Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)
(рівень вищої освіти)

«Методи та моделі Text Mining в задачах аналізу проектної документації»
.
.
(тема)

Виконав:
студент 2 курсу, групи СШІм-18-1
Гулько Д.І.
(прізвище, ініціали)
Спеціальність 122 – Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-професійна
(освітньо-професійна або освітньо -наукова)

Освітня програма Системи штучного інтелекту (СШІ)
(повна назва освітньої програми)
Керівник проф. Рябова Н.В.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ В.О. Філатов
(підпис) (прізвище, ініціали)

2019 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук

Кафедра Штучного інтелекту

Рівень вищої освіти другий (магістерський)

Спеціальність 122 – Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-професійна
(освітньо-професійна або освітньо -наукова)

Освітня програма Системи штучного інтелекту (СШІ)
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 20 ___ р.

ЗАВДАННЯ
НА АТЕСТАЦІЙНУ РОБОТУ

студентові Гулько Дарії Ігорівни
(прізвище, ім'я, по батькові)

1. Тема роботи «Методи та моделі Text Mining в задачах аналізу проектної документації»
затверджена наказом по університету від 4 листопада 2019 р. № 1623Ст
2. Термін подання студентом роботи до екзаменаційної комісії 11 грудня 2019 р.
3. Вихідні дані до роботи Науково-технічні публікації, дані Інтернет джерел та наукових проектів щодо розробки та дослідження методів Text Mining
4. Перелік питань, що потрібно опрацювати в роботі Subject field analysis, Analysis of the project documentation, Types of the project documentation, Classification methods, Naïve Bayes Algorithm, Rationale for the choice of the algorithm, Main technical and business domain knowlage

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів) _Рисунок 1 – Концептуальна карта Text Mining; Рисунок 2 – Типи IT документації; Рисунок 3 – Составі частини проектного плану; Рисунок 4 – Результат роботи алгоритму._____
_____
_____
_____
_____

6. Консультанти розділів роботи (п.6 включається до завдання за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Основний розділ	проф. Рябова Н.В.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Аналіз літератури та Інтернет джерел	05.11.2019 – 08.11.2019	
2	Аналіз предметної області	08.11.2019 – 12.11.2019	
3	Постановка задачі, узгодження з керівником	12.11.2019 – 15.11.2019	
4	Аналіз видів проектної документації	15.11.2019 – 20.11.2019	
5	Вибір інструментів	20.11.2019 – 21.11.2019	
6	Розробка алгоритму класифікації	21.11.2019 – 26.11.2019	
7	Написання пояснювальної записки	26.11.2019 – 08.12.2019	
8	Демонстрація роботи програми	09.12.2019 – 09.12.2019	
9	Рецензування, проходження нормоконтролю	10.12.2019 – 12.12.2019	
10	Захист роботи	19.12.2019	

Дата видачі завдання __04 листопада_____2019 р.

Студент _____
(підпис)

Керівник роботи _____ _____
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Записка пояснювальна: 106 с., 20 рис., 2 табл., 6 дод., 20 джерел.

КЛАСИФІКАЦІЯ, ПРОЕКТНИЙ ПЛАН, ПРОЕКТНИЙ МЕНЕДЖМЕНТ, DOMAIN KNOWLAGE, NAÏVE BAYES , TEXT MINING

Об'єкт дослідження – процесси обробки проектної документації за допомогою методів штучного інтелекту.

Предмет дослідження – алгоритми Data Mining для обробки та роботи із текстовими документами.

Мета роботи – застосування методів класифікації з метою автоматтизації процессу визначення доменної області проекта.

Методи дослідження – використання методу Naïve Bayes для обчислення вірогідностей приналежності об'єкта до тієї чи іншої предметної області.

РЕФЕРАТ

Пояснительная записка: 106 с., 20 рис., 2 табл., 6 прил., 20 источников.

КЛАСИФИКАЦИЯ, ПРОЕКТНАЯ ДОКУМЕНТАЦИЯ, ПРОЕКТНЫЙ МЕНЕДЖЕМЕНТ, ПРОЕКТНЫЙ ПЛАН, DOMAIN KNOWLAGE, NAÏVE BAYES , TEXT MINING

Объект исследования – процессы обработки проектной документации с помощью методов искусственного интеллекта.

Цель работы – применение методов классификации с целью автоматтизации процесса определения доменной области проекта.

Методы исследования – использование метода Naïve Bayes для вычисления вероятностей принадлежности объекта к той или иной предметной области.

ABSTRACT

Explanatory note: 106 p., 20 fig., 2 tabl., 6 ann., 20 sources.

CLASSIFICATION, DOMAIN KNOWLAGE, NAÏVE BAYES, PROJECT DOCUMENTATION, PROJECT PLAN, TEXT MINING

The object of study is the processing of project documentation using artificial intelligence methods.

The purpose of the work is the application of classification methods to automate the process of determining the domain area of a project.

Research methods – using the Naïve Bayes method to calculate the probabilities of an object belonging to a particular subject area.

CONTENT

GLOSSARY, ABBREVIATIONS, ACRONYMS

API – Application programming interface;

ARM – Association rule mining;

IT – Information technology;

CRM – Customer relationship management;

DM – Data Mining;

ERM – Environmental Resources Management;

IoT – Internet of things;

JVM – Java Virtual Machine;

LSA – Latent semantic analysis;

LSI – Latent semantic indexing;

MRP – Manufacturing resource planning;

NMF – Non-negative matrix factorization;

SVM – Support Vector Machine;

TDM – Terminology Document Matrix;

TDM – Text Data Mining;

UML – Unified Modeling Language;

WBS – Work breakdown structure.

INTRODUCTION

The amount of text created every day is increasing dramatically. This vast amount of mostly unstructured text can not simply be processed and understood by computers. Therefore, effective methods and algorithms are required to identify useful samples. Text development is the task of extracting meaningful information from the text that has received considerable attention in recent years. In this paper, we outline some of the most important tasks and methods for text development, including pre-processing, classification, and clustering. Besides, I will briefly explain the extraction of texts in the biomedical and medical fields.

Text data is a good example of unstructured information, which is one of the simplest forms of data that can be created in most scenarios. Unstructured text is easily processed and perceived by humans, but it is much harder for machines to understand. It must be said that this volume of text is an invaluable source of information and knowledge. As a result, there is a desperate need to develop methods and algorithms to effectively handle this avalanche of text in a wide variety of applications. The analysis of structured information stored in databases requires preliminary processing: database design, the input of the information by dedicated rules, its placement in special structures (eg, relational tables), etc. Thus, directly to analyze this information and retrieve from her new knowledge needs more effort. However, they are not always related to the analysis and do not necessarily lead to the desired result. As a result, the quality of structured information analysis is reduced. Also, not all types of data can be structured without losing useful information.

Today the Network is the main source of text (documents), the amount of textual information available to us. About 80% of the total organization information is stored in an unstructured form (reports, emails, opinions, news, etc.) This shows that approximately 90% of the world’s data is stored in unstructured formats. The need to automatically extract useful information from a large body of textual data to aid human analysis is quite obvious [1].

Large amounts of data have been collected routinely in the cource of day-to-day management in business, administration, banking, the dilivery of social and health services, environmental protection, security and politics. However, more and more information, such as textual information, is becoming unstructured and is simply trying to figure it out. Manually analyzing such textual information is becoming increasingly impractical, and as a result, technologies for extracting textual materials are being developed.

1 SUBJECT FIELD ANALYSIS

1.1 Relevance of the work

Textual information is one of the essential components of almost every sphere of human activity. Text can be stored as anything, text messages are used for correspondence, text documents are an integral part of business, medical records are consecutive entries, any article in the media necessarily contains at least a little text. All this vast array of heterogeneous information undoubtedly contains some hidden or somehow useful knowledge, which, due to the volume of data sources, cannot be manually found, no matter how diligent and hardworking an analyst or team of analysts.

