

POLITECNICO DI TORINO

Corso di Laurea Magistrale

in Ingegneria del Cinema e dei Mezzi di Comunicazione

Tesi di Laurea Magistrale

Text Mining extraction from videos
in a learning environment through
Educational Data Mining



Relatore

Prof.ssa Laura Farinetti

Co-Relatore

Prof. Wolfgang Müller

Candidato

Giovanni Filippo Caruso

Index

Abstract	i
Introduction	ii
Research questions	iv
1 - Educational Data Mining	1
1.1 Data Mining: a step in the origins of Educational Data Mining	1
1.1.1 Data Mining Techniques	2
1.2 Educational Data Mining:	3
1.2.1 Educational Data Mining Stackholders	5
1.2.2 Educational Data Mining Usages	6
1.2.3 Improvements.....	10
1.3 Text Mining.....	11
1.3.1 Text Mining Usages.....	12
2 – Designing an Educational Data Mining Tool	14
2.1 Framework	14
2.2 Technologies.....	15
2.3 Scraping the Web	16
2.3.1 Ruby	17
2.4 Conceptual Maps	18
2.4.1 Ontology Map	19
2.4.2 CMap	22
2.5 Natural Processing Language	22
2.5.1 N-grams.....	23
3 – Implementing an Educational Data Mining Tool	25
3.1 First Step: Data Inputs.....	25
3.1.1 Ontology map.....	25

3.1.2 Video Scripts	26
3.2 Second Step: Matches	28
3.3 Third Step: Filtering the matches	31
3.3.1 N-grams cardinality	31
3.3.2 Creation of bi-grams from the transcripts	32
3.3.3 Filtering with the COCA's dataset	32
3.3.4 Creation of N-grams dataset	33
4 – Study's Results	37
4.1 Evaluation of the contents	37
4.2 First Matches	39
4.3 Bigrams Dataset	39
4.4 Final Results	40
4.5 Evaluations	42
4.5.1 Experts Evaluation	42
4.5.2 Evaluation through Entities	43
4.5 Acknowledgements	46
5 – Conclusions	48
5.1 Improvements and future works	50
Bibliography	53
List of pictures	55

Abstract

The World Wide Web was born as a place for everyone and for sharing contents, according to the “Hacker Ethics”. During the past years access to Internet and related technologies improved, developing new challenges in society. According to these times new computer sciences raised such as Data Mining and Big Data.

As the society changes, learning changes as well. Computer sciences created new ways of learning more focused on the single user and his own possibilities and problems. Following this path some general computer sciences as Data Mining become specialized in Education. Educational Data Mining (EDM) is a science which studies data generated from educational environments.

It’s safe to say that YouTube is one of the oldest reasons why Internet spreaded between people. Entertainment and connection between people were and still are the main goals of the company. Right now, YouTube is a big videos container of knowledge of each kind and, most importantly, is a free tool for learning.

Discovering useful contents through the Web is a valid support for teachers and students for improving their teaching and learning skills. The technologies and methods provided by EDM and Text Mining can help through a deep research on meaningful information for support the people involved in the educational environment.

The reseach questions behind the study is: *how good can we evaluate amateur materials on the web, such as a Youtube’s tutorial, comparing it to the material given by experts and how can we avoid not useful data?*

Keywords:

Educational Data Mining, Conceptual Maps, Web Scapring, Text Mining.

Introduction

The thesis aims to study and to find new learning scenarios for teachers according to new technologies and possibilities of the world wide web, taking care of the basics of pedagogy recommendend by the experts of the University of Education of Weingarten.

Education Data Mining is an interdisciplinary new field of Computer Sciences strongly connected to education, the core and the purpose of the project is related to that new interesting science.

Mixing educational tools, such as conceptual maps, and free information already spreaded on the web, such as video-tutorials on YouTube, the project wants to study the best ways to find reliable resources trough user's generated free contents.

The reason behind the decision to use a service as YouTube is because it is a free on-line service, well knowed by everyone and already plenty of information about many topics like coding and languages. Often, the contents published on this platform is provided by amateur users of a certain topic so it's interesting to find how valuable are those tutorials for a curious user that want to improve his knowledge.

The starting point of the study is a conceptual map supplied and developed by Professor Wolfgang Müller and Academic Assistant Sandra Rebholz regarding the main concepts of programming. The map resumes in an ontological way the basics of programming.

After a research of a Youtube's tutorial that explains the concepts showed the map, the goal is to download the subtitles of the videos and find the word-matches between the conceptual map and the script of the videos.

The last part of the project is to evaluate the possible errors in the matches and go deeper for a better evalutation of the sentences, avoiding the presence of unrelated information.

The research is fully developed at the University of Education of Weingarten, a pedagogic university in which there are studies for implementing new practices for learning.

The first chapter of the thesis is about Educational Data Mining, the main topic of the study. The chapter starts with a presentation of Data Mining, which is the Educational Data Mining super-category, and follows with how far Educational Data Mining went so far and his tasks and methods.

The second chapter is the presentation of the tool and the framework adopted. Following there is an explanation of the technologies used for the development of the project.

The third chapter is about the developed project. All the phases of the data-processing are explained and showed. It includes limits and failures regarding the project and the different approaches to reach the different goals of the study.

The fourth chapter analyzes the results of the study, for each step. The chapter analyzes the final output and evaluates them according to the expectations and limits.

The fifth and last chapter includes the conclusions about the project and ideas about future work regarding this topic and this study, in particular.

Research questions

After a general view of the main themes of the study, it's important to find which reasons and objectives moved the will to create a tool for Text Mining from videos subtitles.

“Can we evaluate amateur materials on the web, such as a Youtube’s tutorial, and can we compare it to the material given by experts and how can we avoid not useful data?” is the main question of the study. This question was generated after several considerations about what can be helpful in the educational environment and what we can realize according to the available time and the resources holded.

The study is developed at University of Education of Weingarten in Germany. This university is a pedagogical school, so since the beginning the goal was to find a solution for improving a part of the educational process according to the new computer sciences.

A valid reason behind the choice of creating something that connect education and Web is that internet was conceived, according to the Hacker Ethics, as a place for everyone where people can share their information and build a community and, in my opinion, education shares part of these values. One of the goals of the project is to build something with data already available and totally free, in the same way education should be.

Often, students need support materials for a better understanding of a certain topic and the reasons behind that might be several: the professor doesn't provide a good documentation of the contents, the given material is not enough or totally understandable for the learner or the student needs a different explanation of the content. Learning process is a complicated interaction of two or more peers and the difference cognitive skills of the people generate a lot of challenges, for examples the same content cannot fit to all the groups of people. One of the goals of EDM is to personalize learning according to the necessities of the learner and this study tries to focus, partially, on this need. In fact, why we cannot take information already placed

somewhere in the web and propose it to the students? With a big amount of data available, it's possible that within this data there is a better or several explanations of a concept that make easier to the student the understanding of that concept.

After this general question, the following step is to define a first domain of application for studying the efficacy of this theory. The second question was: where we can find a big amount of educational contents? But the answer was too generical, so for limiting the domain was necessary to go deeper in the habits of the nowadays society. Videos are a big part of the contents in the Web and a lot of educational institutions agree that videos are a big part of the learning process: on-line courses are spreading all over the web. Following this trend, the question then became: can we extract text data from these videos for finding meaningful information?

Afterwards deciding which type of content was good to take in exam, still the domain was a huge amount of data and possibilities. A lot of universities and companies are creating dedicated on-line courses but often these courses are not available for everyone, for pricing and degrees of knowledge. So YouTube.com became the obvious choice for two main reasons: first of all, it's well known by everyone and, secondly, it's plenty of channels with educational purposes. It's worthy to mention that often these videos are made by amateur users so it's more challenging to validate the effectiveness of these contents and find explanations of the concept from people that tries to explain their knowledge to basic users, so without a requested degree of education.

Found a proper domain, the next question is: how can we evaluate this amateur material and find a right field of knowledge in which can we prototype the project? The answer is given by some of the experts of the University of Education of Weingarten. Dr. Wolfgang Mueller and Dr. Sandra Rebholz developed an ontology map regarding Ruby programming. The map is designed as a conceptual map and it shows all the relationships between the concepts of the general programming with a special attention to Ruby code language. Thanks to this tool, it is found the topic that the Youtube's tutorial should cover, totally or partially, and a reliable source for comparing the contents of the videos. The conceptual map furnished a first dictionary for the Text Mining study and it provides a solid structure of how the topics should be linked between them.

Following these steps, we had enough hypothesis to elaborate a proper research question. The possibility of developing a tool that can answer to this question is given by the already cited methods given by Data Mining, Educational Data Mining and Text Mining.

Chapter 1

Education Data Mining

1.1 Data Mining: a step in the origins of Educational Data Mining

Educational Data Mining is considered a specialized usage of Data Mining so it's worthed to introduce first Data Mining for a better understanding.

Data Mining is a Computer Science's subfield which studies the process and methodology of extracting data from huge amounts of it and discover and/or create relationships in large data sets. The usage of Data Mining, knowed as well as DM, is common in many fields such as Business, Health and Research. Combining different disciplines and sciences, the main goal of the DM is to provide meaningful information from raw data to predict the future, according to the patterns discovered through the data.

The spreading of Information Technologies has raised the necessity and opportunity of evaluating the data stored in databases around the world for generating meaningful output for companies and researchers. Even though Data Mining is a somehow recent science, nowadays is already a main component in decision making [1].

