	4.4	0
33	69.78	45.34	0	4.7	0
28	66.20	38.07	1	5	0
31	64.81	33.57	0	4.2	1
14	62.86	74.16	0	4.4	0
2	59.62	18.04	0	4.1	0
35	58.69	90.56	0	5	4
29	57.88	83.29	0	4.3	0
11	57.54	44.97	0	4.1	0
36	56.95	62.12	0	4.8	1
22	56.23	6.48	0	3.9	0
20	55.66	3.01	0	0	0
39	54.45	62.03	1	4.8	0
30	53.95	129.50	0	4.8	0
37	53.36	57.39	1	4	3
23	52.08	16.86	0	3.8	0
25	51.66	10.16	0	0	0
46	48.16	97.73	5	0	0
34	45.93	76.58	0	3.8	0
38	45.93	81.75	0	3.8	0
17	45.54	3.51	5	3.3	0
40	43.76	86.46	0	3.7	0
19	40.42	3.04	8	3	5
24	40.37	16.65	1	2.9	0
27	34.86	10.35	5	3.1	0
10	19.29	122.07	0	2.8	0

È possibile notare come cambiando l'attività di riabilitazione e la specializzazione alcune strutture non compaiono più nella lista; cambiando l'ordine d'importanza degli input o la posizione dell'utente la classifica si aggiorna mostrando un diverso ordine delle strutture. In particolare tali risultati sono evidenti osservando i cambiamenti della struttura con identificativo 40 nelle varie classifiche. Nella

Tabella 4.4 la struttura non è presente perché non possiede la specializzazione in cardio-respiratoria. Nella Tabella 4.6 si trova in fondo alla classifica poiché si è data maggior importanza alla valutazione. Nella Tabella 4.7 è al primo posto poiché si è data maggior importanza alla distanza.

Tabella 4.7: Classifica strutture con setting secondo livello e con indicazione di maggior importanza su distanza e minore su valutazione e servizi

Struttura (Ident.)	Compatibilità (%)	Distanza (Km)	Parcheggi (N.)	Valutazione (V. Medio)	Ristorazione (N.)
40	92.74	0.26	0	3.7	0
46	58.10	29.04	5	0	0
29	57.62	18.44	0	4.3	0
36	40.87	37.98	0	4.8	1
39	39.62	45.81	1	4.8	0
18	8.77	84.43	14	4.3	5
21	6.81	83.24	7	4.2	2
1	5.99	88.29	2	4.4	1
12	5.79	100.34	0	5	1
7	5.56	91.58	0	4.4	1
31	5.11	119.99	0	4.2	1
35	4.98	88.90	0	5	4
19	4.96	83.69	8	3	5
26	4.92	149.39	0	5	0
37	4.87	122.05	1	4	3
32	4.12	97.37	0	4.6	0
33	4.11	67.67	0	4.7	0
28	4.11	51.59	1	5	0
5	4.06	82.56	0	4.4	0
14	3.93	83.59	0	4.4	0
30	3.37	63.40	0	4.8	0
2	3.35	104.62	0	4.1	0
11	3.35	69.64	0	4.1	0
17	3.31	83.92	5	3.3	0
22	3.01	80.64	0	3.9	0
27	2.90	85.77	5	3.1	0
23	2.88	101.51	0	3.8	0
34	2.87	55.44	0	3.8	0
38	2.87	126.81	0	3.8	0
20	2.85	85.21	0	0	0
25	2.85	76.78	0	0	0
24	2.37	81.93	1	2.9	0
10	1.21	59.95	0	2.8	0

Capitolo 5

Tecniche di Text Mining

5.1 Introduzione al text mining

Nel mondo di internet, la rivoluzione testuale ha visto un enorme cambiamento nella disponibilità di informazioni online, trovare informazioni per qualsiasi esigenza non è mai stato così automatico. Se da un lato, però, la digitalizzazione e la creazione di materiali testuali continua alla velocità della luce, la capacità di navigare, estrarre o sfogliare casualmente documenti, sempre troppo numerosi per poterli leggere, si sviluppa a rilento.

Con il termine *text mining* si fa riferimento ad un processo di *knowledge discovery*, in cui un utente interagisce con una raccolta di documenti utilizzando una suite di strumenti di analisi. Il campo del text mining ha guadagnato molta attenzione negli ultimi anni a causa dell'enorme quantità di dati testuali, che vengono creati in una varietà di forme come i social network, le cartelle cliniche, i dati delle assicurazioni sanitarie, le agenzie di stampa, ecc. I dati testuali sono un buon esempio di informazione non strutturata, che è una delle forme più semplici di dati che possono essere generati nella maggior parte degli scenari. Il testo non strutturato viene facilmente elaborato e percepito dagli esseri umani, ma è molto più difficile da capire per le macchine.

Molto spesso il concetto di text mining viene associato al concetto di *data mining* [10]. Il data mining è meglio caratterizzato come l'estrazione dai dati di informazioni precedentemente sconosciute e potenzialmente utili. L'informazione è implicita, non poteva essere estratta senza ricorrere ad opportune tecniche di analisi. Nel caso di text mining, invece, le informazioni da estrarre sono esplicitamente e chiaramente indicate nel testo. Il problema è che le informazioni non sono formulate in un modo strutturato e quindi non è possibile utilizzare direttamente elaborazioni automatiche.

Al fine di sfruttare appieno i dati codificati nel linguaggio, è necessario pensare al linguaggio non come intuitivo e naturale, ma come arbitrario e ambiguo. Questo perché le parole non hanno un significato fisso e universale, indipendentemente da contesti come la cultura e la lingua. Piuttosto che essere definiti da regole, i linguaggi naturali sono definiti dall'uso, sono dinamici, in rapida evoluzione per rappresentare l'espressività umana e ciò implica ridondanza, ambiguità e prospettiva nell'interpretazione di un testo.

Il text mining viene spesso utilizzato per analizzare raccolte di documenti. Una raccolta di documenti può essere qualsiasi raggruppamento di dati testuali e la maggior parte delle tecniche in questo campo hanno lo scopo di individuare informazioni rilevanti in grandi collezioni. Il numero di documenti in tali collezioni può variare da molte migliaia a decine di milioni. Le raccolte di documenti possono essere sia statiche, nel qual caso la raccolta iniziale rimane invariata, o dinamiche, col trascorrere del tempo i documenti possono essere aggiornati ed è possibile inserirne di nuovi.

Ogni raccolta è formata da un numero specifico di documenti. A fini pratici, un documento può essere definito come un'unità di dati testuali discreti all'interno di una raccolta che solitamente, ma non necessariamente, è correlata ad un argomento comune. Tuttavia, dato il numero potenzialmente elevato di parole, frasi, espressioni, elementi tipografici ed elementi di impaginazione, che anche un breve documento può avere, per non parlare del numero potenzialmente vasto di sensi diversi che ciascuno di questi elementi può avere in vari contesti e combinazioni, un compito essenziale per la maggior parte dei sistemi di text mining è l'identificazione di un sottoinsieme semplificato di caratteristiche del documento che può essere utilizzato per rappresentare un particolare documento nel suo complesso.

Anche se diverse sono le caratteristiche che possono essere utilizzate per rappresentare i documenti, i quattro tipi più comunemente utilizzati, indicati in [10], sono i seguenti:

- Caratteri. I singoli componenti: lettere, numeri, caratteri speciali e spazi sono gli elementi costitutivi di caratteristiche semantiche di livello superiore come parole, termini e concetti. Una rappresentazione a livello di carattere può includere l'insieme completo di tutti i caratteri per un documento o qualche sottoinsieme filtrato. Questa tipologia di rappresentazione è spesso di utilità molto limitata per le applicazioni di text mining.
- Parole. Utilizzare parole selezionate direttamente da un documento porta ad una adeguata ricchezza semantica.
- Termini. Si tratta di singole parole e/o frasi composte da più parole selezionate, direttamente dal documento per mezzo di tecniche di estrazione dei termini.
- Concetti. Si tratta di caratteristiche generate da un documento per mezzo di metodologie di categorizzazione che possono essere: manuali, statistiche, basate su regole o ibride. Tale categorizzazione ha lo scopo di individuare

singole parole, espressioni multiparola o unità sintattiche ancora più grandi che sono poi collegate a specifici identificatori di concetto.