Text Mining Systems Research Topic is considered relevant, so information will be doubled every two years. One of the main factors behind this increase is the increase in the proportion of automatically generated data. Most of the textual information is generated, and as a rule, it is large enough unstructured texts. Therefore, some people, especially those working in the IT field, have trouble processing them. To solve this problem, special programs were created for Text Mining. Text mining or text mining is the process of automatically analyzing ordinary unstructured text documents by a computer to obtain high quality structured information.

1.2 Data Mining

Over the years data science as the discipline of analysing and extracting data display us more and more future events. Term of data science was invented in the early 60s by Peter Naur and was used as a usual synonym to computer science. In spite of this, the matter of it started to make sense only with the beginning of technological progress. This is not surprising because data science was developing in parallel with such domains as medicine, marketing, banking,

and other fields where now it could gain some new insights. The cutting edge meaning of «data science» was first outlined during the subsequent Japanese-French statistics symposium organized in 1992. The participants recognized the rise of another term with a particular spotlight on information from different starting points, measurements, types and structures.

As a rule, people always confuse the concept of data mining and statistics because of its related idea–extraction of the new data. The statistic is a mathematically-based area that is designed to collect and interpret information. Whereas data science is a holistic approach in the sense that it supports the entire process including data sensing and collection, data storing, data processing and feature extraction, data mining and knowledge discovery. But traditional mathematical statistics, which for a long time claimed to be the main tool for data analysis, also often fail when solving problems from real complex life. It operates with averaged characteristics of the sample, which are often fictitious values (such as the average temperature of patients in a hospital, the average height of a house on a street consisting of palaces and shacks, etc.). Therefore, methods of mathematical statistics are useful mainly for testing pre-formulated hypotheses (verification-driven data mining). So statistics is a related concept to data-intensive activities – collecting, processing and interpretation of processed data.

Techniques of data mining are based on inductive learning which means the approach offers one solution to different problems. Such an approach relate data mining can be considered as a model which constructed explicitly or implicitly by generalizing from a sufficient number of training examples. The main idea of this method is that the trained model could be useful to future hidden examples [12].

Today, the purpose of Data Mining is to identify hidden rules and patterns in data sets. The fact is that the human’s mind itself is not adapted to the perception of large arrays of heterogeneous information. In addition, a person is not able to capture more than two or three relationships even in small samples. Therefore, most of the organisations – irrespective of their domain – are looking

to monetize on their Big Data and use complex analytical strategies. Most of the organisations are looking to capitalize on their Big Data and are hence using sophisticated analytical methods. As the consumption of Big Data grew, so did the need for data mining.

The field of data mining has been growing due to its tremendous success in terms of broad-ranging application accomplishments and scientific progress, understanding. Several data mining applications have been successfully in force in diverse areas like health care, finance, retail, telecommunication, fraud detection and risk analysis etc. The ever increasing complexities in many fields and enhancements in technology have posed new challenges embrace totally different information formats, information from disparate locations, advances in computation and networking resources, analysis and scientific fields (Figure 1.1).

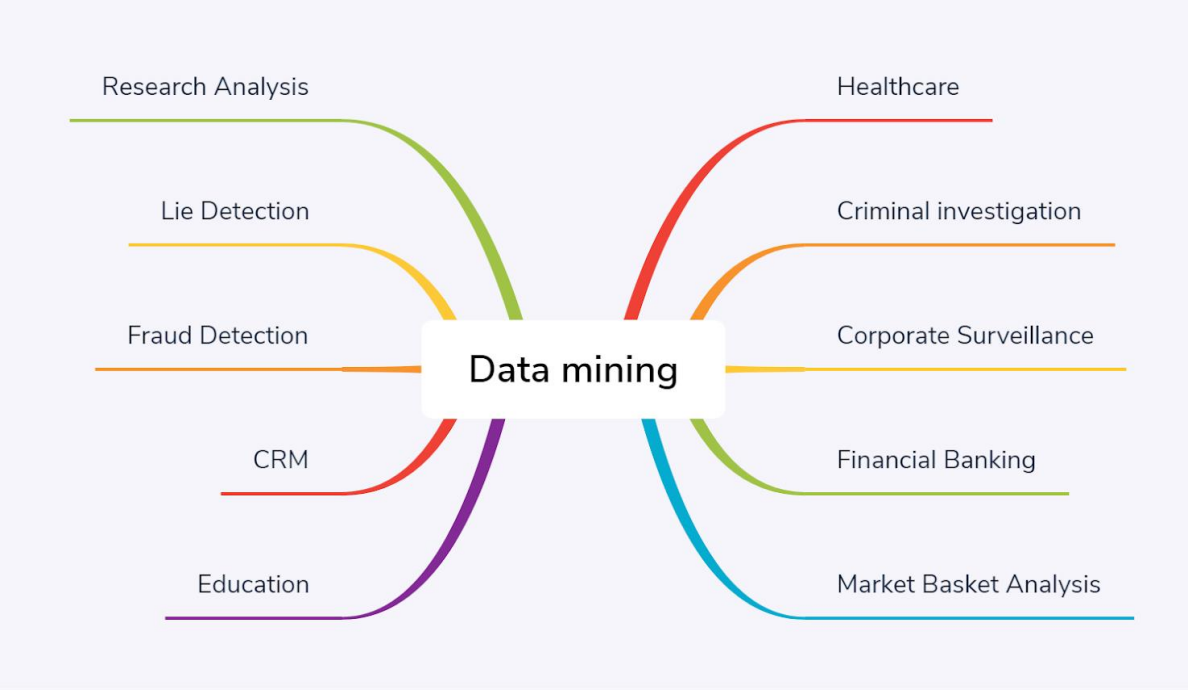

Figure 1.1 – Application of Data mining

Modern technologies of Data Mining (discovery-driven data mining) have already shovel information in order to automatically search for patterns (patterns) characteristic of any fragments of heterogeneous multidimensional data. In

contrast to online analytical processing (OLAP) in Data Mining, the burden of formulating hypotheses and identifying unusual (unexpected) patterns is transferred from a person to a computer. With the advent of the Information Era, accumulating and storing data has become easier and cheaper. Every day millions of bytes of data are created and 90 per cent of all the data in the world today were produced within the past years. It has been estimated that the amount of stored information doubles approximately every year. Unfortunately, as the number of machine-readable information increases, the ability to understand and make use of it does not keep pace with its growth.

1.3 Text Mining

Progressive field of data mining becomes more and more easier to use because of the immense advances in hardware and software technology which has to lead to the availability of different kinds of data. And this is really true case for one of the varieties of data mining–text analytics and text data mining. That is because the development of web platforms and social networks has enabled the rapid creation of large repositories of different kinds of data. In particular, the web is a technological enabler which encourages the creation of a large amount of text content by different users in a form which is easy to store and process. Finding information for just about any need has never been more automatic–just a keystroke or mouse click away. So I believe that it is a big change in the availability of online information.

As in other methods of data mining notion of Text Mining was invented in the mid-1980s, but technological advances have enabled the field to advance swiftly during the past decade. Information extraction is an important text mining problem and has been extensively studied in areas such as natural language processing, information retrieval and Web mining. People usually arrogate the field of text mining to text data mining which is roughly equivalent to text analytics, applies to the process of obtaining quality information from the text.

High-quality information is typically derived through the divining of patterns and trends through means such as statistical pattern learning.

Text mining typically involves the method of structuring the input text (usually parsing, at the side of the addition of some derived linguistic options and therefore the removal of others, and future insertion into a database), etymologizing patterns inside the structured knowledge, and at last analysis and interpretation of the output. Actually, we could say that text mining is a very complex task as it includes dealing with different fuzzy and unstructured data [13]. A key component is that the linking along of the extracted info along to make new facts or new hypotheses to be explored any by additional standard suggests that of experimentation.

A general framework of the text mining could be presented with two components: Text refining that transforms text documents in any form into an intermediate form and knowledge distillation that assumes patterns or knowledge data from the intermediate form (Figure 1.2).