Data Mining combines a lot of different disciplines such as statistics, pattern recognition and databases scraping and his versatile usage allows to go deeper in other disciplines like health, education, business and all those fields who have big amount of data and they need to generate output from that. The mixture of DM with other study fields requests interdisciplinary knowledge from the experts.

Independently of the field of application of DM, three main steps are requested:

- Exploration: the data is taken by a database, filtered and converted to the needed form;

- Pattern identification: the goal is to find relationships between data and choose or find patterns inside them;
- Deployment: the processed data is disposed according to the desired output.

1.1.1 Data Mining Techniques

For extrapolating knowledge from databases, according to Data Mining, there are several methods which can be used and all these methods are used for Educational Data Mining purposes as well.

These techniques are sustained by different algorithms and often they are mixed between them for providing new hybrid solutions, according the desired output and goal of the study. It's relevant to introduce, at least, some of them [2].

- Classification

The goal of classification is to classify some existent examples and develop a model for analyzing big amount of data according to the previous samples. Usually, it is one of the most common used between the DM techniques.

- Clustering

Clustering groups data in, so called, "cluster" which are groups of data with one or more similarities. All the clusters, usually, share in a large way some common features but the data inside each cluster shares a closer and deeper meaning between his peers.

- Anomaly Detection

This technique discovers the unusual data in large datasets for further investigations on the phenomenon.

- Predication

Predication is used to estimate relationships between dependent and independent variables in order to achieve a prediction of what could happen next.

- Association Rules

In large sets it is used to find meaningful relationships between the data and it's helpful for finding uncategorized data.

- Summarization

Summarization is a technique for auto-generating quick reports about the data and it offers graphical visualization of the data. It is useful for having an overview about the information.

1.2 Educational Data Mining

Educational Data Mining is a branch of Data Mining. Education Data Mining, or simply EDM, applies computational approaches to study questions about education [3].

The multidisciplinary of EDM is composed by the computer science's approach for finding data through databases and the educational needs to improve and find new learning scenarios and opportunities for making the acquisition of knowledge easier. If the Data Mining component is purely computer science, the educational component must take care of the human learning skills. The complexity sentimental part of education doesn't make this topic a perfect science so there are many variables and tricks to observe. The meeting point between these two field is found in the possibility of Data Mining to personalize outputs according to the different inputs generated from the different parts involved.

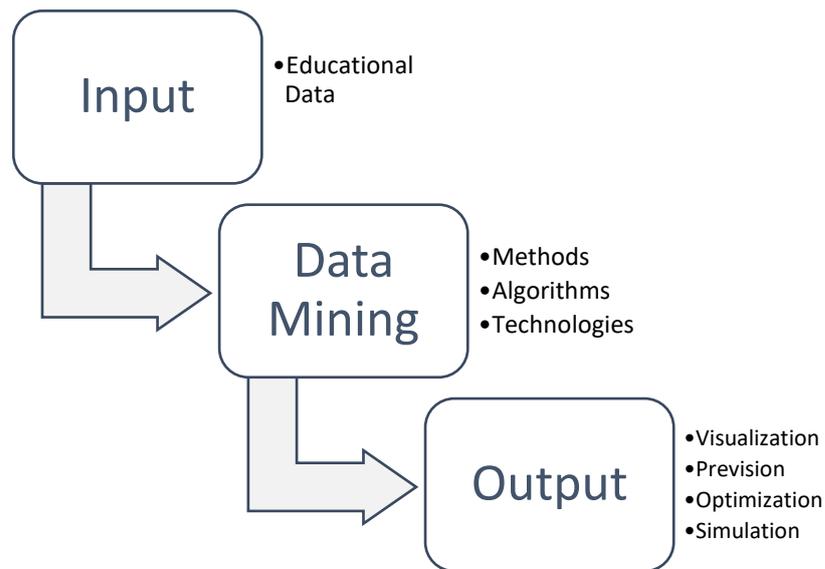
The input data for EDM studies can be from different nature and mixed, depending on which kind of output is looked for.

Nowadays, a lot of institutions are adopting management systems and softwares for easily administrate all the parts involved in the educational environment and, according to a computer science's view, these system's databases are the perfect data input for studies about education. The education environment is a big place in which a lot of different parts and factors take place: students, professors, exams, grades and so on [4]. Tracking all those components and find new relationships between them, creates the

huge chance to study, predict and improve the institution's efficiency in an educational sense. For example, with this kind of data is possible to predict the performance of a student according to his school performance history.

The spread of the web has created new dynamics between the classic teaching-learning model. A lot of information is contained through the internet in different forms and ways and thanks to the new technologies all this data is readable and processable for machines, regardless if it is text, image, video or audio. There are websites with learning contents organized in very different ways and conceived for different degree of learning, the different format of this information allows to find learning materials suited for each student with different needs. The not-located nature of internet creates a new scenario of learning: E-learning. This scenario provides the presence in the web of a lot of learning data and new possibilities for students. A good interpolation of all this data can make a learning tool for students which can improve their self-learning and a starting point for teachers for organizing their knowledge.

In the same way that Data Mining provides methods to access and process raw data in meaningful information, Education Data Mining finds his domain of interest in the educational data for generating outputs of the same type [5]. Visualization and prevision are two of the main outputs usually expected by research in this category: the first one gives to the users an easy understanding of the processed content and the last one allows to avoid or prevent situations not good for the learner/teacher.



[Figure 1: Educational Data Mining's steps for generating information]

Educational Data Mining is an evolving science, there is still a lot of work and studies to do about this topic and experts like C. Romero and S. Ventura in one of their publications explain some limits of the actual state of EDM [4]. For the researchers, a standardization of data and models would allow an easier usage for the developers and, since EDM is mainly supplied by the classical DM algorithms, would be better to include the educational domain knowledge in the algorithms for improving their efficiency.

1.2.1 Educational Data Mining Stakeholders

Educational Data Mining tries to improve “Offline Education”, the classical way to teach through the face-to-face contact, with the new possibilities that technologies give to the people. E-learning, Intelligent Tutoring and Adaptive Educational Hypermedia System are learning methodologies that can be easily improved and integrated by an EDM System for guarantee a better experience to the users.

As already mentioned, there are a lot of different people that interact in a different way and reasons to the teaching/learning process. Learners are the ones whom want to improve their knowledge and EDM can suggest them new contents and resources for

finding new materials according their studies, in order to don't follow only one premeditated path that might not fit with their learning's skills. Educators and Course Developers can receive feedbacks, find new way of teaching and search for common student's mistakes and find the effectiveness of the teaching proposal. Universities and Administrators are able to improve the quality of their studies and make some predictions about the students and improves the service they offer. The multitude of people involved in the school's enviroment makes different needs from each side involved and a lot of data to analyze and process for a meaningful purpose.

1.2.2 Educational Data Mining Usages

The main goal of Education Data Mining is to support the actors of the educational environment through the teaching/learning process, making it easier and more focused on their possible lacks. The identification of this phenomenon includes a lot of different factors and people involved and it implies a deep understanding of the education theory, in order to discover the real needs and which type data is necessary to process for reach the expected goals.

For designing an EDM framework is necessary to take notice of the human part related to the educational process. Humans own different ways to communicate and interact according mainly to their education and cognitive processes, that's why is important to find the right questions and answers. The final users need to be taken in consideration and the purpose of the study need targeted on that specific category of people. EDM, for definition of itself, implies a big set of data to find and process and it's important to identify or suppose the right relationships between them in order to achieve the right output, without any pointless or inaccurate features.

Another usage of Educational Data Mining system is the simple research intention, more in line with the classical and general usage of Data Mining. This approach is more about statistics and the goal is to monitor and anticipate potential problems in the teaching

environment, avoiding for example students at risk of failure or about drop the school prematurely.

According to experts of the field, the usages of Education Data Mining can be divided in several main categories. Some of these categories are similar to the common usages of Data Mining in the different disciplines, but some of them are really pedagogical-oriented. It's important to mention that it's really hard to divide EDM's usages in one strict category because often some of these divisions are somehow related or hybrids. It's important to cite, at least, some of them.

- Analysis and Visualization of data

Analyze and visualize data is, generally, one of the main usages of Data Mining. The purpose of this class is to process all the information contained in the educational environment and create a visualization of the results in order to create reports about the status of the organization. In the same way companies use Data Mining for having a general overview about each department, selling, consumer satisfaction and the all related concepts, EDM can be used for having a general overview about what is happening in the educational institution. All the algorithms used in this group are more likely statistically oriented since statistics is the mathematical science of studying large amount of data. A first usage of this type of output can be a pedagogical study of the data, for example for trying to understand the behavior of the students processing the data related to their career. In this way it's possible to generate reports studying information like grades, timing, classes, failures and demographical aspects. A deeper and more complex usage is to find relationships between more parts of the data and expand the domain of interest, going further on what the students do in the Management System of the Institution and understand where the learners find new resources and how much time they spent for learn something new. This kind of feedback is very helpful for tutors for improving their teaching skills. Regardless the type of data studied only a clear and understandable graphic can really give an important result for the community, mainly because often the final users of these visual data are not-experts of EDM but just normal users.