5.2 Approcci al text mining

L'obiettivo generale del text mining è di quello estrarre informazioni di alta qualità da un testo. Tale campo copre una vasta gamma di argomenti correlati, di algoritmi per l'analisi testuale, tra cui la ricerca di informazioni, ed elaborazione del linguaggio naturale, tenendo presente che un testo può essere strutturato, semi-strutturato o non strutturato.

Di seguito vengono elencati brevemente i possibili approcci al text mining [2]:

- *Information Retrieval*. Si tratta dell'attività di ricerca di risorse informative, nella forma di documenti, da una raccolta di insiemi di dati non strutturati che contengano informazioni di interesse. Il recupero di informazioni è concentrato principalmente su facilitare l'accesso alle informazioni piuttosto che analizzarle e cercare modelli nascosti.

- *Natural Language Processing*. Si tratta di un sottocampo di informatica, intelligenza artificiale e linguistica che mira alla comprensione del linguaggio naturale utilizzando computer.

- *Information Extraction from text*. Ha il compito di estrarre automaticamente informazioni da documenti non strutturati o semi-strutturati. È tipicamente usato come punto di partenza per l'applicazione di molti algoritmi di text mining.

- *Text Summarization*. Ha lo scopo di riassumere i documenti di testo per ottenere una panoramica concisa sul contenuto di un grande documento o di una raccolta di documenti.

- *Unsupervised Learning Methods*. Si tratta di tecniche che cercano di trovare strutture nascoste all'interno di dati non etichettati. Il *clustering* e il *topic modeling* sono i due metodi principali di apprendimento non supervisionato, comunemente utilizzati nel contesto dei dati di testo. Il clustering ha il compito di dividere una raccolta di documenti in *clusters*, ovvero gruppi in cui i documenti nello stesso gruppo sono più simili l'uno all'altro rispetto a quelli in altri cluster. Il topic modeling è un tipo di tecnica in cui scansionando un insieme di documenti, si esaminano come parole e frasi coesistono in essi, e automaticamente si individuano gruppi di parole che meglio caratterizzano i documenti. Mentre il clustering è un'analisi di tipo deduttivo, poiché cerca di stabilire gruppi di documenti all'interno di un insieme di documenti, il topic modeling è un'analisi di tipo induttivo, poiché mira ad astrarre i temi centrali.

- *Supervised Learning Methods*. Sono tecniche di *machine learning* che permettono di dedurre una funzione o un classificatore dai *training data* per effettuare previsioni su dati nuovi. Esiste un'ampia gamma di metodi supervisionati come i classificatori *nearest neighbor*, *decision trees* ecc.

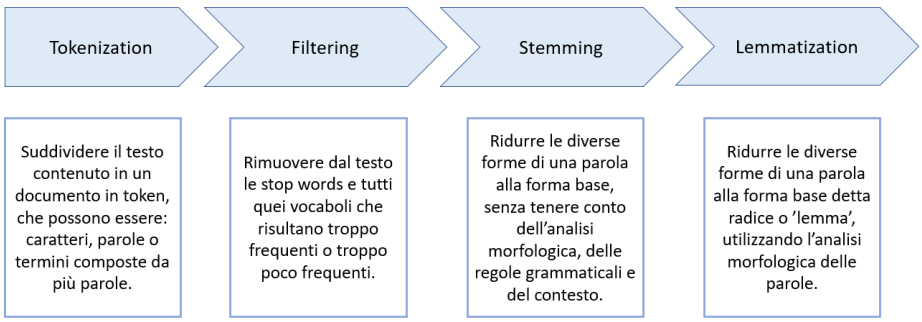
- *Opinion Mining and Sentiment Analysis*. Con l'avvento dell'e-commerce e dello shopping online, viene creata un'enorme quantità di testo riguardante le recensioni di prodotti o opinioni degli utenti. Attraverso la raccolta di questi dati troviamo importanti informazioni di fondamentale utilità nella pubblicità e nel marketing online.

5.3 Definizione di corpus e preprocessing del testo

L'insieme dei documenti sui quali si vuole applicare le tecniche di text mining è detto *corpus*. Un corpus può essere grande o piccolo, anche se generalmente consiste di decine o addirittura centinaia di gigabyte di dati all'interno di migliaia di documenti. I documenti contenuti in un corpus possono variare nelle dimensioni, dai tweet ai libri, e contengono testo e talvolta metadati. Il corpus può essere annotato, in cui il testo o i documenti sono etichettati con le informazioni corrette per gli algoritmi di apprendimento supervisionato, o non annotato, rendendolo candidato per le tecniche di *topic modeling* e di *clustering*.

Qualsiasi corpus reale nella sua forma grezza è completamente inutilizzabile per l'analisi senza una significativa fase di *preprocessing*, ovvero pre-elaborazione e compressione. Tale fase è uno dei componenti chiave di molti algoritmi di text mining e consiste solitamente in compiti come [14]: *tokenization*, *filtering*, *stemming* e *lemmatization*, come riportato in Figura 5.1.

Figura 5.1: Preprocessing



5.3.1 Tokenization

Nella fase di preprocessing, il primo passo è quello di estrarre un insieme di elementi rilevanti che possono essere, come abbiamo visto in precedenza: caratteri, parole, termini ecc. La tecnica più usata per eseguire questa operazione è detta tokenization, la quale si occupa di suddividere il testo contenuto in un documento in token.

Il token è definibile come un blocco di testo atomico, tipicamente formato da caratteri indivisibili separato da delimitatori. Individuare un delimitatore opportuno non è un'operazione semplice. Anche se in un primo momento si potrebbe pensare che gli spazi, le tabulazioni e i caratteri di ritorno a capo possano bastare, in realtà in molte lingue ci sono moltissime eccezione associate a ciascuno dei delimitatori appena indicati. Basti pensare al punto nella lingua italiana che può essere usato per dividere due frasi oppure nella rappresentazione di un orario per dividere l'ora dai minuti.

5.3.2 Filtering

La fase di filtering viene solitamente effettuata sui documenti per rimuovere un insieme di parole. Un possibile filtro consiste nella rimozione delle *stop words*. Quest'insieme contiene un elenco di termini che non devono essere considerati poiché non sono rilevanti per i fini che s'intendono realizzare, come numeri, articoli, congiunzioni, preposizioni, avverbi, caratteri speciali e tutti quei vocaboli che, dopo una scrupolosa analisi delle frequenze, risultano essere comuni.

5.3.3 Stemming

L'obiettivo di questa tecnica è quello di ridurre le diverse forme di una parola alla forma base, senza tenere conto però dell'analisi morfologica, delle regole grammaticali e del contesto. Tutto ciò può generare un immenso numero di errori causati dall'ambiguità del linguaggio. In lingua inglese l'algoritmo di stemming più utilizzato è il *l'inflectional stemming*, dove le forme plurali sono convertite in quelle singolari ("cars" diventa "car") e le forme coniugate dei verbi trasformate nella relativa forma base ("fished" diventa "fish"). Questa tecnica è tipicamente realizzata tramite la rimozione dei suffissi.

5.3.4 Lemmatization

Simile allo stemming, tale tecnica ha lo scopo di ridurre le diverse forme di una parola alla forma base detta radice o "lemma", utilizzando l'analisi morfologica delle parole. Si cerca di raggruppare le varie forme inflesse di una parola in modo che possano essere analizzate come un unico elemento. In parole semplice i metodi

di lemmatization cercano di mappare le forme verbali all'infinito e i sostantivi al singolare. Tipicamente stemming e lemmatization non sono utilizzate entrambi, si decide quale dei due sia opportuno utilizzare in base alle caratteristiche del testo.

5.4 Rappresentazione del testo

Una volta terminata la fase di preprocessing è necessario trasformare i documenti in rappresentazioni vettoriali. Questo processo è chiamato *feature extraction* o più semplicemente vettorizzazione, ed è un primo passo essenziale verso un'analisi linguistica consapevole.