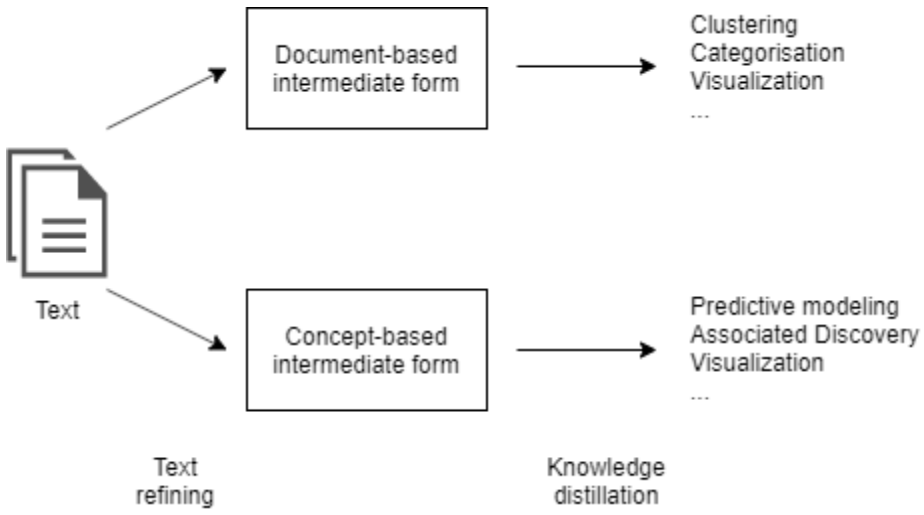

Figure 1.2 – A text mining framework

The intermediate form can be semi-structured like the conceptual graph representation or structured like the relational data representation. It also can be document-based where each entity represents an object or concept of interests in

a specific area. Mining a document-based intermediate form gets patterns and relationship across documents. Clustering, categorization and visualization are examples of mining result from a document-based intermediate form.

Another form of mining is a concept-based intermediate form which determines pattern and relationship across objects or concepts. Operations, such as predictive modelling or associative discovery, fall into this category. A document-based intermediate form also can be transformed into a concept-based intermediate form by reconstructing or extracting the relevant information according to the objects of purposes in a specific domain.

For data discovery in a very specific domain, the document-based intermediate form of the news is projected onto a concept-based intermediate form looking on the task demand. As an example, one will extract info associated with «device» from the document-based intermediate form and form a device database. Knowledge distillation can then be performed on the device database (device-based intermediate form) to derive device-related knowledge. It follows that first one (document-based intermediate form) is usually domain-independent, but concept-based intermediate form not.

The figure 1.3 shows three aspects of the Text Mining: Techniques, Spheres and Problems. The first part of the concept map contains elements such as Text Processing, Dimensionality Reduction, Text summarization, Classification, Text Clustering and Association rule mining. Small numbers near each of the techniques means amount of child elements. As I mentioned before there are various applications and they are also presented in the figure below. Moreover, the figure allows you to get acquainted with individual problems such as Information retrieval, Natural language processing(NLP), Named entity recognition, Disambiguation, Recognition of the Pattern Identified Entities, Document clustering, Coreference, Relationship and event extraction, Sentiment analysis and Quantitative text analysis. All these aspects presented in the figure make it possible to see the integrity of the picture and to organize knowledge about Text Data mining. A full conceptual map is presented in the Attachment A.

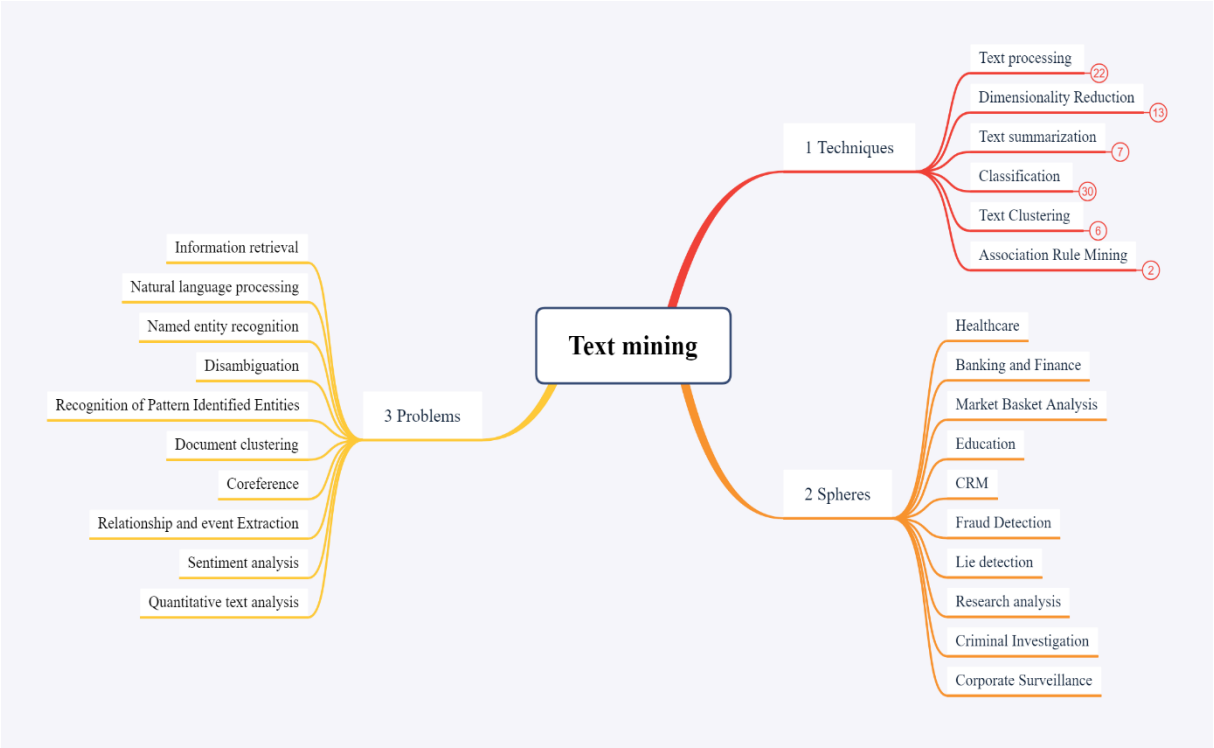

Figure 1.3 – Concept map of text mining

With the fast growth of textual data on the Web, it is expected that future work on information extraction will need to deal with even more diverse and noisy text. Weakly supervised and unsupervised methods will play a more substantial role in information extraction. The various user-generated content on the Web such as Wikipedia articles will also become important resources to provide some kind of supervision. So that regarded by many as the next wave of knowledge discovery, text mining has very high commercial values.

1.4 Tasks of the text mining

As a matter of fact, text mining algorithms are specific data processing algorithms within the domain of natural language text. There is no point to search out an efficient and effective technique for text categorization because varied techniques of text categorization are recently developed. It can be any kind of content -simple business documents, articles, news stream, blogs, emails or even

comments below posts in social media and alternative varieties of unstructured data.

But there are some criteria that affect the quality of mining in methods such as classification and clustering. It is highly dependent on the noisiness of the features that are used for the clustering process. For example, commonly used articles or conjunctions, may not be very useful in improving the clustering quality. Therefore, it's essential to pick the options effectively, in order that the strident words within the corpus are removed before the cluster. In addition to feature selection, a number of feature transformation methods such as Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Analysis (PLSA), and Non-negative Matrix Factorization (NMF) are available to improve the quality of the document representation and make it more amenable to clustering.

Algorithms for text mining include various methods, such as text classification, categorization, and clustering. All of them are aimed at identifying hidden relationships, trends, and patterns which are a solid base for making business decisions.

In this part, the focus is text mining process, a different method of text categorization, cluster analysis for text documents, the basic differences between relative terminologies on the basis of process, models, tools and the algorithms used, a comparison between text mining techniques on the basis of algorithms. Thus, the tasks of the Text Mining algorithms can be divided into the following categories:

- text processing: assigning documents to predefined categories (for example, induction of decision trees);
- text clustering: a descriptive activity that groups similar documents together (such as self-organizing maps);
- information extraction: modeling and discovering concepts, sometimes combining categorization and clustering of approaches to concepts / logical ideas to find concepts and their relationships with text collections (for example, a formal approach to concept analysis to build a concept hierarchy);

- information retrieval: receive documents related to a user's request;
- withdrawal of information: answer questions.