- Prediction of student performance

One of the oldest usages of EDM is the prediction of student performance. The main goal of this usage is to predict the results of the student using data such as mark, performance, activity on the university website and all the variables related to the university environment that somehow can explain something about the student's life in school. Nowadays with the introduction of On-line Courses and Web-based materials, which allows to keep track on the user's actions, it's easier to find new variables that can indirectly explain the student's learning routine. The prediction of the student performance can occur finding the links between a student's dependent variable and one or more independent variables, otherwise grouping individual elements in groups based on some inherent characteristics and compare them to elements already classified. For examples a good study is to predict the possibly success or insuccess of a student during the course, tracing all his activities and interactions with the course's materials, his previous marks and his attitude in course's projects and homeworks.

- Providing Feedback

If analyzing data and visualize it is more about to extrapolate information straight from data, providing feedback is to find good tools for decision making. Applying the most common Data Mining's models, cited at the beginning of this chapter, the goal of receive feedback is to find facts and problems for be aware of them and eventually find a solution. The given feedback can be adressed to all the parts involved in the EDM context: for teachers it's might be helpful find an automatic way for evaluating the students work and for students would be meaningful to have advices about how to study and where they can receive suggestions for additional materials. In this category Text Mining finds a huge application of himself, because it's important to find not only the data but contextualize it and find meaning in it. For examples if there is a tool of automatic evaluation of the students works, it's important to find the concepts showed by the students in the teacher's materials. Since human language is plenty of synonyms

and the mean of a word can change according to the sentence, it's essential a deep understanding of the sentence and that's where Text Mining takes place.

- Social Network Analysis

Social Network Analysis provides a different approach to the research of meaningful information for educational's purpose. Instead of finding information about the single learner entity, it provides to study the social relationships between the entities. Studying how a person interacts to the surrounding environment and his peers, allows to find interesting facts such as which kind of content the student would like to see more often or which suggestions would improve his interest in a certain topic. In the same way the Social Networks like Facebook or Instagram suggest us which friends, pictures, pages or posts we are more likely interested to see according to our previous actions on the platforms, this usage of EDM can give a suggestions and awareness about the educational life surrounding us. Social network analysis is a community-oriented approach where the information of the single peer not linked to other's peer information is useless.

- Developing concept maps

An important usage of Educational Data Mining is the automatic or semi-automatic creation of concept maps. Since ever, in pedagogy conceptual maps are important tools for learners and teachers for having an overview about the concepts of a certain topic. A concept map is a simple graphic representation of the subject divided in nodes, the key-concepts, and links, the relationships between nodes. For the construction of concept maps, Text Mining algorithms are applied to extrapolate the essentials from educational papers, teaching materials and all the media related to an educational context.

A better explanation regarding conceptual maps and their usages are reported in the following chapter.

- Constructing courseware

Courseware's construction is the automatic creation of the learning contents or recognize it in other sources. The purpose of this task is to help teachers in the development of the course's materials and find new resources for improving the already existent contents.

- Other Educational Data Mining Tasks

Educational Data Mining is a really versatile science that can be helpful in a lot of education contexts and be related to all the educational environment peers. Many usages can be still studied and discovered in this topic and besides the ones already cited previously, there are still few tasks that should be briefly mentioned in this thesis:

- Recommendations for students: a good way for give students recommendations according to their necessities and habits;
- Student Modeling: creations of models of the student's status related to their knowledge and skills;
- Detecting undesirable student behaviors: the goal is to find the learner at risk or who are likely to assume an inappropriate behavior;
- Planning and scheduling: developing of course schedule and all the daily routine in university of the learners/teachers in an automatic way;
- Grouping students: creation of groups of students with similar features in order to create a targeted learning process for them.

1.2.3 Improvements

It's safe to say that a lot of studies are made about EDM and is a developing science that improves according to the new technologies and discovers about pedagogy.

A lot of studies need to be done to improve the actual state of EDM or for finding new important usages of this science. So far, all the Educational Data Mining studies are supported by general Data Mining algorithms and methods so would be a good improvement to optimize these algorithms in the EDM's direction. Another problem related to this science is the lack of educational data, since educational in the digital world is a brand new thing and since not many years now the universities and institutions are starting to use digital systems. Regarding the distribution of data there is another interruption for the spreading of Educational Data Mining: privacy. For many, the usage of people's data for Data Mining is against the privacy of the single person because it monitors all the actions of the users and, directly or indirectly, it might afflict the person and it analyzes personal data for comparing it with other information. Another point that should be taken care, for the experts, is that Educational Data Mining tools should be easy to use and understand because often the final users are people without any computer sciences and statistical knowledge and this science is conceived for a lot of different users with different necessities.

1.3 Text Mining

Text Mining is another important computer science strictly related to Data Mining and Educational Data Mining. If Text Mining can be seen as an extension of DM, for EDM is core piece for finding important data through educational papers and all the contents that can be translated in text format [6].

The main challenge for Text Mining experts is to find information in unstructured data. Unstructured data is all the data not organized in a pre-defined data-model and, nowadays, is the mainly part of the available data in the Web. If structured data implies a model to reference, the not structured data needs to find his own path to find affordable patterns for understanding which data is meaningful and which is not.

Similar to DM and EDM, Text Mining tries to find and extract data for processing it. The main different with the first two sciences is that Text Mining needs to deal with the complexity of human communication. Communication between humans implies the presence of synonyms and words that have different meanings according to the context they are in and the machine needs to properly recognize all this language's shades for a good analysis of the content. Another issue is that mostly every language follows his own grammatical pattern, so it's really hard to find an algorithm that can match with all the languages of the world [7].

Text Mining can be used either for finding words or categories in text files either for finding concepts in them. The first option is simpler and it doesn't require a semantic study of the words, something that the second option needs to find.

The applications of Text Mining are huge and useful for a lot of disciplines. For examples, for understanding in a Social Network what the users are talking about. The goal of the project is to find meaning connections between a given list of words and the proper usage of these words in the video's subtitles, so a deeper and practical presentation of this science will be showed in the following chapters.

1.3.1 Text Mining Usages

Text Mining usages are several and all of them differentiate according to the desired output [8].

- Categorization is one of the oldest usages of Text Mining and it concerns the classification of text data in categories. These categories can be pre-assigned or automatically found according to the text content.
- Information Retrieval aims to find text materials on the Web or in a general database for analyzing them.
- Clustering is an automatic process in which the text files are sorted by keywords or groups with similarities, so called "clusters".

- Summarization extrapolates the core of the text from big files, generating a summarize.
- Sentiment Analysis provides to extract meaningful text in a subjectively way. The goal of this usage is to understand the text's author opinions and feelings regarding a certain topic.
- Natural Language Processing studies the best ways for processing the human language and how successfully the machine can understand and learn the language. This topic will be discusse in the following chapter.

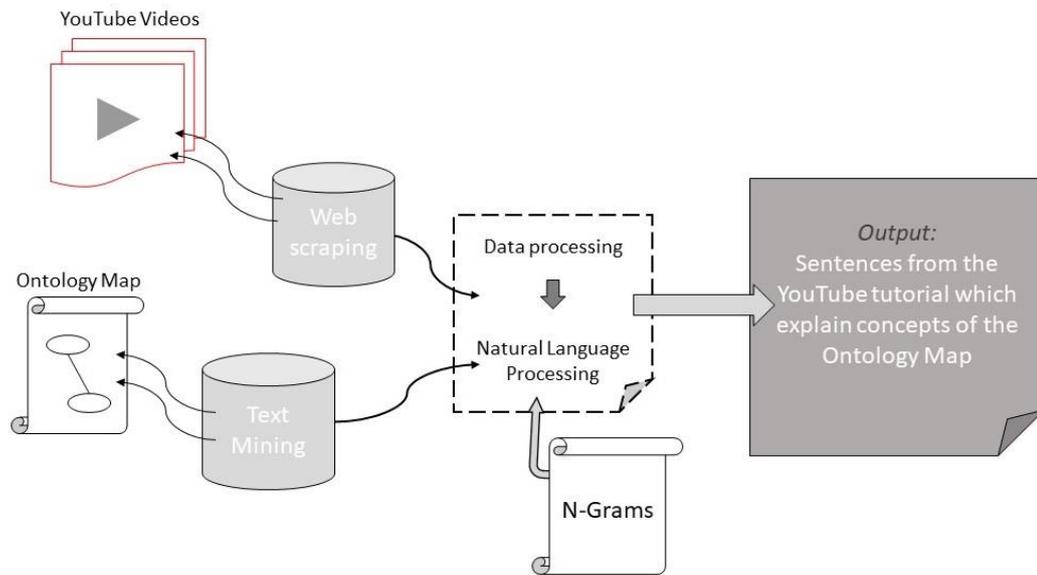
Chapter 2

Designing an Education Data Mining Tool

2.1 Framework

The project is developed according a design model that can be divided in six steps:

- 1- In a preliminary phase, the only data owned is the ontology map about Ruby programming;
- 2- The following step is to find something related to compare the map with, in this case a Youtube tutorial about Ruby Programming;
- 3- Scrape the data from the Web for downloading the subtitles of the videos;
- 4- Compare the keywords of the map with the scripts of the tutorials;
- 5- Process the similitaries between the two inputs for finding matches and meaning through the data;
- 6- Generate an output.



[Figure 2: Framework]

2.2 Technologies

For the development of the project, several technologies are used. The whole project is developed in Ruby code language, which is really useful for scraping the web and easy to learn. The coding part is written in RubyMine by JetStorm. The ontology map is provided in the CMap software and the matching concepts between the map and the subtitles are showed on Cmap as well.