L'obiettivo è quello di dare come input ad algoritmi di data mining un vettore bidimensionale dove le righe rappresentano i token e le colonne rappresentano caratteristiche dei token. Per questo motivo, è necessario fare un cambiamento critico nel modo in cui si pensa al linguaggio, si passa da una sequenza di parole ad una serie di punti che occupano uno spazio semantico di grandi dimensioni. I punti nello spazio possono essere vicini o distanti tra loro, strettamente raggruppati o distribuiti uniformemente. Lo spazio semantico è quindi mappato in modo tale che documenti con significati simili siano più vicini tra loro e quelli che sono diversi siano più distanti. Codificando la somiglianza come distanza, possiamo iniziare a ricavare le componenti primarie dei documenti e a tracciare i confini decisionali nel nostro spazio semantico.

Si definiscono allora una raccolta di documenti $D = \{d_1, d_2, ..., d_D\}$ e l'insieme di parole e o termini distinti nella collezione $V = \{w_1, w_2, ..., w_v\}$, chiamato vocabolario. L'obiettivo della vettorizzazione è creare un vettore in cui ogni parola presente nel vocabolario è rappresentata da una variabile che indica il peso, l'importanza della parola nel documento. La più semplice codifica dello spazio semantico è il modello della *bag of words*, la cui intuizione primaria è che il significato e la somiglianza tra due parole siano codificati nel vocabolario.

Si parte dall'idea di rappresentare un documento come vettore in cui ogni chiave, o indice, rappresenta un token, e ogni cella rappresenta una caratteristica del token nel documento di testo dato. Quindi un testo è rappresentato come il set dei token che lo compongono, come riportato in figura Figura 5.2. Tale rappresentazione non si cura del significato, del contesto e dell'ordine in cui i token appaiono. Ad ogni parola è possibile associare una caratteristica di interesse, che rappresenta la codifica vettoriale del token. I principali tipi di codifica vettoriale sono: *one-hot*, basata sulla frequenza e *TF-IDF* [4].

5.4.1 Vettorizzazione One-hot

La vettorizzazione one-hot è una codifica vettoriale booleana, in cui alla posizione corrispondente ad un dato token si trova un valore uno se il token è presente nel

documento e zero se non è presente. Si rappresenta semplicemente la presenza o l'assenza del token nel documento, come riportato in Figura 5.3.

Figura 5.2: Esempio di bag of words

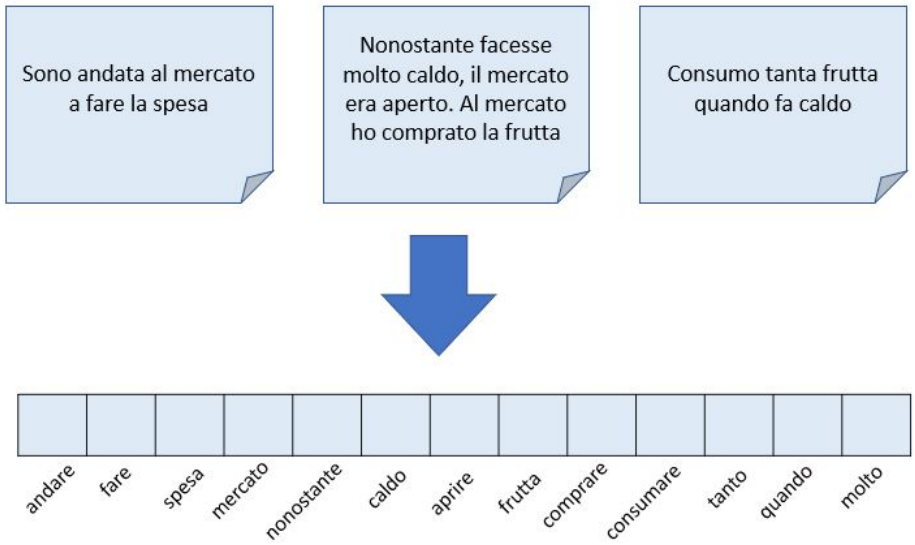
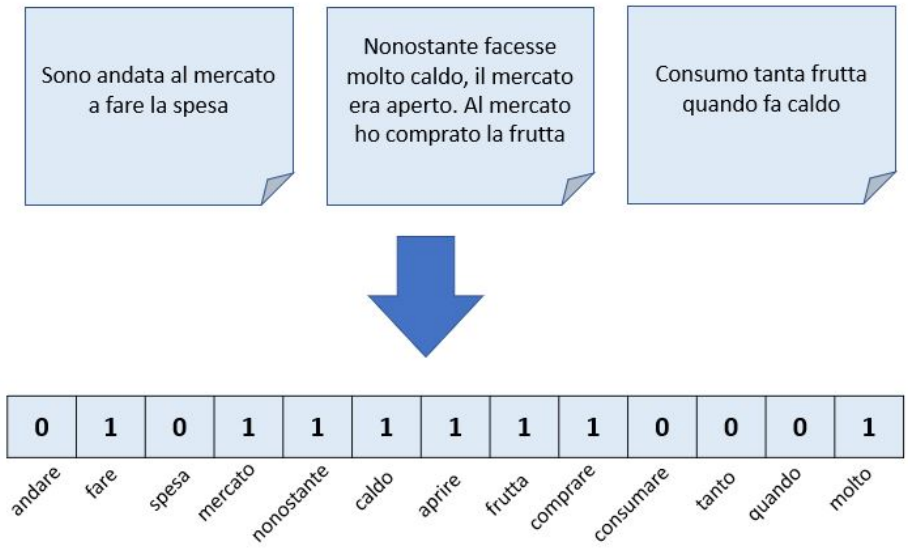


Figura 5.3: Esempio di vettorizzazione One-hot



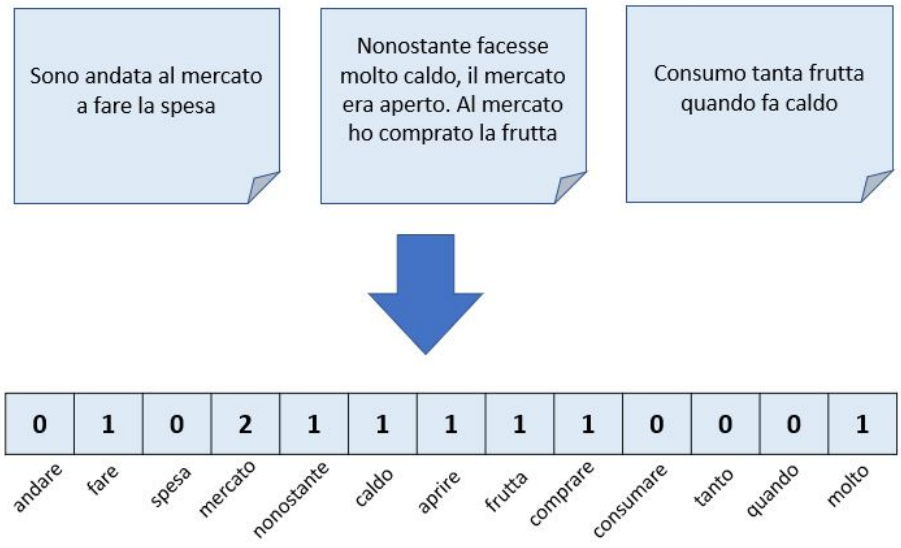
La vettorizzazione one-hot è più efficace per documenti molto piccoli (frasi, tweet) che non contengono molti elementi ripetuti.

5.4.2 Vettorizzazione basata sulla frequenza

Per realizzare la vettorizzazione di un corpus si rappresenta ogni documento come un vettore la cui lunghezza è pari alla grandezza del vocabolario del corpus e si riempie il vettore con la frequenza con cui il token appare nel documento. In questo schema di codifica, ogni documento è rappresentato tramite una bag of words in cui alla posizione corrispondente ad un dato token troveremo una rappresentazione della frequenza del token all'interno del testo, un esempio è riportato in Figura 5.4.

La rappresentazione della frequenza può essere un conteggio diretto oppure può essere normalizzata, nell'ultimo caso ogni token è ponderato dal numero totale di token presenti nel documento. I metodi di codifica basati sulla frequenza soffrono del problema del *long tail*, ovvero i token che si verificano molto spesso sono di ordini di grandezza più "significativi" di altri meno frequenti, che rappresentano una coda lunga di valori non considerati.