1.4.1 Classification

The classification downside is one among the foremost elementary issues in the machine learning and data mining. Therefore, text mining techniques ought to be designed, to effectively manage massive numbers of parts with varied frequencies. Almost all the known techniques for classification such as decision trees, rules, Bayes methods, nearest neighbor classifiers, SVM classifiers, and neural networks have been extended to the case of text data. Recently, considerable attention has been paid to linear classifiers, such as neural networks and SVM classifiers, the latter being significantly appropriate for characterizing text information.

During the ensuing years, in view of the advancement of web and social network technologies have lead to a tremendous interest in the classification of text documents containing links or other meta-information. The process is presented in the form of assigning a given text into groups of entities in which items are in some way similar to each other. The Classification problem can be stated as a training data set consisting of records. Each record is identified by an unique record id, and consist of fields corresponding to the attributes. An attribute with a continuous domain is called a continuous attribute. An attribute with a finite domain of discrete values is called a categorical attribute. The goal of text classification is to assign a category to a new document (Figure 1.4). By reducing the load on memory and facilitating the efficient storage and retrieval of information, classification serves as the fundamental cognitive mechanism that simplifies the individual’s experience of the environment

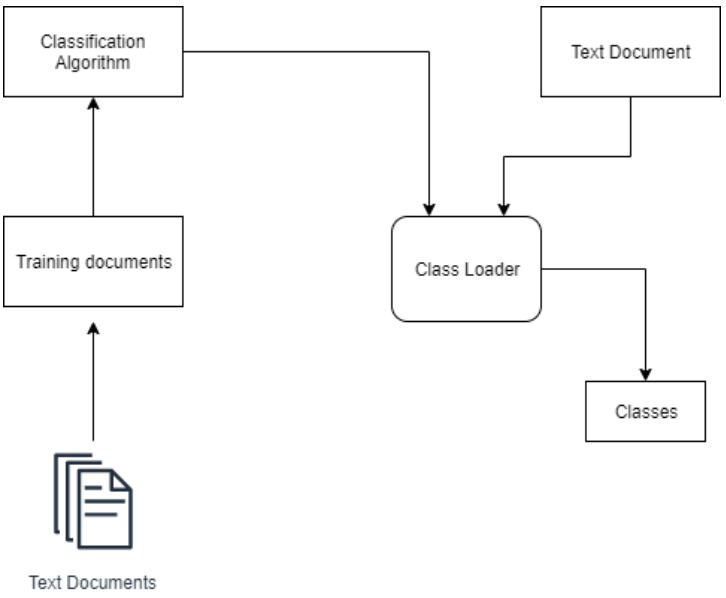

Figure 1.4 – Text mining classification principle

Text classification concerns not only the area of text mining, but also in areas such as data mining, database, machine learning and information retrieval. It used in large range of applications in numerous domains like image process, diagnosing, document organization, etc. Text classification aims to assign predefined categories to text documents [13]. So the aim of text classification is to assign predefined classes to text documents.

1.4.2 Clustering

The problem of clustering has been studied widely in the database and statistics literature in the context of a wide variety of data mining tasks. The clustering problem is defined to be that of finding groups of similar objects in the data. The similarity between the objects is measured with the use of a similarity function.

Clustering is the process to find groups of documents with similar content. It is a process of forming groups (clusters) of similar objects from a given set of inputs. Good clustershave the characteristic that objects belonging to the same cluster are «similar» to each other, while objects from different clusters are

«dissimilar». The idea of clustering originates from statistics where it was applied to numerical data. However, computer science and data mining in particular, extended the notion to other types of data such as text or multimedia. Clustering is an unsupervised process through which objects are classified into groups called clusters. In the case of clustering, the problem is to group the given unlabeled collection into meaningful clusters without any prior information. Any labels associated with objects are obtained solely from the data. An advantage of clustering is that documents can emerge in multiple subtopics, thus ensuring that a useful document will not be absent from search results.

Even though clustering technique used to group similar documents it differs from classification because in clustering documents are clustered on the fly instead of use of predefined set of documents. It is based on unsupervised learning. In data mining, K-means clustering is frequently used clustering algorithm, in text mining field it also gives good results.

It can be very useful in the text domain, where the objects to be clusters can be of different granularities such as documents, paragraphs, sentences or terms. Clustering is especially useful for organizing documents to improve retrieval and support browsing. The study of the clustering problem precedes its applicability to the text domain. Traditional methods for clustering have generally focussed on the case of quantitative data, in which the attributes of the data are numeric. The problem has also been studied for the case of categorical data, in which the attributes may take on nominal values [14]. A good clustering of text requires effective feature selection and a proper choice of the algorithm for the task at hand. Among the different classes of algorithms, the distance-based methods are among the most popular in a wide variety of applications.

1.4.3 Association rules

Associative rule search is one of the most popular Data Mining applications. The essence of the task is to identify common sets of objects, a large

collection of such sets. This task is a special case of the classification problem. Association Rule Mining (ARM) is a method used to identify the relationships between a large set of variables in a dataset.

Initially, the problem was solved in the analysis of trends in the behaviour of customers in supermarkets. Analyzed were the data on the purchases they made, which the buyers put in a cart (basket). This was the reason for the second common name – Basket Analysis. When analyzing these data, the first and foremost interest is the information about what products are purchased together, in what sequence, what categories of consumers are which products are preferred, at what periods. The task of finding associative rules is relevant not only in the field of trade. For example, in the service sector, it is of interest to know what services customers prefer to use in the aggregate. To obtain this information, the problem is solved concerning the data on the services used by one client during the time limit (months, years). This helps to determine, for example, how best to put together the service packages offered to the client. In medicine, the symptoms and illnesses observed in patients may be exposed. In this case, the knowledge of which combinations of illnesses and symptoms are the most common help in the future to make the right diagnosis [2].

ARM refers to finding relationships between a large set of variables, that is, providing a database of records, each containing two or more variables and their respective values, ARM defines common combinations of variable values. It is similar to the idea of correlation A study in which relationships between two variables are uncovered, ARM is also used to identify relationship variables, but each relationship (also known as an association rule) may contain two or more variables.

The main difference between the sequential analysis task and the search for associative rules is to establish the relationship of order between the studied sets. This relationship can be defined in various ways [15]. When analyzing the sequence of events that take place over time, the objects of such sets are events, and the order relation corresponds to the chronology of their occurrence.

1.5 Types of Text Mining algorithms

One of the main reasons for applying data retrieval methods to text document collections is to structure them. And the structure can greatly simplify access to a collection of documents for the user. Well-known structure access is library catalogues or book indexes. However, the problem of manual indexes is the time it takes to maintain them. Therefore, they are very often not up-to-date and therefore cannot be used for recent publications or frequently changing sources of information such as the World Wide Web. Existing methods of structuring a collection either attempt to assign keywords to documents based on a given set of keywords (classification or categorization methods) or to automatically structure document collections into groups of similar documents (clustering). Thus, the problem of text mining Classification of the dataset and detection associations among the data. To overcome the problems of Data Mining, consider the following algorithms [3].

On the figure 1.5 we can see classification of the Text Mining techniques which divided into five parts: Text classification, Text categorization, Association rule mining, Text summarization, Text Processing, and Text Clustering. The biggest block is Text Classification and it is not a surprize. That is because Classification problem is one of the most popular tasks in Text Mining. There some algorithms which could be part of different categories at the same time, so trees like this one are always special. More details about some algorithms are provided below.