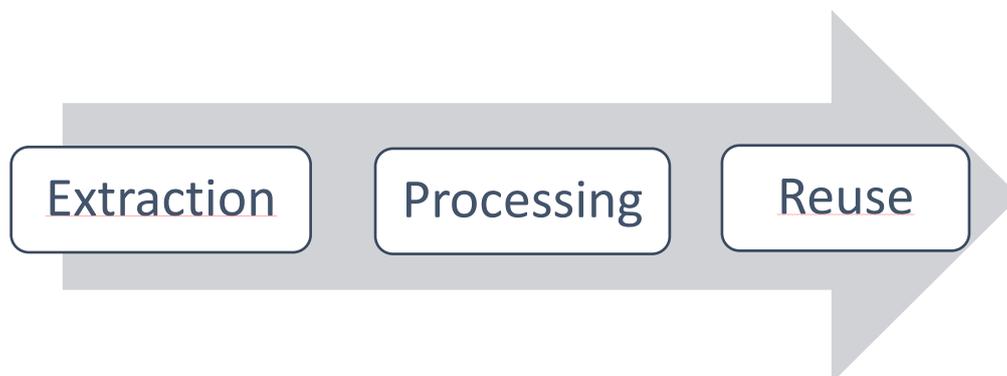
During the production of the tool different approaches were considered. Some failures and inaccuracies found during the development led to other choices for problem solving.

In the following paragraphs are explained, theoretically, all the technologies and methodologies considered during the study.

2.3 Scraping the Web

Web Scraping is a methodology for extracting data from the Web through a software or tool in automatic way. Several code languages can be used for this purpose, Ruby and Java are one of those.

This feature is often used in business tasks and it is strictly related to Data Mining. Web Scraping tools simulate the human web navigation and they access to information through Browser or directly from HTTP protocol. [9] One of the goals of this procedure is to avoid human errors and repetitive tasks for data retrieval. It finds his application in the HTML structure and from that extrapolates data to process in other contexts. One of the applications is to analyze and download data from a web server and re-use it for another purpose that include that previous data.



[Figure 3: Web Scraping phases]

There are several debates about the legality of Web Scraping because this technique doesn't require any permissions from the owner of the Website and, often, are unclear the usages of the processed data. For its definition Web Scraping extracts data from something that potentially can be seen and used by every human because it doesn't mine through databases or sensible data but it only takes what is included in the HTML

and related source code, so basically it can extract data from everything a normal user can experience during his navigation through the Web page. In the general opinion it's not allowed to scrape data for commercial or illegal purposes, but there are not strict rules about this field and it basically depends on the privacy laws of the country.

For example, a web scraping application related to this project is to copy and paste YouTube videos subtitles directly from the web page. This possibility is given to every user on the platform, in fact it's possible to visualize the scripts of the videos in a section of the web page.

An easy way to scrape a website is through API, usually furnished by the databases owners. Google, Facebook and Wikipedia give their API partially for free and, sometimes, the problem with this alternative is that the usage of the data is limited and not many actions are allowed.

2.3.1 Ruby

Ruby is an object-oriented programming language, it was developed in Japan in 1996 and it was written in C language. Lately, Ruby is getting known by the mass because it's easy to learn and it doesn't need particular previous programming knowledge. The syntax is easy to memorize and understand because, as the Ruby's creator said, it is thought for people and not for machines. The structure is similar to Python but it offers a lot of features that makes the programming language really user-friendly. Ruby gives a lot of freedom to the coders and that's why this programming language is easy to learn for everybody.

An important feature of Ruby is the "Gems", these gems are libraries usually developed by the Ruby community and shared for free on the www.rubygems.org website. The objective of the gems is to furnish to the people a better tool for develop in an easiest way Ruby softwares. For integrating a gem, or more, in a script is necessary to find the gem, download it and install it. The installation can be done from the command prompt and it is necessary to include and install in the script the gem needed plus all the gems that the chosen gem use. For Web Scraping, there are two useful gems: Nokogiri which

allows to parse HTML and XML files in Ruby and Mechanize which derives from the Nokogiri and it makes easy automated web interactions.

Ruby is a good programming language for rapid prototyping and for creating scripts and with the Framework extension “Ruby on Rails” it’s easy to develop web applications as well. A first prototype of Twitter was written in Ruby on Rails.

It’s possible to compile a Ruby scripts in a normal Text Editor and run it thanks the command prompt. For the project, RubyMine by JetBrains was used as development environment.

2.4 Conceptual Maps

A conceptual map is a tool for visualizing the relationships between concepts. The concepts are usually expressed as nouns and they are linked each others through relationships defined usually by verbs. The link between two concepts is binary, each key word is connected to the other one thanks to unique relationships.

The goal of concept maps is to express knowledge with the minimum number of words needed and connect them through concise expressions that explain the relationship between the words.

A conceptual map is meant to be readed from the top to the bottom: the top, usually, are the general concepts of a topic and going down to the hierarchy became more specific.

There are no limits regarding the field of usages of Concept Mapping, it’s a tool that can express every type of knowledge and it can be used by users of all ages. The applications of conceptual maps are different and usable in different contexts.

The theory behind concept mapping is that humans learn new concepts and connect them with other concepts already understood. Concept maps are the graphical representation of these cognitive connections. This assumption implies that each

human being can understand differently a concept and represent it in a different way even though the topic is the same and that's why it's important to keep the concept map as easier as possible in order to make easier the understanding of the map to the general user.

An important value of conceptual map is that, according the experts, it can show properly the knowledge of a person regarding a topic and the different perception of the key words and/or relationships can create good discussion about the explanation of the topic itself. For a learner, this tool can be a good way for fix concepts in his mind and stimulate his creativity for finding meaningful relationships between concepts.

2.4.1 Ontology Map

Ontology is a formal and explicit representation of a certain type of knowledge limited in a domain. Ontology needs a definite domain because it aims to model the domain of application for a better understanding and for a better performance, instead of finding a common domain for everything.

Finding similarities in ontology means to find strictly comparable features between concepts, according to the domain of interest [10]. Xiaomeng Su gives a clear and concise definition of what ontology applied on maps is:

“Given two ontologies A and B, mapping one ontology with another means that for each concept (node) in ontology A, we try to find a corresponding concept (node), which has the same or similar semantics, in ontology B and viceverse.” - [Xiaomeng Su] [11]

This sentence explains the direct relationship between concept maps and ontology. Ontology maps, similar to Concept maps, show how concepts are related according to ontology's fundamentals.

In computer sciences, Ontology has a lot of applications regarding Artificial Intelligence, Semantic Web and Text Mining [12].

For the development of this project, the ontology is represented by an ontology map about Ruby Programming developed by the experts of University of Education of

Weingarten in Cmap. As mentioned by Professor W. Mueller and Professor S. Rebholz, in the map there are some relationships not well defined because there might be different ways to define these relationships and they were not sure about which one should fit more.

2.4.2 CMap

CMap is a free software for creating concept maps. The user interface it's easy to understand and it allows to format in different the content of the map. The software provides the creation of complex concept maps and it supports the hypermedia links. Different exporting format are supplied by CMap, for example HTML, JPG, PDF and txt. This feauture makes CMap versatile and usable in different research contexts.

It's very useful in the context of education, a lot of schools already use this software for improve the learning process. The extension CmapServer allows to share the concept maps with everybody and supports the co-creation of the maps.

It is developed by Florida Institue for Human and Machine Cognition, a no-profit organization, and the download is available on <https://cmap.ihmc.us/> for free.

2.5 Natural Language Processing

Natural Language Processing (NLP) analyzes text data through computational methodologies for finding linguistic meaning within it. The analysis can be semantic or syntatic according to the purpose of the study and it finds structures in unstructured information, such as the human language [13].

Several methodologies and algorithms are applied for extracting the desidereted outoput. Often those techniques are supplied by statics and deep learning.

A fist approach used for this analysis is the "tokenization" which is the division of the text is single words in order to analyze them as a single entity. From this first processing many considerations can be done, it's possible to count the frequency of a single word in order to study how much this word is important in the context or analyze the adjacent words of a choosen words for understanding in which sequence this word is often used.

Other study about this field can be done thanks to deep and shallow parsing which elaborate, in different levels, the grammatical relations between the words or thanks to entities which are relevant words taken from a pre-setted dictionary.

For processing the data in an efficient way, “stemming” is used for morphologically normalize the words [13]. Stemming allows to analyze and associate for each word, all the tenses and plural forms related to that word.

Natural Language Processing often is supported by other computer sciences like Neural Networks, Big Data, Machine Learning and N-grams.

2.5.1 N-grams

N-grams analysis is text-analysis technique used for comparing sentences of a text with arrays of strings called N-grams.

A N-gram is an array of N-element containing a sequence of splitted words that usually appear sequentially in the natural language [16]. The cardinality of a N-gram changes according to the necessities of the study, roughly more words are included in the sequence of strings and more accurate the result will be but it means a bigger time of processing and well defined ontology of arrays. According to the cardinility of the N-gram, the name assumes a different form.

N	Name
1	Uni-gram
2	Bi-gram
3	Tri-gram
4	Four-gram

[Figure 5: N-grams classification]

For example, the sentence “Ruby is a code language” if it is splitted as Bi-grams is: Ruby is – is a – a code – code language.