Figura 5.4: Esempio di vettorizzazione basata su frequenza



5.4.3 Vettorizzazione Term Frequency–Inverse Document Frequency

Nelle vettorizzazioni precedenti il problema più evidente era quello dello sbilanciamento nella distribuzione dei token. La soluzione migliore è quella di considerare la frequenza relativa o la rarità dei token nel documento rispetto alla loro frequenza in altri documenti.

La vettorizzazione basata su *term frequency–inverse document frequency (TF-IDF)*, normalizza la frequenza dei token in un documento rispetto alla loro frequenza nel resto del corpus. Questa normalizzazione diminuisce il peso dei termini che si verificano più frequentemente nella raccolta di documenti, facendo in modo che il contenuto dei documenti sia maggiormente rappresentato da parole distinte che hanno frequenza medio-bassa nella raccolta. È necessario però dare delle definizioni matematiche per comprendere appieno tale schema.

Si individuino innanzitutto due termini: *term frequency (TF)* e *inverse document frequency (IDF)*. Si riprenda la definizione fatta in precedenza di $D = \{d_1, d_2, ..., d_D\}$, raccolta di documenti, e $V = \{w_1, w_2, ..., w_v\}$, insieme di token distinti nella collezione.

Si definisce term frequency, indicata con $f_d(w)$, la frequenza del token $w \in V$ nel documento $d \in D$, con la seguente formula:

$$f_d(w) = \frac{n}{|d|},$$

in cui

- n è il numero di occorrenze del token w nel documento d;
- $|d|$ indica il numero di token presenti in d.

La term frequency è un valore espresso considerando un solo documento e rappresenta la frequenza del token all'interno del documento.

Si definisce *inverse document frequency (IDF)* con la seguente formula:

$$idf(w) = \log \frac{|\mathcal{D}|}{f_{\mathcal{D}}(w)},$$

in cui:

- $|\mathcal{D}|$ è il numero totale di documenti presenti nel corpus;
- $f_{\mathcal{D}}(w)$ indica il numero di documenti contenti il token w, $f_{\mathcal{D}}(w) = |d \in \mathcal{D} : w \in d|$.

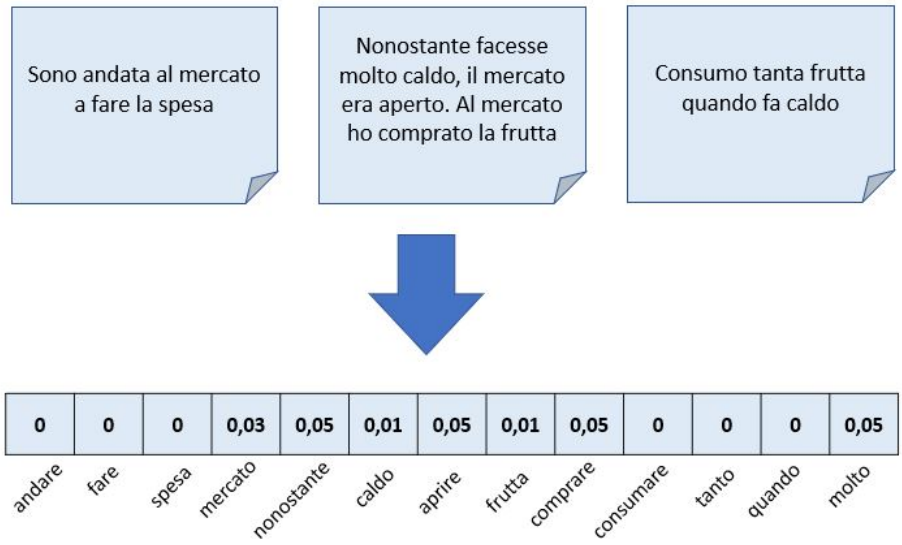
Il calcolo del IDF vuole mettere in evidenza che una parola non è di grande utilità se compare in tutti i documenti.

La term frequency–inverse document frequency è il prodotto di questi due termini ed è rappresentato dalla seguente formula finale:

$$q(w) = f_d(w) * \log \frac{|\mathcal{D}|}{f_{\mathcal{D}}(w)}.$$

La combinazione di TF e IDF ha lo scopo di fornire un peso per il token che dipenda sia dalla rilevanza del token in un documento, sia dall'inverso dell'uso del token nell'intero corpus, come riportato nell'esempio in figura Figura 5.5.

Figura 5.5: Esempio di vettorizzazione TF-IDF



5.5 Unsupervised Learning on Text

Spesso l'unsupervised learning può essere incredibilmente utile per l'analisi di testi esplorativi. I corpus tipicamente non arrivano con etichette pronte per la classificazione e quindi l'unica scelta possibile è adottare un approccio non supervisionato. Il clustering e il topic modeling sono i due algoritmi di apprendimento non supervisionato comunemente utilizzati nel contesto e sono analizzati nel dettaglio di seguito.

5.6 Clustering

Gli algoritmi di clustering mirano a scoprire la struttura latente o i temi nascosti in dati non etichettati, utilizzando le caratteristiche del dato per organizzare le istanze in gruppi significativamente diversi.

Il clustering, in ambito text mining, consiste sostanzialmente nell'individuare gruppi di documenti simili in una raccolta di documenti. I documenti che sono simili tra loro sono raggruppati insieme e i gruppi che ne risultano descrivono ampiamente i temi, gli argomenti e gli schemi generali all'interno del corpus. I modelli ottenuti attraverso il clustering possono essere *exclusive*, quando i gruppi non si sovrappongono affatto, oppure *non-exclusive*, quando c'è molta somiglianza e i documenti sono difficili da distinguere. In entrambi i casi, i gruppi risultanti rappresentano un modello del contenuto di tutti i documenti. Il raggruppamento del testo può avvenire a diversi livelli di granulometria, ovvero i cluster possono essere documenti, paragrafi, frasi o termini.

5.6.1 Metriche di similarità

Dopo aver rappresentato il corpus attraverso uno spazio semantico i documenti sono definiti come punti nello spazio, quindi la relativa vicinanza dei due documenti è una misura della loro somiglianza. Questo ci permette di determinare la similarità tramite l'utilizzo di una funzione ben definita.

Le metriche di distanze più utilizzare sono [12]: *euclidean distance*, *cosine distance* e *Jaccard distance*.

L'euclidean distance è la lunghezza del segmento che collega i due punti, più i due punti sono vicini nello spazio più i documenti sono simili. Dati due documenti d_a e d_b, rappresentati tramite i loro vettori di termini $\vec{t_a}$ e $\vec{t_b}$, l'euclidean distance è definita come:

$$D_E(\vec{t_a}, \vec{t_b}) = \sqrt{\sum_{t=1}^{m} |w_{t,a} - w_{t,b}|^2},$$

in cui $T = \{t1, ..., tm\}$ è il set di termini presenti nei documenti e w_t, a e w_t, b sono i pesi calcolati con la TF-IDF.

La *cosine distance* è una misura di distanza tra vettori in cui si utilizza il coseno dell'angolo tra i due vettori per valutare in che misura essi condividono lo stesso orientamento. Quanto più paralleli sono i due vettori, tanto più simili saranno i documenti. Date le rappresentazioni vettoriali dei due documenti $\vec{t_a}$ e $\vec{t_b}$, la cosine distance è data da:

$$SIM_C(\vec{t_a}, \vec{t_b}) = \frac{\vec{t_a} \cdot \vec{t_b}}{|\vec{t_a}| \times |\vec{t_b}|},$$

in cui $T = \{t1, ..., tm\}$ è il set di termini presenti nei documenti. *Jaccard distance*. La distanza definisce la somiglianza tra insiemi finiti come il quoziente della loro intersezione e la loro unione. Gli insiemi in questo caso sono i documenti, definiti come insiemi di token. Il coefficiente di jaccard è definito come:

$$SIM_J(\vec{t_a}, \vec{t_b}) = \frac{\vec{t_a} \cdot \vec{t_b}}{|\vec{t_a}|^2 + |\vec{t_b}|^2 - \vec{t_a} \cdot \vec{t_b}}.$$

Il coefficiente di Jaccard è una misura di similarità e assume valori tra zero e uno, in cui uno significa che i documenti sono identici e zero che i documenti sono differenti.