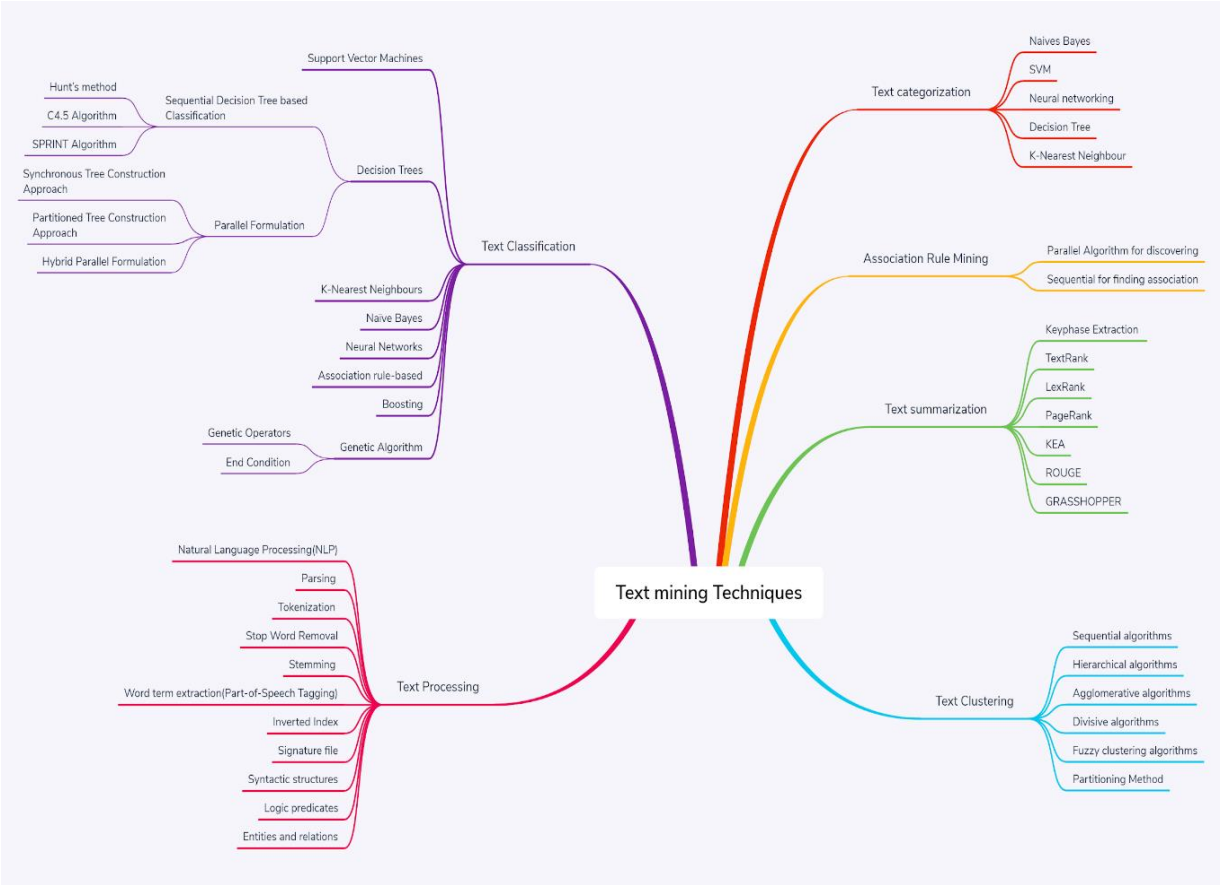

Figure 1.5 – Concept map of Text Mining Techniques

1.5.1 Naïve Bayes

The Naive Bayes classifier is probably the simplest and most widely used classifier. It models the distribution of documents in each class using a probabilistic model, assuming that the distribution of different terms is independent of each other. Although this assumption of «naive Bayes» is false in many real programs, naive Bayes works surprisingly well. Two basic models are commonly used to classify naive Bayes [3]. Both models aim to find the posterior probability used by the classification method, which uses the word probability to classify textual data. In this classification scheme, based on the previous text and templates, the data is evaluated and the class capability is measured.

When constructing the classifier, the assumption is that the words of an individual document depend on some probabilistic mechanism of distribution, and

the category of the document, in turn, is associated with the words found in the document. The following is a formula describing this dependency that returns the likelihood that a document that includes certain words belongs to a category. The document assigns the category for which the formula has the highest probability.

1.5.2 K-nearest

Instead of constructing a formula for the category of a document from its content, this method takes as its basis the fact that similar documents belong to the same categories. There are many different indicators of similarity between texts, most often based on entering into the same documents of the same words [5].

Document similarity is calculated for the entire case, then k most similar documents are selected. All unclassified documents in this group are assigned the most common category in the group.

In the area of text mining, the nearest neighbour algorithm k is a classic and commonly used technique. To find the query text, k the nearest neighbour classifier exceeds the results. This method estimates the distance between two lines for comparison and classifies the text based on the distance.

$$d_a(x, y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2}. \quad (1.1)$$

Where x and y are data instances and d is the distance between x and y. The main advantage of this algorithm is the high accurate classification. On the other hand, the main drawback is the consumption of resources such as memory and time [16].

1.5.3 Decision Trees

The decision tree is a special type of classifier based on the consistent application of predefined classification rules to a set of documents. The classification rules are as follows: in the training set of documents M, a word is selected that can best define the category of documents. The whole set is divided into two new ones – M + and M -, respectively, containing and not containing the selected word documents. The procedure is recursively repeated for new groups until the documents in all received groups belong to the same category.

1.5.4 K-means clustering

This technique is also a classic approach to categorizing text. This uses the distance function as k nearest neighbour classifier for data clustering. This is an effective method of extracting text to save resources, but the precision of this cluster approach is sensitive due to the initial process of selecting the cluster centre[6]. Besides, hierarchical text categorization schemes are available that are not very effective for clustering or categorizing.

Randomly select k points in the space of document vectors. These points represent the centres of mass of the clusters. All documents are broken down into k clusters according to which centre of mass they are closer to, after which the centrists are transferred. Documents are redistributed over distances to new centres of mass. This process takes place until the centres of mass of each cluster become stable, ie they will no longer change after recalculation.

1.5.5 Support vector machine (SVM)

This approach is one of the most efficient and accurate classification algorithms. This approach uses a concept that uses hyperplane-based and sizing techniques to identify or classify data. The main advantage of this algorithm is the

achievement of high accurate classification results.

The basic principle of SVM is to define separators in the search space that can best divide different classes. The advantage of the SVM method is that since it tries to determine the optimal direction of discrimination in the feature space by examining the appropriate combination of features, it is sufficiently robust to high dimensionality. The SVM classifier has also proven useful in large scale scenarios where a large amount of labelled data and a small amount of labelled data are available [6]. It is supervised surveillance through its use of unmarked data in the classification process. This technique is also quite scalable due to the use of some modified quasi-Newtonian methods, which are generally effective in practice.

1.5.6 Neural networks

The basic unit in a neural network is a neuron or unit. Each block receives a set of inputs denoted by the vector X_i, which in this case corresponds to the frequency terms in the i-th document. Each neuron is also associated with a set of weights A, which are used to calculate the function of its inputs. A common feature often used on a neural network is a linear function:

$$p_i = A \cdot \overline{X_i}, \tag{1.2}$$

Thus, for a vector consisting of a vocabulary d words, the vector weight A must also contain d elements. The effect of such multiple layers is to induce multiple linear boundaries that can be used to approximate nested areas belonging to a particular class. In such a network, the outputs of neurons on earlier layers are fed to neurons in later layers. The learning process of such networks is more complex because errors need to be spread at different levels. However, a general observation of the text is that linear classifiers usually give comparable results to nonlinear data, and improvements in nonlinear classification methods are relatively small. This speaks to the added complexity

1.6 Formulation of the problem

As a result of the analysis of existing text processing systems, the preconditions for simplifying and increasing the speed of extracting useful information with the help of Text Mining were created.

Due to the rapid development of technologies and the expansion of managerial responsibilities in the field of IT, it was decided to stop applying classification methods. Given the shortcomings and advantages of existing systems, it was decided to simulate an algorithm for determining the affiliation of project plans to particular domains.