For choosing the right cardinality of the N-grams is necessary to understand how accurate our analysis need to be. Bigger is “N” and stricter is the sequence of words that composes the N-gram, it generates a really specific string and any small variations in the sentence wouldn’t be recognize as the same as the N-grams array content. It means that the information is lost it because wouldn’t match with the N-gram. On the other hand, a small “N” can introduce a generical restriction that can provide a certain amount of error in the analysis.

This division allows to find which words follow and anticipate a certain word and from there is possible to understand the most common sequences regarding the word.

The usages of N-grams are to predict a following word of a choosen word, for example for anticipating what a user is about to type or say, and for understanding in which context is used a word for giving a semantic study of the phenomenon, this application is part of the aim of the study.

The processing of the data usually is a strings matching, since the simple nature of the N-grams.

It’s important to notice that N-grams cannot recognize automatically the semantic meaning of a sentence, they only can recognize what is inside the sequences of strings. For example, if a Tri-gram is the string array “is a color” and this sequence is found in the sentence “a bottle is a color”, a system that uses N-grams wouldn’t recognize that the sentence doesn’t have any logical sense but according to its dictionary the words sequence is right. In the same way if a Four-gram is “the house is new”, the system wouldn’t recognize that sentences like “the house is barely new” or “the house is totally new” are pretty similar to the Four-gram content because the sequence order changes.

Chapter 3

Implementing an Education Data Mining Tool

3.1 First Step: Data Inputs

This chapter will describe the development process of the realized tool and it includes the failures or other tries made during the process. The process is divided in three main steps.

As already cited, the data inputs of the tool are: the ontology map about Ruby Programming and the scripts of a series of YouTube's video tutorials about Ruby.

3.1.1 Ontology Map

As preliminary step, the ontology map given by Mr. Müller and Mrs. Rebholz was translated from german to english in order to compare it with the videos which are in english as well. From Cmap, the map was exported in txt format for making easier the data processing in RubyMine.

Method	has	Visibility	
Program	represents	Algorithm	
public	regulates	Visibility	
public	controls	Permission	
structogram	visualizes	Algorithm	
Flow chart	visualizes	Algorithm	
Code Block	is a	Function	
Procedural	based on	Procedure	
Float	Data type		
for	is a	Loop	
Instance Variable	is a	attribute	
Algorithm	consists of	Statement	
Operand	is used	Expression	
machine language	is representable		Assembler
SubProgram	supports	recursion	

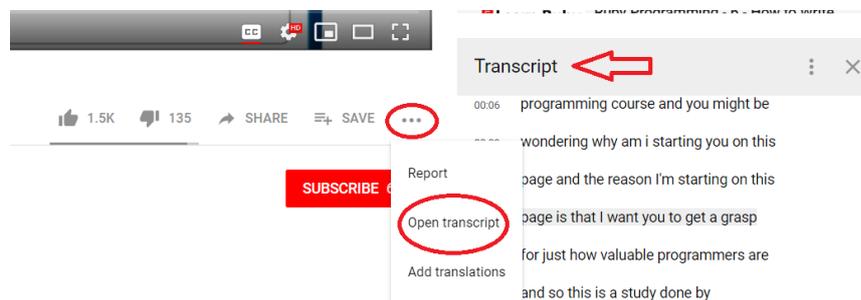
[Figure 6: Ontology Map converted]

3.1.2 Video scripts

The chosen tutorial is a set of 37 videos by Jake Day Williams, reachable at the following link:

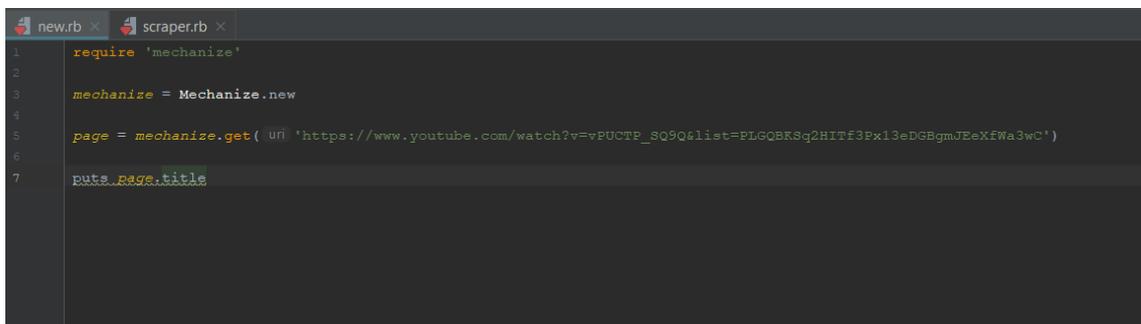
<https://www.youtube.com/watch?v=8I539U5IXWY&list=PLMK2xMz5H5Zv8eC8b4K6tMaE1-Z9FqSOp>

YouTube allows to visualize the script of a video clicking first on the three dots below the right-bottom corner of the video, then click on “Open transcript” and the transcript will be showed on the right.



[Figure 7: Access to Youtube's transcripts]

The first idea was to create a script with Ruby for automatically downloading the script. As already mentioned, thanks to Mechanize is easy to navigate in Website for scraping it. In the same way a user can find the transcript, copy and paste it, Ruby methods and Mechanize can work for the same goal [18]. Unfortunately, on the first try was discovered that YouTube is developed with the AJAX technique and Ruby doesn't support this format. The following picture show a script for scraping the element "title" in the HTML code of the web page and already from this was clear that were some problems.

A screenshot of a code editor showing a Ruby script. The script is named 'scraper.rb' and is located in a file named 'new.rb'. The code consists of seven lines: 1. 'require 'mechanize'', 2. 'mechanize = Mechanize.new', 3. 'page = mechanize.get(URI 'https://www.youtube.com/watch?v=vPUCTP_SQ9Q&list=PLGQBKSq2HITf3Px13eDGBgmJEeXfWa3wC')', 4. 'puts page.title'. The code is written in a dark-themed editor with syntax highlighting. The URI is a YouTube video link. The script is intended to fetch the title of the page using the Mechanize library.

[Figure 8: Web Scraping code example]

At the beginning was unclear why the code was returning empty arrays and no-errors were found in the script. After a research on the web, the problem was found in the usage of AJAX from YouTube and the discover that Ruby doesn't support this format, because Ruby can only scrape from the HTML source code and not from server, as AJAX technique provides. This part of the study is developed during the month of November 2019, so the considerations are related to this time period.

As second approach YouTube APIs were considered but the limits of usages and the lack of time and economic funds the hypothesis was dropped. There are high chances that this approach is the good one for download subtitles from YouTube with Ruby.

As final solution, the transcripts were copied and pasted manually in a txt format, in order to be compared easily with the ontology map's words. The format of the text files is divided by lines: video's title, video's URL at the beginning of each video and then time

code and related sentence for each line of the scripts. The video's title and URL are introduced by the word "INFOVIDEO".

```
|INFOVIDEO Install Ruby and Editor  
INFOVIDEO https://www.youtube.com/watch?v=8I539U5lXwY&list=PLMK2xMz5H5Zv8eC8b4K6tMaE1-Z9FgSOp  
00:00  
hey guys it's Jake and today we are  
00:03  
going to be starting our learning Ruby  
00:06  
programming course and you might be  
00:09  
wondering why am i starting you on this  
00:11  
page and the reason I'm starting on this  
00:14  
page is that I want you to get a grasp  
00:17  
for just how valuable programmers are
```

[Figure 9: Transcripts sample]

3.2 Second Step: Matches

The second step of the development of the tool is to find some matches between the ontology and the subtitles. This step is divided in two smaller steps: the first step is a quick comparison between the inputs for understanding if the video tutorials chosen are comparable to the map and the second step is the research of the sentences which include the matching words.

Since the ontology used is restricted regarding a specific field, it's necessary to find a series of videos that cover partially or totally the concepts presented in the map. Different Youtube's playlists were initially searched for the study and for choosing the most suitable between them, it was developed a script for finding the words of the ontology in the transcripts. The set of videos chosen is the one with the most significant matches.

Subsequently the first processing, in the matches there are a lot of "Stop-words". Generally, stop-words are those words that are not useful on a research because of their common use or their meaning not appropriate for the study field. In this particular case, the set of words not useful are the verbs that explains the relationships between the

concept in the ontology map, adverbs, prepositions and conjunctions. This list of words is not meaningful for the study because the objective of the tool is to find matches in the concepts where our information is contained, independently of how they are related. Thanks to the Ruby method “.delete_if”, those words are found and deleted after a first processing of the words.

```
map_words.delete_if {|i| i == "has" || i == "is" || i == "of" || i ==
"a" || i == "as" || i == "in" || i == "based" || i == "an" || i ==
"on" || i == "place" || i == "define" || i == "from" || i ==
"interpreted" || i == "allows" || i == "takes" || i == "written" || i
== "worth" || i == "uses" || i == "creates"}
```

[Figure 10: Deleting stop words]

At the end, the chosen tutorial includes 34 word matches between words without stop words, for this reason part of the ontology map is covered and there are enough samples to study.

Method	Hash
Program	return
Function	value
type	Machine
for	code
Loop	times
Variable	actual
used	global
Expression	until
machine	each
language	local
Class	Array
Ruby	Operator
conditional	if
description	elsif
case	else
while	end

[Figure 11: Matching concepts]

It's important to notice that some words such as “for” and “if”, usually considered as stop-words, are necessary to keep for this study because these words admit a

programming meaning as well. A better consideration of this group of words will be done later on the research.