5.6.2 Hierarchical Clustering

Gli algoritmi di *hierarchical clustering* costruiscono un gruppo di cluster che può essere rappresentato come una gerarchia di cluster. La gerarchia può essere costruita in modo top-down o bottom-up.

Nel caso di costruzione top-down o *divisive*, i dati vengono gradualmente divisi, iniziando con tutte le istanze come unico gruppo e finendo come singole istanze. Nel caso di costruzione bottom-up o *agglomerative*, i cluster iniziano come singole istanze che si aggregano iterativamente per similarità fino a quando tutti appartengono ad un unico gruppo.

Gli algoritmi di clustering gerarchico sono basati sulla distanza, cioè utilizzando una funzione di similarità per misurare la vicinanza tra i documenti di testo.

5.6.3 Partitive Clustering e K-means Clustering

Il *k-means clustering* è uno degli algoritmi di *partitive clustering* ampiamente utilizzato nel text mining. Gli algoritmi di partitive clustering separano i documenti in gruppi i cui punti condividono la massima somiglianza, definita da una certa distanza metrica. I cluster individuati sono rappresentati da un vettore centrale, dentro il centroide che rappresenta un valore aggregato, ad esempio medio o mediano, di tutti i documenti del cluster e permette quindi di descrivere i documenti di quel cluster.

Il k-means clustering, partiziona n documenti in k cluster rappresentativi. L'algoritmo è diviso nelle seguenti fasi [3]: si definiscono un set di documenti D, una misura della similarità S, numero k di cluster, si selezionano in modo casuale k punti come centroidi di partenza, finché non si ha convergenza, si assegnano i documenti ad un cluster in base al centroide più vicino, ovvero più simile.

5.7 Topic modeling

Il *topic modeling* è una tecnica di apprendimento non supervisionato per astrarre argomenti lessicali co-occorrenti in una collezione di documenti sulla base di calcoli statistico-probabilistici. Ogni algoritmo di topic modeling parte dal presupposto che i documenti siano costituiti da un numero fisso di *topics*, ovvero argomenti. Il modello valuta quindi la struttura di base delle parole all'interno dei dati e cerca di trovare i gruppi di parole che meglio rappresentano il corpus in base a tale vincolo. Il topic può essere caratterizzato come un insieme di parole con una data distribuzione.

L'idea principale del topic modeling è quella di creare un modello probabilistico per il corpus, in cui i documenti sono una distribuzione di probabilità su un insieme di topic ed un topic, è una distribuzione di probabilità su un insieme di token. Le due principali tecniche di topic modeling sono: *Latent Semantic Analysis (LSA)* e *Latent Dirichlet Allocation (LDA)*

5.7.1 Latent Dirichlet Allocation

Introdotta per la prima volta nel 2003 da Blei et al. [5], l'idea di base di questa tecnica è che i documenti sono rappresentati come una combinazione casuale di topic latenti, dove ogni topic è una distribuzione di probabilità sui token. Quindi i topic sono rappresentati come la probabilità che ognuno degli insieme dei token si verifichi ed i documenti sono a loro volta rappresentati in termini di una combinazione di questi topic.

L'algoritmo dietro LDA è descritto in parole semplici di seguito. Si definisce la bag of word ed ogni documento viene rappresentato come un vettore a n dimensioni, con n pari alla grandezza del vocabolario, in cui sono indicate le frequenze dei token. Occorre specificare quanti topic possono essere presenti nel corpus (non si discutono in questo elaborato le tecniche con le quali si possa determinare tale numero). In prima istanza l'algoritmo assegnerà ogni token ad uno o più topic temporanei. Queste assegnazioni vengono effettuate in modo semi-casuale, secondo una distribuzione di Dirichlet, ciò significa che se una parola appare n volte, essa può essere assegnata a diversi topic. Successivamente all'assegnazione iniziale, l'algoritmo in modo iterativo verifica e aggiorna le assegnazioni delle parole ai topic, analizzando ogni parola in ogni documento. Per ciascun token w, l'assegnazione al topic t nel documento d, ovvero il calcolo della probabilità che essa appartenga al documento, viene aggiornata in base a due criteri: la frequenza relativa con cui w appare in t e la frequenza assoluta dei restanti termini di t in d. Alla fine del ciclo di iterazioni la distribuzione si assesta e di conseguenza è possibile calcolare la presenza relativa dei topic in ciascun documento.

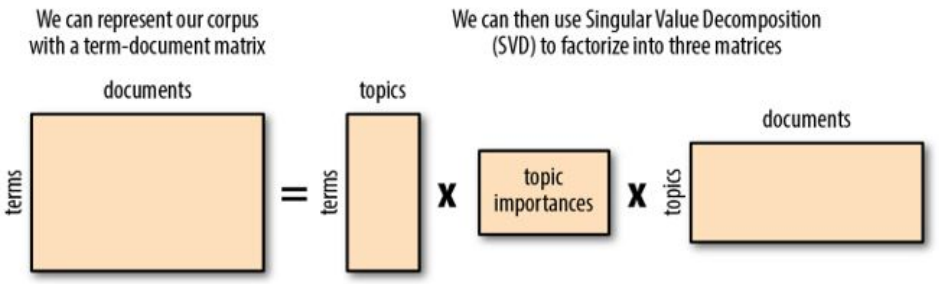
Una caratteristica unica dei modelli LDA è che gli argomenti non devono essere distinti, e le parole possono essere presenti in più argomenti; questo permette una sorta di relazione tra gli argomenti che è utile per gestire la flessibilità del linguaggio.

5.7.2 Latent Semantic Analysis

Latent Semantic Analysis (LSA) è un approccio di topic modeling suggerito per la prima volta come tecnica da Deerwester et al. [9] nel 1990. Questa tecnica individua gruppi di documenti contenenti gli stessi termini. LSA identifica i topic all'interno di un corpus creando una matrice sparsa *document-token*, in cui ogni riga rappresenta un token e ogni colonna indica un documento. Ogni valore della matrice corrisponde alla frequenza con cui il token indicato appare in quel documento e può essere normalizzato utilizzando TF-IDF.

Il *Singular Value Decomposition (SVD)* [10] viene applicato alla matrice *document-token* per fattorizzarla in matrici che rappresentano: la relazione tra topic e token, l'importanza dei topic e la relazione tra documenti e topic. Il risultato finale di tale tecnica saranno dunque due matrici: la matrice *topic-token*, che suddivide i topic in base alle loro componenti testuali, ovvero i token, e la matrice *document-topic*, che descrive i documenti in termini di topic (Figura 5.6). Utilizzando la matrice diagonale di importanza dei topic così ottenuta, possiamo identificare i topic che sono più significativi nel nostro corpus, e rimuovere dalla matrice le righe che corrispondono a topic meno importanti.

Figura 5.6: LSA



Fonte: Bengfort et al. [4]

Capitolo 6

Text Mining Applicato alle Recensione dei Pazienti

Dopo lo studio e l'implementazione del sistema di raccomandazione si è deciso di integrare il lavoro svolto sul sistema effettuando un'analisi sulle recensioni degli utenti sulle diverse strutture. L'obiettivo è quello di ampliare la conoscenza dei contenuti presenti nel dataset, al fine di esaminare la percezione degli utenti per individuarne le necessità e le preferenze.

I commenti degli utenti sono un tipo di dato non strutturato, ovvero il dato non è etichettato o classificato. Su questa tipologia di dato è possibile applicare solo tecniche di unsupervised learning. L'unsupervised learning cerca una relazione tra i dati per capire se e come essi siano collegati tra di loro. Questo modello, non contenendo alcuna informazione preimpostata, è chiamato a creare "nuova conoscenza".

6.1 Fonte dato: Qsalute

La fonte dato utilizzata per lo sviluppo delle successive analisi è il portale Qsalute. Come si è accennato nei capitoli precedenti da tale portale è possibile estrapolare informazioni di diversa granularità come: una valutazione sintetica della struttura, che tiene conto delle recensioni di tutti gli utente ed una specifica recensione sulla struttura espressa da un utente.

Nella prima parte del lavoro di tesi si è analizzata ed utilizzata la valutazione generale della struttura, in questa fase invece si è data importanza alle recensioni dei singoli utenti. Note le strutture di riabilitazione sul territorio di interesse, si è provveduto ad ottenere la lista delle recensioni relative ad ogni singola struttura.