The main purpose is to research and benchmark the basic methods, models and algorithms of Text To achieve the selected goals, the following tasks must be solved:

- conduct a comparative analysis of Text Mining methods and models;
- identify tasks that are solved with Text Mining;
- to carry out the analysis of tools;
- to implement software implementation;
- mining that is used in intelligent systems.

The algorithm is developed should be geared towards IT companies that have a large amount of project documentation and use it in online repositories. In the framework of the work, it is necessary to demonstrate the operation of the developed algorithm within the built-in system Atlassian Jira / Confluence.

2 ANALYSIS OF THE PROJECT DOCUMENTATION

2.1 Project Documentation

Documentation is the workspace of the project. At any time, any employee can quickly find the information he needs and understands how to solve specific problems. And also documents will save you from unnecessary conflicts because they clearly spell out who does what, who is responsible for what, how the system works and what to do if something went wrong.

Documentation is necessary if the company's goal is:

- to ensure the coordinated and competent work of all employees;
- to gain confidence that in the event of a failure, the system and all connected services will be quickly restored;
- to keep records of IT assets, plan and optimize the costs of their maintenance;
- to update the regulatory framework of the company, replace obsolete regulations and instructions in time.

Describing and documenting IT projects helps better understand how it works. Seeing the functions of the individual components and their relationship with each other can help the team make really effective management decisions. There are many documents which depend on other aspects such as methodology or architecture. Documents are different, and some of them are of paramount importance, and some of which are impossible to do without. Large products often develop a whole range of documentation. Further on the picture, you can see the main types into which all project documentation can be divided.

On picture 2.1 we can see tree-diagram which show four main types of documentation in IT sphere: Pre-sale Documentation, Project Management Documentation, Program Documentation and Maintenance documentation. Pre-sale documentation includes development concepts, terms of reference, various contracts. Project management documentation is a description of the list of tasks,

a project plan, a risk plan and other documents related to the management. Program documentation, unlike other types, can be very different, but often it is a description of the configuration and specification. Maintenance documentation consists of such documents as user manual and other instructions. In any case, each company has the right to determine its own list of mandatory project documentation.

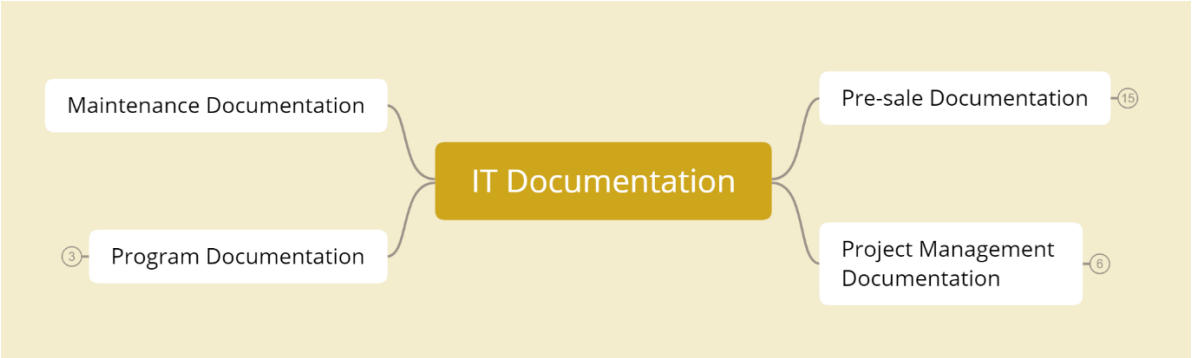

Figure 2.1–Types of the IT Documentation

2.2 The aim of the project documentation

Currently, no one needs to be convinced that during the creation of the automated system for its further effective operation and development, it is necessary to develop a set of design documentation. However, during the development process and at the first stages of the system's operation, when the developer decides all the issues of implementation, configuration and maintenance, the Customers sometimes do not pay enough attention to the design documentation, which entails risks for the system's operation in the future, including its possible early replacement.

Difficulties in the operation of the system may arise when the design documentation is missing or developed «formally» at the same time:

- the development company no longer supports its system;
- technical specialists who accompanied the system no longer work in the

company;

- third-party specialized companies do not want to take someone else's system to support.

In such conditions, it is very difficult to fully maintain and operate the system, since the specialists involved (company`s or third-party organizations). So it will not be able to quickly realize the task to solve the problems that arise or develop the functionality of the system. If there is a detailed description of all design and organizational solutions for the creation and functioning of the system, the implementation of these tasks becomes more likely.

Thus, if there is a full-fledged set of documentation for the system, the customer will be able to further effectively independently develop the functionality and operate the system, minimizing the dependence on his specialists and third-party developers and operating organizations.

2.3 Task of the project management documentation

As we are able to perceive that project documentation is applicable all told areas throughout a project. This is part of all areas of knowledge that covers all stages of the process. A project manager must always provide due importance to sensible documentation. Project success is outlined by meeting project objectives inside project constraints. However, to assure success, you would like the support of excellent project documentation. The success of the project depends not only on the preparation of the finished documentation but also on the changes, reviewing and re-consideration. change documents is additionally an important part of honest documentation. It provides you with a period of time data to create excellent selections. It prepares you to require any course correction, whenever an associated anomaly is proved.

The project, as an object of management, has a set of features that require the use of special techniques and methods for managing it. Over the past forty years, project management has emerged as a special professional field and an

independent discipline that equips project managers with technologies and tools for planning, monitoring and coordinating project implementation. There is no doubt that the project is an important part of project management. Even small firms implementing relatively small projects are increasingly beginning to systematically approach the preparation, planning and monitoring of their projects using project management methods and tools.

The two main functions of the documentation confirm this: make sure that the requirements of the project are met, and establish traceability as to what was done, who did it and when it was done. The application of methods and means of project management allows not only to achieve the project results of the required quality but also to save money, time, other resources, reduces risks and increases reliability, as it helps to:

- determine the objectives of the project and carry out its justification;
- identify the structure of the project (subgoals, main stages of work, etc.);
- determine the necessary volumes and sources of financing;
- to select performers, in particular, through bidding and tendering procedures;
- prepare and conclude contracts;
- determine the timing of the project, draw up a schedule for its implementation, calculate the necessary resources;
- make a calculation and cost analysis;
- plan and consider risks;
- organize the implementation of the project, including selecting the «project team»;
- provide control over the progress of the project.

The choice of appropriate project management methods and tools is determined, first of all, by the complexity, scope and type of project. Moreover, the main difficulties, in the general case, arise in the initial phases of the project, when the main decisions must be taken, requiring non-traditional methods and means.

As we have now understood the importance of documentation, it really helps a project manager during projects. There are many indirect advantages of project documentation. For example, this is one of the development methods for project managers. There is no skip to the documentation if your goal is to aspire to be a promising project manager. Project documentation helps to deal with upcoming situations with ease. It sets the platform for thorough communication and understanding. Another argument is that documentation helps a project manager to make sound and informed decisions because it paves the way for your project success indeed. Writing down their successes and worst incidents, they are given an excellent opportunity to learn from their mistakes. Thus, good documentation makes them a reliable resource and increases their credibility among project management professionals.

2.4 Types of the Project Documentation

Project documentation serves the purpose of providing reference of something done. Therefore it is very important for the success of any project. There are many project documents created through the life cycle of a project. As an example of the most requested documents developed at all project stages, the following can be cited in the Attachment D.

As mentioned earlier, all project documentation for IT projects can be divided into several types: Pre-Sale Documentation, Program Documentation, Project Management Documentation and Maintenance Documentation. Each of these types includes a certain set of documents.

Pre-sale stage includes activities such as meetings, negotiations, presentations, identification of current and potential customer requirements, as well as preparation of the organization of the project. Therefore, for further effective work of the team, it is necessary to have legal documents and documents related to the business analysis of the client. First of all, these are prepared requirements for the project, which can be provided as technical specifications, or

broken down into separate documents (Product Vision, User Roles, Feature list, Glossary, Assumptions). The evaluation of the project should also be documented, as it often depends heavily on what type of contract will be. Depending on the project, business analysis documents may be supplemented by Diagrams (Use Case Diagram, ERD, IDEF0, IDEF1), User Stories and the like. Signing a contract and further billing is also negotiated and documented at the Pre-sale stage.