At this point the data is filtered according to the research necessities and the next step is to find those words in the scripts of the videos.

For the comparison of the two text files is necessary to split each word in the ontology map and save each of them in a stand-alone array and to split the scripts lines in arrays which contain only one line each.

Furthermore, it is indispensable to delete the first two lines at the beginning of the script of each video. The title and url of the video are introduced by the string "INFOVIDEO" and for this part of the study is not necessary this information and it is not in the interest of the tool to find matches between the concepts and the Youtube video's title because the goal is to find where the word is mentioned in the video. Similar to the stop-words, the line containing the video's info are not considered for the matches analysis.

The matches are provided by a loop which compare the word in the ontology map with the sentences of the scripts. It's notable to say that the intention is to find the concepts in the scripts and not the opposite. In a practical way the code is asking to find the concept word in the scripts sentence, scanning through all the words that compose the sentence.

```
map_words.each do |element|
  tutorial_sentences.each do |index|
    if index.include? element and unless index.include? "INFOVIDEO"
      matches.push(index)
    end
  end
end
end
```

[Figure 12: Code example for filtering the sentences]

The output generates a collection of 2049 sentences taken from the videos. The huge number of sentences suggests that it's necessary a better processing of the data for finding meaningful matches. From a deep look of the outputs is clear that a lot of words like, "if", "else" and "type", are partially used in a different context of the ontology domain so better text mining techniques need to be used.

3.3 Third Step: Filtering the matches

The first elaboration of the data didn't provide good results for the study, it is necessary to go deeper with the mining of text. The second step is necessary to have a first look on the validity of the input and if is worthy to work with a tutorial instead of another one, but now is crucial to find the meaning of these words in the context. The huge amounts of synonyms and similarities in a language make harder the total comprehension of the meaning of a text for a machine. In this case, in particular, there a lot of similarities because the programming terms are taken from already-existing words that have their own meaning in the normal communication language vocabulary, so it's important to fully understand where those words are used in the programming related meaning and where those words are used out of this context.

Natural Language Processing is the approach adopted for the resolution of this problem, particularly N-grams methodology is used for reaching the goal.

3.3.1 N-grams cardinality

The cardinality of the N-grams can be from 1 to N and it is necessary to understand which cardinality is better for the study. Generally, a bigger cardinality can give a better result but it implies a bigger dataset of N-grams entries and it's hard to find a proper dataset for a particular category, such as programming, with a lot of different sequences of words that can give an accurate study of the phenomenon. Two is the

cardinality chosen for the study, because in this context a sequence of two words can already represent a good comparing tool.

3.3.2 Creation of bi-grams from the transcripts

For comparing the chosen bi-grams with the sentences of the video tutorial is indispensable to generate string arrays of two words elements from the transcript file. The sentences of the transcripts are splitted in a sequence of two words each, each sequence contains the last element of the previous sequence and the following word according to the sentence order.

For example, the sentence “each value method” generates two string arrays: [“each”, “value”] and [“value”, “method”].

3.3.3 Filtering with the COCA’s dataset

As a first approach, since the topic is really restricted, the idea is to find a big dataset of bi-grams to use as a “black-list”. Usually, N-grams are used as the set of strings that are needed to find in the ontology but since programming is a specific field it is worth to find the sequences of words that are not related to programming for keeping only the sentences that are related to this field.

frequency	word1	word2	word3
1419	much	the	same
461	much	more	likely
432	much	better	than
266	much	more	difficult
235	much	of	the
226	much	more	than

[Figure 13: COCA list example]

The sets of bigrams used are taken from the Corpus of Contemporary American English (COCA) website [14]. This repository has the biggest database of N-grams and it's possible to choose different N-grams categories, based on the cardinality or particular features such as case or not case sensitive. This dataset of N-grams includes the most common words of the English language, so the aim of the "black-list" is to delete, at least, the words like "if" that often are used in the normal construction of a sentence.

Comparing the bi-grams provided by the COCA and the bi-grams arrays from the transcript, no matches are found. This result shows that the COCA's dataset doesn't include bi-grams that can be helpful for the study.

No progress is made for the study, so it's necessary to find another way for filtering the data.

3.3.4 Creation of N-grams dataset

So far, the methodologies adopted for filtering the data didn't provide the expected results. On the Web there are not free N-grams dataset that can help to reach the aim of the study so it's necessary to build a dataset containing the sequences of words that can improve the system [15].

If a first approach was to create a "black-list" of words to delete from the transcripts, now is worthy to find a "white-list" of words that can match with the strings of the video sentences. The creation of N-grams is given by big repositories where some words are found in order to statistically see which words are often used in a context and which words follow and precede usually the needed word.

As already mentioned for processing the data, bigrams are used. A list of two words sequence is made with the most common usages of the words included in the ontology map. Thanks to some research on the web and the help of experts, the list is made but after a first running of the tool, still there are some matches that are not properly fitted in the study.

From the bigrams of the matching words generated from the transcripts file, the frequency of these bigrams is calculated in order to see which bigrams are common in the domain of the study. Those bigrams that are often recurring in the transcripts can improve the dataset created for the evaluation of the text. After an evaluation of the new generated bigrams and the ones already made, a reliable bigrams dataset is made.

Processing the bigrams made and the strings of the transcripts, the number of the sentences decrease from 2409 lines to 153 lines. Those new matches are all related to the ontology and it's safe to say that the remaining lines are related to the programming field and the matching words from the ontology and from the video's transcripts are used in the same semantic way.

Script Name	Usage	Output
1_evaluation_word_matches.rb	Finds the first matches between the ontology map and the transcripts (only words)	1_evaluation_tutorial_scripts.txt
2_matches_finder.rb	Finds the full sentences which include the matching words in the transcripts	2_matches.txt
3_matching_word_frequency.rb	Calculates the frequency of the matching words in the transcripts	3_matching_words_frequency.txt
4_matches_bigrams_generator.rb	Generates bigrams of the matching sentences taken from the transcripts	4_matches_bigrams_generator.txt
5_tutorial_bigrams_frequency.rb	Calculates the frequency of the bigrams in the transcripts	5_tutorial_bigrams_frequency.txt
6_matching_sentences_frequency.rb	Shows the bigrams taken from the transcripts which include the concepts of the ontology map	6_matching_sentences_frequency.txt
7_final_matches.rb	Finds the final matches between the transcripts and the ontology map, according to the bigrams	7_final_matches.txt

[Figure 14: Script's list]

File name	Usage
tutorial.txt	Includes all the transcripts of the videos, divided by title and video url
map.txt	Includes all the words of the ontology map
sorted_grams.txt	Includes the bigrams alphabetically sorted
bigrams_DB.txt	Includes all the bigrams

[Figure 15: Files list]

Chapter 4

Study's Results

4.1 Evaluation of the contents

The first results of the study are the matching words between the ontology map's concepts and the words in the YouTube's videos transcripts. The tutorial with more matches in this phase is the one used during the study.

The concepts that are apparently discussed in the videos are enough to the study goal: those matches represent almost the 40% of the total concepts showed in the ontology map.

The outputs cover the basics and most common words in the programming field and this result is acceptable since the chosen tutorial is an introduction to the Ruby programming.

The following figure shows, in light orange, the matching words which are those concepts where the research goes deeper.

4.2 First Matches

After the matching words are found, it is needed to find the sentences in which these words are included. A first elaboration of the data didn't generate an acceptable result. Many sentences are out of the programming context because of the usage of the matching words in a colloquial context or these sentences are not properly satisfying to be included in the study research.

For example the sentence "because we've named our character Ray's", taken from the list of matching lines, uses the ontology word "character" not for explain the programming variable named "character" but for referring to a fictional person called Ray.

This analysis delivers the obligation to improve the study and move to the Natural Language Processing techniques.

4.3 Bigrams Dataset

Following the basic of the Natural Language Processing and of N-grams, a dataset of bigrams is generated. The development of this set of bigrams is provided by research and scraping on the web, experts opinions and analyzing part of the transcripts of the videos.

The following list shows an extract on the Bigrams dataset developed.

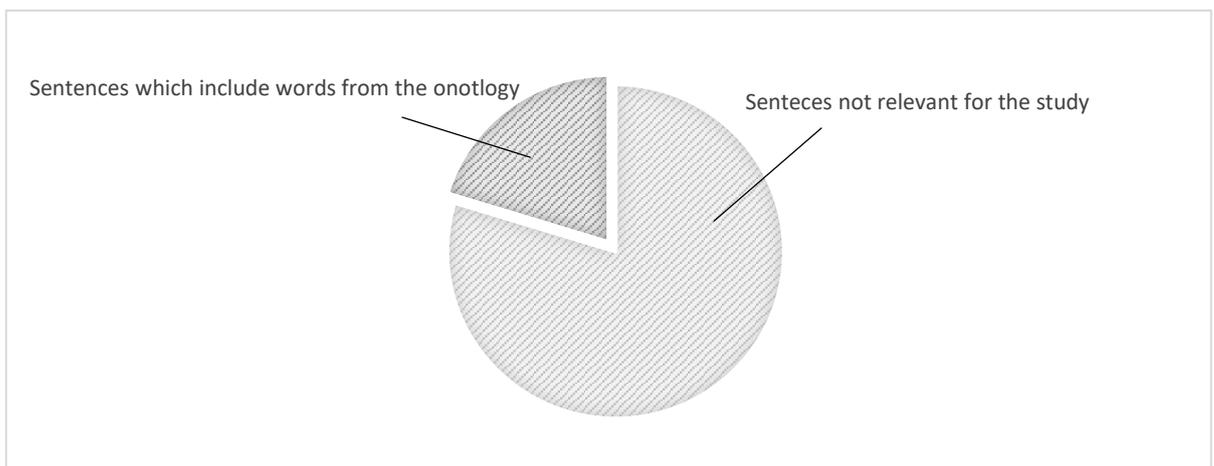
```
value character
value float
value hash
value integer
value method
value statement
value string
value type
```

[Figure 17: Developed Bigrams dataset sample]

4.4 Final Results

The study shows that N-grams, in particular Bi-grams, can help to evaluate the content of a text taken from videos. The generated output matches with the concepts included in the ontology map developed by Mr. Wolfgang Müller and Mrs. Sandra Rebholz and it's safe to say that most of the matches found are related to the programming field.