Un esempio di recensione presente su Qsalute è riportata in Figura 6.1. Si può osservare la presenza di:

- una parte destinata a contenere la recensione testuale, in cui bisogna indicare il titolo della recensione e il contenuto della recensione (in figura tale parte corrisponde al riquadro rosso);
- un riquadro contenti le valutazioni a cinque punti espresse su ognuno degli ambiti analizzati in precedenza;
- il nome dell'utente, che non corrisponde ad un profilo quindi può essere anche anonimo;
- la patologia che è stata trattata durante il ricovero/visita del paziente in struttura;
- la data in cui la recensione è stata pubblicata.

Si è osservato già in precedenza la caratteristica di QSalute che definisce gli ambiti sui quali esprimere una valutazione numerica ovvero: competenza, assistenza, pulizia e servizi.

Figura 6.1: Esempio di recensione QSalute

Fonte: QSalute [20]

6.2 Tool utilizzati

Nelle varie fasi di preprocessing e applicazione delle tecniche di text mining si sono utilizzati i tool descritti brevemente di seguito.

Scikit-Learn [19] è un'estensione di SciPy (Scientific Python) che fornisce un'API per l'applicazione di tecniche di machine learning. Essa combina alte prestazioni con la facilità d'uso per analizzare insiemi di dati di piccole e medie dimensioni. Aperto e commercialmente utilizzabile, fornisce un'unica interfaccia per molti modelli di regressione, classificazione, clustering e riduzione dimensionale, oltre ad una utility per la cross-validazione e la regolazione dell'iperparametrizzazione.

NLTK [15], acronimo di *Natural Language Tool-Kit*, è una suite di librerie e programmi per l'analisi simbolica e statistica nel campo dell'elaborazione del *Natural Language Processing (NLP)*, scritta principalmente per lingua inglese in linguaggio Python. Nato originariamente come strumento accademico per l'insegnamento del NLP, contiene esempi di corpus, risorse lessicali e algoritmi di elaborazione linguistica che permettono ai programmatori Python di iniziare rapidamente l'elaborazione dei dati di testo.

Gensim [24] è una libreria efficiente che si concentra sulla modellazione semantica non supervisionata del testo. Originariamente progettata per trovare la somiglianza tra i documenti, ora include altre librerie e consente l'accesso a metodi di topic modeling per tecniche di unsupervised learning.

TreeTagger[1] è uno strumento per annotare il testo con informazioni sui lemmi. È stato sviluppato da Helmut Schmid ed è stato integrato con successo nel tempo per essere utilizzato in diverse lingue.

WordCloud[2] è una libreria che permette la realizzazione di *word cloud*.

6.3 Preprocessing del dato

È necessaria una fase di pulizia e di pre-elaborazione dei dati, per convertire i dati testuali in un formato appropriato per gli algoritmi di text mining analizzati. La fase di preprocessing si divide in: *tokenization*, *filtering* e *lemmatization*, come riportato in Figura 6.2

6.3.1 Tokenization

Il token è l'unità atomica di analisi del testo, essere composto da: un singolo carattere, una singola parola o termini composti da più parole. Nel seguente lavoro di tesi si è deciso di utilizzare le parole come token poiché esse portano un'adeguata ricchezza semantica. Il documento dunque è diviso in token in cui ogni token rappresenta

[1] TreeTagger - https://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger

[2] WordCloud - https://github.com/amueller/wordcloud

Figura 6.2: Fase di preprocessing del testo

una parola del documento. Tale operazione è eseguita tramite il tool NLTK, che dato il corpus, la tipologia di token scelto, trasforma la frase di un documento in un vettore di token.

6.3.2 Filtering

La fase di filtering ha lo scopo di rimuovere dal documento le cosiddette *stop words*, ovvero un elenco di termini che non devono essere considerati poiché non sono rilevanti per i fini che s'intendono realizzare, come numeri, articoli, congiunzioni, preposizioni, avverbi, caratteri speciali ecc.

Durante tale fase si sono rimossi tutti i simboli di punteggiatura e tutti i caratteri maiuscoli sono stati riscritti in minuscolo. Anche in questo caso si è utilizzato il tool NLTK che contiene una libreria di stop word italiane.

6.3.3 Lemmatization

L'obiettivo è quello di ridurre le diverse forme di una parola alla forma base detta radice o "lemma", utilizzando l'analisi morfologica delle parole, in questo modo ad esempio tutti verbi sono riportati alla forma infinita. Per fare tale operazione si è utilizzato il tool TreeTagger che data una parola è in grado di determinare la radice della parola e la corrispondente parte del discorso, ovvero se la parola è un verbo, avverbio, pronome ecc.

6.4 Rappresentazione del testo

Nel seguente lavoro di tesi sono utilizzate due delle possibili rappresentazioni: quella basata sulla frequenza e quella basata su TF-IDF.

6.5 Unsupervised Learning Methods

Con l'obiettivo di svolgere una fase conoscitiva preliminare, durante la fase di preprocessing, si sono analizzate le parole più frequenti presenti nel corpus, per avere un'idea dei contenuti presenti nel dataset. In questa fase si è utilizzata la libreria *wordcloud*, che non richiede dataset già pre-erabolati, dato un dataset divide il corpus in parole e ne calcola la frequenza all'interno dell'intero corpus, creando automaticamente una rappresentazione vettoriale basata sulla frequenza.

In particolare, in Figura 6.3 è riportata la word cloud delle parole più frequenti dopo aver applicato tokenization e filtering. Si può osservare che tutte le parole sono riportate in minuscolo e sono presenti vari verbi in vari tempi. Tra le parole presenti quelle più di interesse sono: personale, struttura, grazie, medico, riabilitazione, professionalità e paziente.

Si è eseguita un'ulteriore analisi delle parole frequenti dopo aver applicato la lemmatization, in Figura 6.4 è stata riportata la word cloud corrispondente. Si può notare come in questo caso siano presenti solo i verbi all'infinito, mentre la lista delle parole più frequenti non è cambiata molto rispetto al precedente caso.

Dopo questa analisi esplorativa è necessario preparare il dateset in modo tale da poterlo usare come input degli algoritmi di clustering e topic modeling. I passaggi descritti di seguito sono stati svolti con l'utilizzo del tool *gensim* che mette a disposizione una serie di metodi per il preprocessing e la rappresentazione dei dati.

Si è determinato il vocabolario, ovvero l'insieme di token distinti all'interno del corpus. Si è determinata la rappresentazione dei documenti trasformandoli in vettori, realizzando la *bag of words*, utilizando il vocabolario sopra definito ed associando ad ogni parola distinta un valore interno, che d'ora in poi rappresenterà la parola. La rappresentazione vettoriale utilizzata è quella basata su TF-ID, con la quale ad ogni parola si associa un valore che rappresenta la frequenza della parola in un documento, normalizzata rispetto alla frequenza della stessa nel resto del corpus. Con l'utilizzo del TF-IDF è possibile diminuire l'importanza dei termini che si verificano più frequentemente nella raccolta di documenti, facendo in modo che il contenuto dei documenti sia maggiormente rappresentato da parole che hanno maggiore importanza localmente al singolo documento. La word cloud riportata in Figura 6.5 rappresenta quanto appena descritto, le parole rilevanti non sono più quelle determinate con la frequenza, che non compaiono più. In particolare è possibile osservare la presenza di parole quale: epilogo, qualitativo, gradito, neuroriabilitazione e relazione.

Figura 6.3: Word cloud parole frequenti del dataset senza lemmatization

Stabilita la rappresentazione vettoriale del testo è possibile applicare le tecniche di unsupervised learning, ma è necessario definire il concetto di documento. All'interno del seguente lavoro di tesi si sono utilizzati due approcci: considerare il documento come l'insieme dei commenti relativi ad una struttura e considerare il documento come il singolo commento.