Project Documentation and its contents directly depend on which methodology was chosen for project development. This is because each methodology includes a different set of documents. Despite this, the most common document can be called Project Plan, since it is the basis and contains all the basic nuances. Initially, this document was mainly used for the Waterfall methodology, but with the advent of flexible methodologies, it can now be applied to almost any methodology.

It should also be noted that a hybrid methodology may include any of the listed documents, since this type of flexible methodology may be a combination of the other two.

Program documentation is project-specific and usually consists of documents related to development, design, and testing. These can be mockups, prototypes, a description of the system settings, code review, Test Cases, Test Plans, Bug reports and other documents related to the project.

Maintenance Documentation is the final documentation submitted directly for exploitation by the user: Specification, Source code, Technical documentation, User Guides. This set of documents also varies, which once again proves that the documentation in IT projects is very dependent on the type of project and subject area.

The careful planning and implementation of the project are based on the talented managers of projects and correct methods and instruments. There are also a few key documents, that allow the sponsors, managers of projects, commands and parties concerned carefully and exactly to manage necessary project actions. The list of these documents is below given, and why each of them is needed for

management by project activity and grants of necessary results.

As was mentioned before, all the project documents are essential and they serve different purposes. All more detailed dependencies of the project documentation are presented in the Attachment B. Project documents must meet the traceability and quality requirements. They must be well arranged and easy to read. Understanding the function of each document provides success in project management. In addition to that establishing a proper document control system is as important as creating and using project documents.

2.5 Content of the Project Plan

A project plan can contain many different components. This section will cover the most common components that are commonly used in a project plan.

The project management plan is a comprehensive document defines how the project is planned, executed, monitored, and controlled. The project management plan includes but not limited to management plan, configuration management plan, baseline schedule, and cost.

It is not a surprise that the biggest document of the project documentation is the most important one. That is because this document used to describe every phase of a project. The components may include initiating, planning, executing, monitoring and controlling, and closing. But the description of any phase depends very much on the chosen development methodology and, accordingly, the content of the project plan. In Figure 2.2, you can see the main elements that should be contained in the plan: Project overview, Scope of the project, Roles and Responsibilities, Key stakeholders, Key milestones, Communication plan, Budget plan, WBS, Risk Matrix, Technical methods tools and techniques [20]. Despite this, none of the points is strictly mandatory, since in the flexible development, which is used in most cases today, the absence of any points in the project plan is not critical.

Figure 2.2 – Components of the project plan

One of the first points in the project plan is Project overview. This item helps to catch the main idea of the project, reveals the main goals and objectives of the project and contains a brief description of it. This short section helps determine the overall vision of the project from both the user and the client.

At the very beginning of the consideration of planning processes, the initial step in the process is to determine the scope of work. At first, their sequence is out of the question. It is important to collect in one place in a structured form the whole set of actions that must be carried out in order for the project task to be fully completed. The first important tool for determining the tasks of a project is a hierarchical structure of work. Theoretically, creating this document does not seem complicated, but in practice, it is specific and requires taking into account the mass of nuances. The hierarchical structure of the work is a breakdown of the project into specific results that must be achieved to achieve the goals of the project. It is important to understand that it is the work results that are collected in WBS, and not the tasks that must be performed to obtain these results.

A Scope of the project document is an agreement on the work you’re going to perform on the project. Project scope refers to the detailed set of deliverables or features of a project. These deliverables are derived from a project’s requirements. The planning process is when an attempt is made to capture and

define the work that needs to be done. In the final process, the closing includes an audit of the project deliverables and an assessment of the outcomes against the original plan.

To clarify the project key milestones should be prescribed. A project milestone is a task of zero duration that shows an important achievement in a project. The milestones should represent a clear sequence of events that incrementally build up until your project is complete. A milestone can be, for example, a transition to a new stage of the project

When designing or changing processes, it is necessary to organize responsibility and the relationship between the roles involved in the process. The Roles and Responsibilities item is for this purpose intended. In some cases, the RACI Matrix may be used instead. Projects require the participation of many people. This document can prevent a situation when people are fighting with each other for the task. No less difficult is the situation when no one becomes the owner and makes decisions, therefore this item is very important.

Equally important is the prescribing of Key stakeholders and Communication plan. In which should be indicated all the contact details of the persons involved in the development of the project, as well as ways of communication between the team and the client.

Project budget – is a plan of costs required for its implementation, in value terms. The project budget includes expenses for the purchase of materials, payment of wages (including contributions to social funds), services of third parties, depreciation of buildings, machinery, equipment and intangible assets. As a rule, the budget is formed in the context of the stages of the project areas of work, the implementation of which is controlled individually. The main parameters affecting the project budget are duration of work, number of participants and equipment used, as well as specific requirements for the result [19].

Moreover, the project manager needs to identify the existing risks and lay a reserve in the project budget. To do this, there is a registry – Risk Matrix,

designed to track the status and likelihood of risk.

The paragraph called Technical methods tools and techniques which may contain additional information about technologies, libraries, software and other aspects that affect the purely technical part.

So, the project management plan is the most important document when creating a project–at the level of stakeholder involvement. Just as a project cannot be implemented without the participation of a stakeholder, so without a well-thought-out management plan, the project will fail.

2.6 Project`s Domain Knowledge

Each project relates to a specific area in which it will be operated in the future domain knowledge. Domain refers to a broad-based understanding of a particular industry or solution. Domain knowledge is now becoming very important in the IT software industry and incredibly valuable for software development as well. It helps to understand the system, which technologies should be used to solve the tasks, which specialists do the work and how much time they will need to complete certain actions. When considering a project from the side of a company, it is quite easy to determine if a project belongs to a certain domain with the help of experts. The knowledge about any sphere builds on the experience, so the more people work with the code, the more areas of documentation they confronted with, the more domain knowledge they get. That’s why the person that has been on the team the longest typically knows the most about the system.

At the moment, we can distinguish the following prevalent domain areas, which you can overview in Attachment C. This is a fairly large list of common domain areas that can be attributed to the development of applications, sites and platforms. All domains can be conditionally divided into two categories: specific domain knowledge and technical domain knowledge. Specify domains are special knowledge about the particular business field, whether technical domain

application skills of the appropriate technology. Technical domain knowledge include such business arias as Artificial Intelligence, BigData, Blockchain, Computer Vision, Customer relationship management (CRM), Environmental Resources Management (ERM), Material requirements planning (MRP), Employee Request (ER), Cryptocurrency, Extensions, Plugins and Internet of Things (Figure 2.3). Whereas special domain knowledge consist of business sectors as Art and Design, Banking and Finance, Beauty and SPA, Car industry, Chemical industry, e-Commerce, e-Learning, Entertainment, Film industry, Food and Drinks, Games, Healthcare, HR and Recruiting, Law, Real estate industry, Shipping and Transportation, Social services, Sports and Outdoors activity, Telecommunication, Travel and Booking, Media tools and other Business apps and services. In any case, any domain structure can be transformed and is variable. Therefore, the structure shown above can be ordered or rearranged with the addition of new areas.