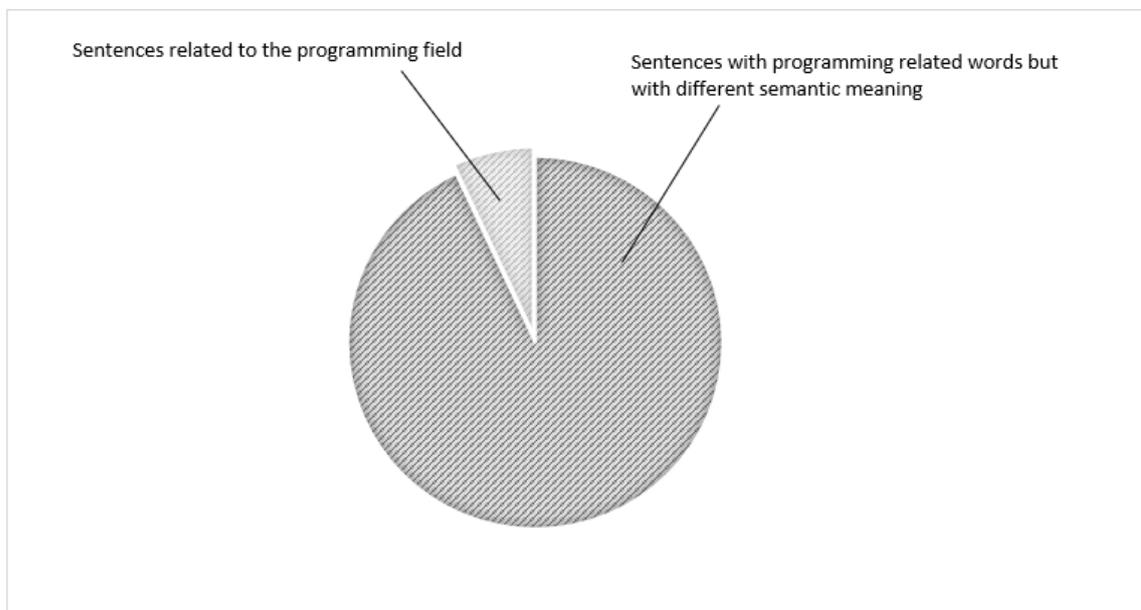
A first processing of the data generates 2049 lines, all of them include at least a concept showed in the ontology map. In this first elaboration of the inputs, there is a lot of inaccuracy through the sentences because several terms are used in a context out of the programming field. From 8046 lines taken the transcripts, the domain of the study is reduced to less the 25% of the total lines and it can be considered a good domain for the study considering that a spoken speech includes a lot of senteces used for driving the flow of speech in an understandable way. It's valuable to say that the transcripts used for the research are made by amateurs so the language used is simplified and often not properly technical, since the target of the videos are people with no specific skills and this adds a percentage of sentences that are not related properly to the field of the study.



[Figure 18: First filtering percentage]

The introduction of N-grams in the study provides better results. The second processing of the data moves the study from a simple comparison of words to a semantic

comparison of them. Thanks to the developed Bi-grams set the found matches are compared not only with the word itself but with the previous one and the following one, so for the machine is possible to understand better in which context the word is used. This implementation of the the tool decreases the domain to the study to 153 lines out of the 2049 lines from the first processing. The final matching sentences furnish a set of lines that are all related to the programming field, the goal of the study. From the first elaboration of the study only the 8% of the results are taken and considered valid.



[Figure 19: Second filtering percentage]

In the total of 8046 lines that compose the transcripts of the videos, 153 sentences are meaningful for the study and it consists of the 2% of the total. Even though the numbers might not seem appealing, it's an important result for the study itself: Data Mining, for its nature, evaluates huge amount of data for using alone a small part of that which is meaningful for the purpose.

The final matches cover properly a part of the ontology map and still the results appear interesting because the ontology map represents a very general and huge topic of concepts and study and the tutorials can cover only a small amount of the concepts. The research is made considering a playlist of video for evaluating the performance of the

tool and the validity of the research question but these results show that is possible to aspire to cover all the concepts of the ontology map including more information taken from the web or other digital resources.

4.5 Evaluation

Following the generation of the outputs, two evaluations are made for proving the value of the research and of the tool. The first evaluation is made by some experts which judged the sentences and gave some thoughts and considerations about the precision of the semantic analysis made, the second evaluation is provided by “TextRazor” which is a specialized software in Natural Language Processing and it provides an impartial analysis of the context, according to the algorithms and knowledge of the platform.

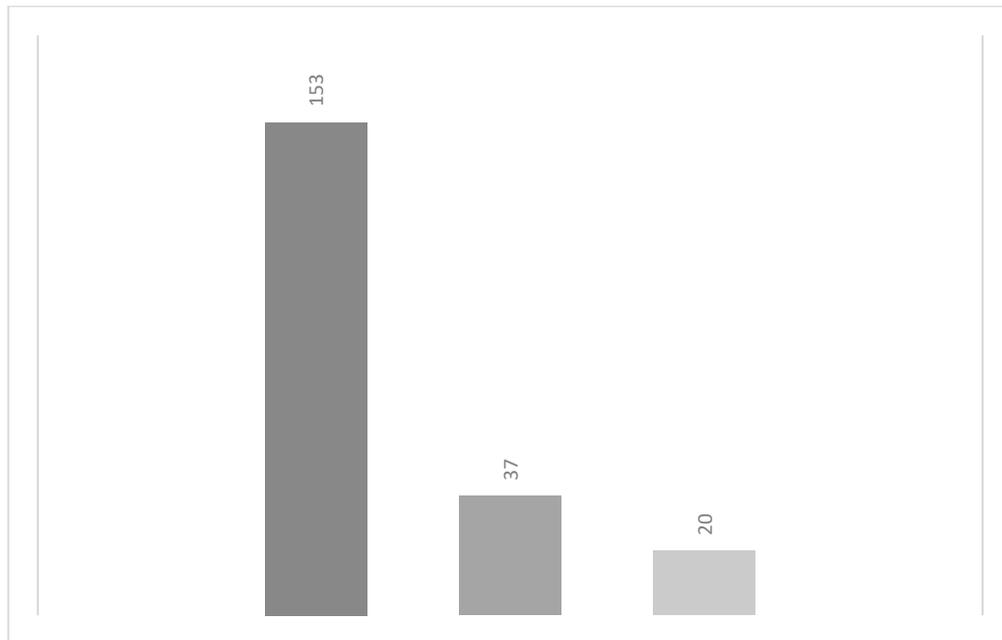
4.5.1 Experts Evaluation

Subsequently the processing of the data, the final results are examined by experts for evaluating the value of the output generated.

The matching lines are divided in three main groups: lines which are explaining or introducing a programming concepts, lines which are explaining programming concepts but used more as an example or in generical way and lines which are not really useful to be added in the ontology map. All those sentences are totally related to the programming field, so the N-grams showed a good performance following this path.

The analyzed sentences are 153, from the total 37 senteces are classified as not satisfying. From those 37 senteces, 20 linses use a recurring sequence of words which are “to program” and “to program in Ruby”. This fact is due by the ripetion of the sentence “Today we are at the lesson number N of our tutorial about how to program in Ruby” and similar, which introduces each lesson. The remaining lines are still carrying some error, related to the usages of Bi-grams and a small dataset of samples for comparing the lines. Totally, from 153 the 24% is considered as not meaningful and from that 24% almost the 55% is the same sentence repetition. The following chart shows in

dark grey the total of sentences generated from the tool (153), the middle grey shows the total sentences considered as inaccurate (37) and the light grey shows how many sentences from those 37 are following the same error pattern (20).



[Figure 20: Inaccuracy analysis of the final results]

4.5.2 Evaluation through Entities

Another evaluation of the generated outputs, for proving the accuracy of the Bi-grams analysis made on the data, is given by entities thanks to the “TextRazor” software.

In Natural Language Processing entities are elements of an unstructured text grouped by categories [19]. The goal of this technique is to provide a reference for each entity and categorize words and sentences according to this reference.

TextRazor is an API developed for text analysis, in particular for Named Entities Recognition [20], it can be integrated to softwares written in several code languages and it supplies a demo of the API, through a Website, for analyzing the text. The demo is reachable at: <https://www.textrazor.com/demo>.