6.6 Clustering

Terminata la fase di rappresentazione del testo, è possibile applicare le tecniche di clusterig, al fine di determinare gruppi di documenti simili tra di loro. Il tool utilizzato è stato *Scikit-Learn*, che mette a disposizione un metodo contenente la realizzazione del modello k-means. Tale metodo necessita di ricevere come parametri: il numero di cluster che si vuole individuare ed il corpus vettorizzato. In questo caso si sono utilizzate sia la rappresentazione basata su frequenza sia quella basata su TF-IDF, al fine di confrontare i risultati ed individuare la soluzione migliore. La metrica di similarità utilizzata per costruire il modello è stabilita dal tool stesso, che

Figura 6.4: Word cloud parole frequenti del dataset con lemmatization

applica la *euclidean distance*, in cui la similarità è data dalla lunghezza del segmento che collega i due punti, più i due punti sono vicini nello spazio più sono simili i documenti. Nell'analisi svolta si è scelto di adottare un numero di cluster pari a tre. Questo numero permette di ottenere un ottimale raggruppamento di documenti. Aumentando il numero di cluster, essi risultano difficilmente interpretabili.

I risultati ottenuti dal modello, utilizzando la vettorizzazione basata su frequenza, sono riportati in Tabella 6.1 e in Tabella 6.2. In particolare nella prima tabella si sono riportati i risultati considerando il documento contente un singolo commento e nella seconda tabella invece considerando il documento contente un insieme di commenti relativi ad una struttura. I cluster sono rappresentati riportando i termini più frequenti all'interno dello specifico cluster.

Si può ipotizzare che il "Cluster 1" rappresenta il paziente e il personale, il "Cluster 2" rappresenta il paziente e la malattia e il "Cluster 3" rappresenta la riabilitazione.

I risultati ottenuti dal modello, utilizzando la vettorizzazione basata su TF-IDF, sono riportati in Tabella 6.3 e in Tabella 6.4. In particolare nella prima tabella si sono riportati i risultati considerando il documento contente un singolo commento

Figura 6.5: Word cloud parole determinato utilizzando tf-idf

Tabella 6.1: Cluster ottenuti con vettorizzazione basata su frequenza e documento contente un singolo commento

Cluster 1	Cluster 2	Cluster 3
stare	personale	ingessare
cura	essere	oggi
problema	medico	cosa
poco	struttura	stato
molto	patologia	trattare
sembrare	degenza	fisioterapico
cortese	giorno	infermieristico
seguire	molto	informare
professionalità	grazie	bene
paziente	dopo	lavoro

e nella seconda tabella invece considerando il documento contente un insieme di commenti relativi ad una struttura. I cluster sono rappresentati riportando i termini con TF-IDF più alti all'interno dello specifico cluster. Si può innanzitutto osservare

Tabella 6.2: Cluster ottenuti con vettorizzazione basata su frequenza e documento contente un insieme di commenti

Cluster 1	Cluster 2	Cluster 3
ottimo	essere	informare
trovare	stare	problema
problema	personale	ancora
professionalità	struttura	oggi
paziente	grazie	seduta
sembrare	ricoverato	fisioterapista
cortese	dopo	anziano
cura	medico	stare
poco	struttura	trattare
stare	molto	bene

che le parole che compaiono sono parole di media frequenza e quindi più inconsuete, inoltre i cluster trovati, considerando il documento come insieme di commenti su una struttura, sono più caratterizzati.

Si può ipotizzare che nella Tabella 6.3 il "Cluster 1" rappresenta il paziente e la malattia, il "Cluster 2" rappresenta la struttura e il "Cluster 3" rappresenta il paziente e le sue sensazioni.

Si può ipotizzare che nella Tabella 6.4 il "Cluster 1" rappresenta la percezione del paziente sul personale, il "Cluster 2" rappresenta la struttura ed il suo contenuto, come luoghi e operatori e il "Cluster 3" le sensazioni del paziente.

Tabella 6.3: Cluster ottenuti con vettorizzazione TF-IDF e documento contente un singolo commento

Cluster 1	Cluster 2	Cluster 2
ringraziare	curare	me
potere	clinico	sensibile
prima	medico	chiedere
problema	tipicamente	base
ottimo	serietà	barriera
cura	fisioterapico	distinguere
anziano	operatore	altro
lavoro	pratico	inizio
giorno	reparto	nessuno

Tabella 6.4: Cluster ottenuti con vettorizzazione TF-IDF e documento contente un insieme di commenti

Cluster 1	Cluster 2	Cluster 3
tutto	esame	me
cortese	bar	staff
professionalità	primario	cura
seguire	struttura	medicina
paziente	stanza	abbandonato
molto	personale	crisi
grazie	trovare	praticamente
essere	medico	mamma
ridere	sito	soggiorno

6.7 Topic modeling

Dopo aver analizzato le tecniche di clustering si è scelto di esplorare le tecniche di topic modeling, per astrarre argomenti co-occorrenti nei documenti. Si è utilizzato il tool gensim che espone metodi per la creazione di entrambi i modelli di seguito descritti. Sono state utilizzate *Latent Dirichlet Allocation*, i cui risultati sono riportati in Tabella 6.5 e in Tabella 6.6, e *Latent Semantic Analysis*, i cui risultati sono riportati in Tabella 6.7 e in Tabella 6.8. Per ogni tecnica utilizzata, la prima tabella riporta i risultati ottenuti considerando il documento contente un singolo commento e la seconda tabella invece considerando il documento contente un insieme di commenti relativi ad una struttura.

I risultati ottenuti sono coerenti con quelli generati tramite clusterizzazione. In particolare, si può osservare la presenza di topic legati principalmente a termini quali struttura, pazienti e personale. Da un'analisi dei commenti degli utenti, le tematiche affrontate sono molto simili nelle diverse recensioni e sono descrivibili con i termini di cui sopra. Le tematiche risultano però molto omogenee e rendono difficile ottenere un gran numero di argomenti distinti: questa osservazione è confermata dall'analisi del comportamento degli algoritimi di topic modeling, i cui risultati sono meglio interpretabili, nel caso in esame, utilizzando un numero piccolo di topic.

Tabella 6.5: Topic ottenuti utilizzando LDA e documento contenente un singolo commento

Topic 1		Topic 2		Topic 3	
Peso	Parola	Peso	Parola	Peso	Parola
0.011	personale	0.012	personale	0.010	struttura
0.010	medico	0.011	essere	0.010	stato
0.010	molto	0.010	giorno	0.009	essere
0.007	essere	0.009	medico	0.008	medico
0.007	dovere	0.009	struttura	0.008	dopo
0.007	paziente	0.007	grazie	0.008	dovere
0.007	struttura	0.007	fare	0.007	personale
0.006	reparto	0.006	ospedale	0.007	giorno
0.006	dopo	0.006	paziente	0.007	dire
0.006	giorno	0.006	trovare	0.006	molto

Tabella 6.6: Topic ottenuti utilizzando LDA e documento contenente un insieme di commenti

Topic 1		Topic 2		Topic 3	
Peso	Parola	Perso	Parola	Peso	Parola
0.010	essere	0.013	personale	0.009	medico
0.008	personale	0.012	struttura	0.009	essere
0.008	molto	0.010	medico	0.009	personale
0.007	medico	0.010	essere	0.008	giorno
0.007	giorno	0.008	stato	0.008	dopo
0.006	dopo	0.008	giorno	0.007	struttura
0.006	struttura	0.007	paziente	0.007	grazie
0.006	reparto	0.007	molto	0.007	stato
0.006	paziente	0.006	dovere	0.006	dott
0.005	stato	0.006	trovare	0.006	dovere

Tabella 6.7: Topic ottenuti utilizzando LSA e documento contenente un singolo commento

Topic 1		Topic 2		Topic 3	
Peso	Parola	Perso	Parola	Peso	Parola
0.251	giorno	0.396	personale	-0.325	giorno
0.242	essere	0.225	dott	0.290	molto
0.216	medico	0.221	grazie	0.236	persona
0.199	personale	-0.196	dire	-0.203	medico
0.194	dopo	0.175	struttura	0.194	personale
0.186	dovere	-0.169	ora	-0.185	stato
0.181	struttura	-0.165	giorno	-0.163	dopo

Tabella 6.8: Topic ottenuti utilizzando LSA e documento contenente un insieme di commenti

Topic 1		Topic 2		Topic 3	
Peso	Parola	Perso	Parola	Peso	Parola
0.261	personale	-0.285	villa	0.218	dott
0.245	essere	-0.188	grazie	-0.179	riabilitazione
0.223	struttura	-0.180	personale	0.161	ringraziare
0.217	medico	-0.178	fisioterapista	-0.154	struttura
0.211	giorno	0.174	dopo	-0.154	essere
0.179	dopo	-0.167	molto	0.145	professionalità
0.166	stato	0.163	dire	0.136	grazie

Capitolo 7

Conclusioni e Sviluppi Futuri

Il lavoro descritto nei capitoli precedenti, rappresenta la progettazione e la realizzazione di un sistema di raccomandazione di strutture riabilitative. Il sistema realizzato, noto il quadro clinico del paziente e le sue preferenze, determina una compatibilità tra le caratteristiche della struttura e le specifiche indicate, generando una classifica. Dopo un'attenta analisi, si sono individuati come parametri su cui l'utente può esperire delle preferenze, la distanza della struttura riabilitativa dalla sua posizione, la valutazione che la struttura ha ricevuto e la presenza di servizi nelle vicinanze.