The 153 sentences processed by the tool are elaborated in the TextRazor demo for having a report about the text meaning found by the application.

The report shows the list of each sentence with the found entities for each words of the line. An example of a line is given in the following picture which proves that related entity is found properly, according the programming field.

the user is going to return .

Words Phrases Relat

string

Normalized Entity Id: String (computer science)

Normalized English Entity Id: String (computer science)

Wikipedia Link:
[http://en.wikipedia.org/wiki/String_\(computer_science\)](http://en.wikipedia.org/wiki/String_(computer_science))

Freebase Id:
</m/06x16>

Wikidata Id:
[Q184754](https://www.wikidata.org/wiki/Q184754)

Confidence Score: 1.354

Relevance Score: 0.5191

[Figure 21: TextRazor analysis example]

Evaluating all the sentences processed is possible to confirm that all the entities found are recognized by the tool as words taken from the programming field, as the ontology map requests. This statement proves that analysis made through the Bi-grams is reliable, since all the sentences and words left are totally related to this field.

TextRazor provide a report about the percentage of the categories which compone the text, the following picture shows all the categories found for the sentences and all of them are categories regarding the programming field.



[Figure 22: Categories found by TextRazor]

Another report demostrates all the topics recognized in the text analysis and once again all of those topics are related to the ontology map field.

TOPICS
1.00
Control flow
1.00
Relational operator
1.00
Array data type
1.00
Ruby (programming language)
1.00
Conditional (computer programming)
1.00
Mathematical logic
1.00
Computing
1.00
Programming constructs
1.00
Areas of computer science
1.00
Software engineering
1.00
Computer programming
1.00
Software development
1.00
Computers
1.00
Theoretical computer science
1.00
Notation

[Figure 23: Topics found by TextRazor]

This final evaluation provided by one of the most reliable tools of Natural Language Processing proves the validity of the text analysis made by the Educational Data Mining tool developed during the study.

4.6 Acknowledgements

The study is made at the University of Education of Weingarten and it is supported by Professor Wolfgang Müller and PhD student Sandra Rebholz which provide the basics for the development of the research. They provide the conceptualization of the study's research, thanks to their knowledge regarding pedagogy and computer sciences. In particular, they furnished the ontology map about programming which represent the main ontology of the study and the element of comparison for the information find on the Web and the validity of the results. A huge thanks to these experts for supporting me during the study and for sharing with me their knowledge.

Another special thanks to the experts that helped me during the study and for evaluating the final outputs.

Chapter 5

Conclusions

Nowadays technologies can improve and facilitate the learning and teaching process, more studies about Educational Data Mining can help to find new educational scenarios. Information spreads on the Web faster than ever, more data is added every second and it's important to keep the meaningful data for using it for useful purposes. The study and the implementation of the tool show that several methods can be used for processing data and use it for educational purposes.

The first part of the study explains how to access to some information through the Web, in particular thanks to Web Scraping. Ruby is good code language which allows easily this goal and that can be improved using APIs provided by many. YouTube, for this research, is one of the best environments for looking for educational data in a free way. More websites, such as Udemy.com, can provide a similar service with even better video tutorials regarding different fields and with the same approach adopted for this study is possible to take information from there and process it.

Following the choose of the inputs, the processing part can be provided by many code languages and supported by associated tools. The developed tool adopts the application of Natural Processing Language, in particular N-grams, for the resolution of the problem but many more technologies can be used such as Machine Learning, Artificial Intelligence and Neural Networks. N-grams is useful in this context because it doesn't require particular resources nor a deep knowledge of databases. Furthermore, more information is added to the tool and more the N-grams dataset can be improved. The final results show the validity of the tools and the technologies used and involved for the development of the study. The framework adopted can work either with text ontologies

and every media format that can allow to extrapolate text after a pre-processing, such as audio format.

The availability of the information and the division of the tool in few simple steps make the tool easy to understand and flexible. The tool can be used:

- By teachers for improving their course materials, referring it to external resources and offer to the students a bigger choice of learning contents;
- By learners for having a place which is possible to find easily all the information related to the concepts that are needed to learn and personalize and organize the main content of the class according to the resources they prefer and find more useful for learning;
- By experts for evaluating how much of information is contained in a text comparing their ontology.

The final results of the study can be considered satisfying because the objectives of research are reached. Starting from the scripts a good filtering of the sentences is made and final matches are mostly related to programming field, avoiding the inaccuracy of the first processing of the data. Furthermore, there aren't on Web set of bigrams about programming and the research helped to find some of them.

The two evaluations made prove the validity of the study and of the tool developed. The final lines are totally related to the programming field, even though with a small amount of error in the accuracy of the semantic meaning.

5.1 Improvements and future works

The tool generated good results and it can be considered a starting point for developing a more complete system.

A script for downloading automatically the subtitles can be really helpful for automating the process. This would make the usage faster and more efficient. YouTube's APIs can be used for improving this part because Ruby, so far, doesn't allow that.

It's necessary to improve the dictionary of the bigrams for allowing a better understanding of the future information that will be added. It's worthy to find N-grams regarding the key-concepts from the ontology that are not found in this first prototype. New N-grams can be found through the Web in files related to programming or manually added if there are not enough contents to analyze. The experts analysis made with the final results prove that a bigger and accurate dataset of N-grams can improve the study and the application of Tri-grams can do the same but it's necessary to notice if there isn't any lost of information already found. Furthermore, other Natural Language Processing methodologies can be used, such as Neural Networks.

Improve the study adding new contents to analyze for making the concepts of the map all-covered, in order to have information about each single concept showed in the ontology map. Another improvement in this direction is to add synonyms in order to find the concepts that are showed in the video but the word used is not the same used in the video but with the same meaning.

A visualization of the concept which allows to move through the map can be developed for making easier for the user to navigate through the contents. For example, the Javascript's library "3D.js" can reach this goal and can be one of the approaches for this resolution. The file containing the lines of the transcripts is furnished of the title, url and time code of each line so with the methods provide by Mechanize in Ruby is possible to create loops which can automatically generate the links in which the sentece is said in the video and with a responsive visualization is possible to create an interface where

you can move through the video clicking on the concepts of the ontology map and navigate the videos according to the chosen concept.

A final improvement that would make the study a complete tool that can be easily used by many, is a user interface for adding the inputs and elaborate all the information for have a complete final output that can support the learning and teaching process.

All these points can be used for going deeper on the understanding of the all the topics discuss in the study and maybe open new research opportunities about Educational Data Mining.

Bibliography

- [1] https://en.wikipedia.org/wiki/Data_mining
- [2] Data Mining techniques and applications, Bharati M. Ramageri / Indian Journal of Computer Science and Engineering, Vol. 1 No. 4 301-305
- [3] https://en.wikipedia.org/wiki/Educational_data_mining
- [4] Educational Data Mining: A Review of the State-of-the-Art, Cristobal Romero, Member, IEEE, Sebastian Ventura, Senior Member, IEEE
- [5] Handbook of Educational Data Mining, Cristobal Romero (Editor), Sebastian Ventura (Editor), Mykola Pechenizkiy (Editor), Ryan S.J.d. Baker (Editor)
- [6] Text Mining: The state of the art and the challenges, *Ah-Hwee Tan*
- [7] <https://lorenzogovoni.com/text-mining/>
- [8] Text Mining: Techniques, Applications and Issues, Ramzan Talib, Muhammad Kashif Hanif, Shaeela Ayesha, Fakeeha Fatima
- [9] WEB SCRAPING, APPLICATIONS AND TOOLS, Osmar Castrillo-Fernández, ePSIplatform Topic Report No. 2015 / 10, December 2015
- [10] <http://cmap.ihmc.us/docs/conceptmap.php>
- [11] A Text Categorization Perspective for Ontology Mapping, Xiaomeng Su
- [12] Ontology Mapping – An Integrated Approach, Marc Ehrig and York Sure, April 28, 2004
- [13] Natural Language Processing, Karin Verspoor, Kevin Bretonnel Cohen, University of Colorado Denver
- [14] <https://www.ngrams.info/>
- [15] <https://www.sitepoint.com/natural-language-processing-ruby-n-grams/>
- [16] N-Gram-Based Text Categorization, William B. Cavnar and John M. Trenkle, Environmental Research Institute of Michigan

- [17] *Ontology Learning for the Semantic Web*, Alexander Maedche and Steffen Staab, University of Karlsruhe
- [18] <https://www.rubydoc.info/gems/mechanize/Mechanize>
- [19] <https://www.kdnuggets.com/2018/08/named-entity-recognition-practitioners-guide-nlp-4.html>
- [20] <https://www.textrazor.com/technology>

List of pictures

1. Educational Data Mining's step for generating information	5
2. Educational Data Mining tool Framework	15
3. Web Scraping phases	16
4. Ontology Map	21
5. N-grams classification	23
6. Ontology Map converted	26
7. Access to Youtube's transcripts	26
8. Web Scraping code example	27
9. Transcripts sample	28
10. Deleting stop words	29
11. Matching concepts	29
12. Code example for filtering the sentences	30
13. COCA list example	32
14. Script's list	35
15. Files list	36
16. Matching concepts on the ontology map	38
17. Developed Bigrams dataset sample	39
18. First filtering percentage	40
19. Second filtering percentage	41
20. Inaccuracy analysis of the final results	43
21. TextRazor analysis example	44
22. Categories found by TextRazor	45
23. Topics found by TextRazor	46