Dai risultati riportati è possibile notare come: cambiando l'attività di riabilitazione e la specializzazione necessarie, alcune strutture non compaiono più nella classifica, mentre cambiando l'ordine d'importanza degli input o la posizione dell'utente la classifica si aggiorna coerentemente, mostrando un diverso ordine delle strutture.

Per arricchire le informazioni sulle strutture nel sistema sviluppato si è eseguita un'analisi sui commenti presenti sul portale QSalute, al fine di esaminare la percezione dell'utente ed individuarne le necessità. Dai risultati ottenuti con il clustering si è notato che i commenti possono essere raggruppati secondo tre tematiche: rapporto paziente-personale, rapporto del paziente con la malattia e opinioni in generale sulla struttura. I risultati ottenuti con il topic modeling confermano la presenza di temi quali: struttura, pazienti e personale, utilizzando un numero minore di topic, i temi risultano meglio caratterizzati.

Nel sistema realizzato sono stati scelti alcuni parametri su cui l'utente può esprimere preferenze, un ampliamento futuro potrebbe essere quello di valutare l'inserimento di ulteriori parametri al fine di migliorare la raccomandazione.

Il seguente lavoro di tesi focalizza la sua attenzione sull'ambito riabilitativo. Un opportuno sviluppo futuro del sistema è quello di considerare non solo il caso in cui sia necessaria la riabilitazione, ma anche altri ambiti di dimissioni. Interessante è,

per esempio, il caso delle Residenze sanitarie assistenziali, in cui le preferenze del paziente hanno un maggior peso.

Bibliografia

[1] Aggarwal, C. C. (2016): *Recommender systems: The Textbook*. Springer-Verlag New York Inc, 1st edition.

[2] Aggarwal, C. C. and C. Zhai (2012): *Mining Text Data*. Springer Science & Business Media.

[3] Allahyari, M., S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut (2017): "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques", *ArXiv Preprint ArXiv:1707.02919*.

[4] Bengfort, B., R. Bilbro, and T. Ojeda (2018): *Applied Text Analysis with Python: Enabling Language-aware Data Products with Machine Learning*. " O'Reilly Media, Inc.".

[5] Blei, D. M., A. Y. Ng, and M. I. Jordan (2003): "Latent dirichlet allocation", *Journal of Machine Learning research*, Vol. 3, No. Jan, pp. 993–1022.

[6] Burke, R. (2007): "Hybrid Web Recommender Systems", in B. P., K. A., and N. W. eds. *The Adaptive Web*. Springer, Berlin, Heidelberg, pp. 377–408.

[7] Chen, L., C.-M. Chan, H.-C. Lee, Y. Chung, and F. Lai (2014): "Development of a Decision Support Engine to Assist Patients with Hospital Selection", *Journal of Medical Systems*, Vol. 38, No. 6, p.5̃9.

[8] Colombo-Mendoza, L. O., R. Valencia-García, A. Rodríguez-González, G. Alor-Hernández, and J. J. Samper-Zapater (2015):

"RecomMetz: A Context-Aware Knowledge-based Mobile Recommender System for Movie Showtimes", *Expert Systems with Applications*, Vol. 42, No. 3, pp. 1202–1222.

[9] Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman (1990): "Indexing by Latent Semantic analysis", *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp. 391–407.

[10] Feldman, R. and J. Sanger (2007): *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

[11] Harry Zisopoulos, G. D. S. A., Savvas Karagiannidis (November 2008): "Content-Based Recommendation Systems".

[12] Huang, A. (2008): "Similarity Measures for Text Document Clustering", in *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand.* , Vol. 4, pp. 9–56.

[13] Isinkaye, F., Y. Folajimi, and B. Ojokoh (2015): "Recommendation Systems: Principles, Methods and Evaluation", *Egyptian Informatics Journal*, Vol. 16, No. 3, pp. 261–273.

[14] Kannan, S. and V. Gurusamy (2014): "Preprocessing techniques for text mining", in *Conference Paper. India.*

[15] Loper, E. and S. Bird (2002): "NLTK: The Natural Language Toolkit", *ArXiv Preprint cs/0205028*.

[16] Ministero della Salute (2014): "Patto per la Salute 2014-2016", http://www.salute.gov.it/portale/pattosalute/dettaglioContenutiPattoSalute.jsp?lingua=italiano&id=1299&area=pattoSalute&menu=vuoto, [Online; controllata Maggio-2019].

[17] Omotosho, A., O. Adegbola, and A. Adebo (2016): "A Patient-Based Hospital Referral Decision Support System", *International Journal of Computer Application*, Vol. 115, No. 10, pp. 38–43.

[18] OpenStreetMap contributors (2017): "Planet dump retrieved from https://planet.osm.org ", https://www.openstreetmap.org.

[19] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg

ET AL. (2011): "Scikit-Learn: Machine Learning in Python", *Journal of Machine Learning Research*, Vol. 12, No. Oct, pp. 2825–2830.

[20] QSALUTE (2008): https://www.qsalute.it, [Online; controllata 2008].

[21] RAIFER, M. (2012): "Overpass Turbo", https://github.com/tyrasd/overpass-turbo, [Online; controllata 2017].

[22] REGIONE PIEMONTE (2007): "Deliberazione della Giunta Regionale 2 aprile 2007, n. 10-5605", Deliberazione della Giunta Regionale.

[23] REGIONE PIEMONTE (2012): "Deliberazione della Giunta Regionale 28 marzo 2012, n. 27-3628", Deliberazione della Giunta Regionale.

[24] ŘEHŘEK, R. AND P. SOJKA (2011): "Gensim—Statistical Semantics in Python", *Statistical Semantics; Gensim; Python; LDA; SVD.*

[25] SALUNKE, A. B. AND S. L. KASAR (2015): "Personalized Recommendation System for Medical Assistance using Hybrid Filtering", *International Journal of Computer Applications*, Vol. 128, No. 9, pp. 6–10.

[26] SCHÄFER, H., S. HORS-FRAILE, R. P. KARUMUR, A. CALERO VALDEZ, A. SAID, H. TORKAMAAN, T. ULMER, AND C. TRATTNER (2017): "Towards Health (Aware) Recommender Systems", in *Proceedings of the 2017 International Conference on Digital Health.* , pp. 157–161, ACM.

[27] SEZGIN, E. AND S. ÖZKAN (2013): "A Systematic Literature Review on Health Recommender Systems", in *2013 E-Health and Bioengineering Conference (EHB).* , pp. 1–4, IEEE.

[28] SUKSOM, N., M. BURANARACH, Y. M. THEIN, T. SUPNITHI, AND P. NETISOPAKUL (2010): "A Knowledge-Based Framework for Development of Personalized Food Recommender System", in *Proc. of the 5th Int. Conf. on Knowledge, Information and Creativity Support Systems.*

[29] WIKIPEDIA (2008): "Shrinkage estimator — Wikipedia, The Free Encyclopedia", https://en.wikipedia.org/wiki/Shrinkage_estimator, [Online; controllata 6-February-2019].

[30] YUAN, X., J.-H. LEE, S.-J. KIM, AND Y.-H. KIM (2013): "Toward a User-Oriented Recommendation System for Real Estate Websites", *Information Systems*, Vol. 38, No. 2, pp. 231–243.