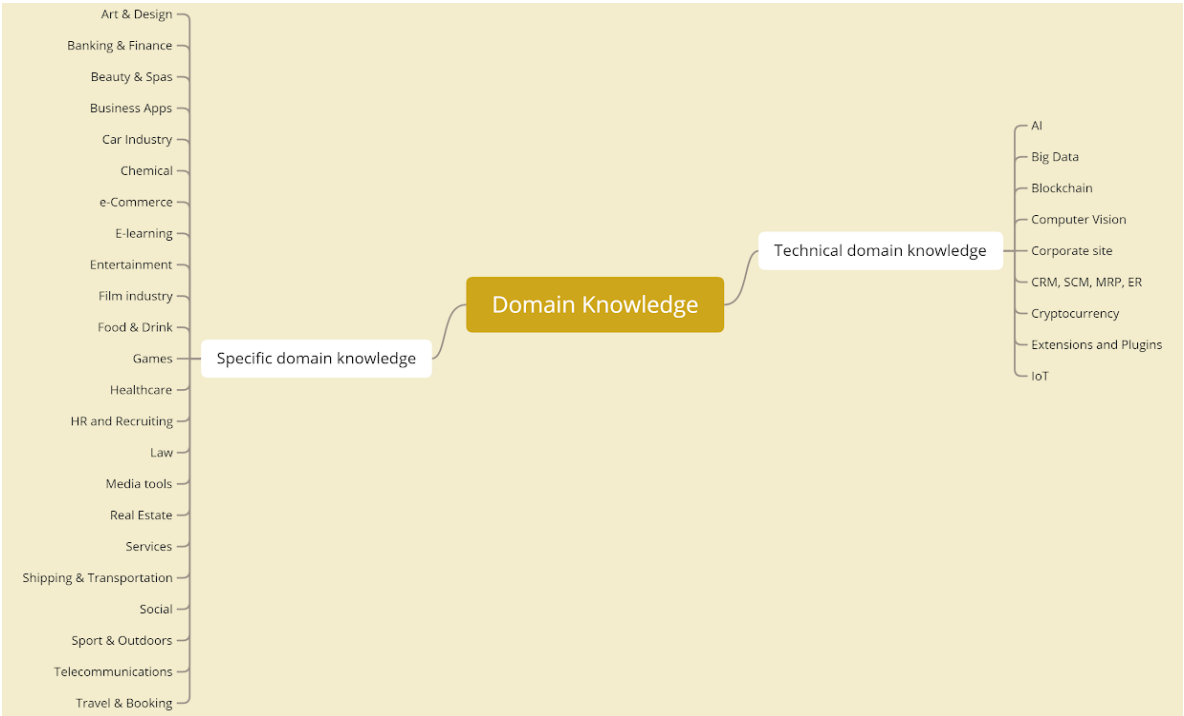

Figure 2.3 – Types of the main domain knowledge for IT Projects

If the number of projects is growing rapidly, determining the domain of the project becomes more difficult. And therefore, there are several reasons, but the most important problem is extremely costly since the expert has to read the project documentation and highlight the main points in order to determine whether he belongs to a particular class.

It is very important to identify the domain and there are some reasons why. Firstly, it is essential to enhance decisions, services, and products by adding intelligence, increasing value and providing flexibility. Secondly, the creation of the learning organization to sustain a competitive position in the market is also critical [18]. Another reason to identify domain is to provide better job security as a specialized resource and help to get into the niche market and stay in demand.

However, other problem may cause difficulties in determining ownership to the aria, because the domain is nothing but knowledge of any particular sector, specific industry or nature of business. Those cases help to understand from the above example that domain knowledge is useful in creating features to apply the classification algorithms.

3 RATIONALE FOR THE CHOICE OF THE ALGORITHM

3.1 Multinomial Naive Bayes Algorithm

Naive Bayes is a group of probabilistic algorithms that benefit from applied mathematics and Bayes' Theorem to predict the tag of a text. they're probabilistic, which suggests that they calculate the likelihood of every tag for a given text, so output the tag with the best one. The method they get these chances is by applying of Bayes' Theorem, that describes the chance of a feature, supported previous information of conditions which may be associated with that feature [8]. Naive Bayes area unit largely used in natural language process (NLP) issues. Naive Bayes predict the tag of a text. They calculate the chance of every tag for a given text so output the tag with the best one.

On the figure 3.1, we can see that the probability of the event A given B is the probability that A and B (intersection) occurred divided by the probability that B happened.

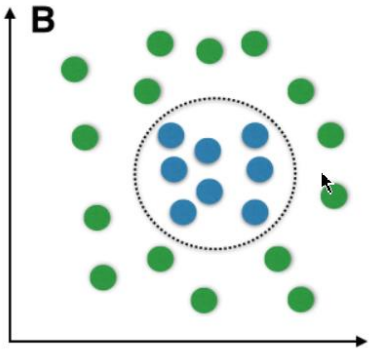

Figure 3.1– non-linear classifier

So that we can get:

$$P(A\,|B) = \frac{P(A \cap B)}{P(B)}, \tag{3.1}$$

Following the previous formula can be obtained the probability that B has happened given that A has also happened. Also, we should remember that from the Venn diagram:

$$P(A \cap B) = P(B \cap A), \tag{3.2}$$

In such a way, by equating formula (3.1) and it's inverse for A event we can get the Bayes theorem:

$$P(A\,|B) = \frac{P(B\,|A)\,P(A)}{P(B)}, \tag{3.3}$$

with these definitions where P(A) – the prior probability of A occurring independently;

P(B) – the prior probability of B occurring independently;

P(A|B) – the farther probability that A occurs given B;

P(B|A) – the likelihood probability of B occurring, given A.

The theorem refers to the strong independence assumptions in the model, rather than the particular distribution of each feature. A Naive Bayes model assumes that each of the features it uses is conditionally independent of one another given some class. More formally, if I want to calculate the probability of observing features f_1 through fn, given some class c, under the Naive Bayes assumption the following holds:

$$p(f_1, \dots, f_n\,|c) = \prod_{i=1}^{n} p(f_1|c). \tag{3.4}$$

This means that when I want to use a Naive Bayes model to classify a new example, the posterior probability is much simpler to work with:

$$p(c|f_1, \dots, f_n|c) \propto p(c)p(f_1) \dots p(f_1|c). \quad (3.5)$$

In alternative words, we've got left p(fi|c) indefinable. The term Multinomial Naive Bayes merely lets us recognize that every p(fi|c) is a multinomial distribution, instead of another distribution. This works well for knowledge which may simply be changed into counts, like word counts in text. To summarize, Naive Bayes classifier is a general term which refers to conditional independence of each of the features in the text, while Multinomial Naive Bayes classifier is a specific instance of a Naive Bayes classifier which uses a multinomial distribution for each of the features.

3.1.1 Tagging

At a basic level, the classification process is simplified. Data classification is of specific importance once it involves risk management, compliance, and information security. And in order to make the data easily accessible for search and tracking, there are data tags as a way of classification. Using tags can reduce storage and backup costs, speeding up the search process.

Tags are a type of meta-information transformations that represent or define some aspect of information, namely textual [10]. Knowledge tags are more than direct non-hierarchical words or terms; they are a type of metadata that collects knowledge in the form of descriptions, classifications, classifications, semantics, comments, notes, annotations, hyperdata, hyperlinks or links that are collected in tag profiles (a kind of ontology). Different types of knowledge can be captured in knowledge tags, including factual knowledge, conceptual knowledge, expected knowledge, and methodological knowledge (results from reasoning and strategies). In most cases, these forms of knowledge exist outside the data itself and are based on personal experience.

Profile tags relate to an information resource located in a distributed and heterogeneous storage. Therefore, this type of data is often the starting point for

many enterprises, followed by additional procedures and markings that determine the data based on their relevance to enterprises, quality and other classifications.

3.2 Types of the document for classification

The structure of this document consists of five main parts. The first one are documents which we get on the Pre-sale stage: Requirements, Estimation, Diagrams, User Stories (Figure 3.2). Contract and invoice are also part of the project documentation and analysing it we can calculate correlation between the type of contract and the benefits of the project or get useful detail information about the budget Development documents are also very important part of the project documentation. It is because there we can find already done cases which will facilitate work on a future project. In management documents, you can also apply classification to obtain any specifics and quickly process large texts. For example search for an assessment of a specific task in the list of the scope of work document or WBS or one and the other at the same time.

But the most interesting document for application of the classification is Project Plan documents and there a few reasons why. The first one is that Project Plan is the biggest and the most informative document. It give us confidence that we could get from it maximum. Basing oneself upon data and knowing all the sources in the plan we can easier predict something on the pre-sale stage of the future project. For example we can define the quality of a specific team performing a certain kind of project. If we know that the problem was in Front-end team we should think about rotating it or something else [17].

But in case of my diploma project I would like to consider more common and everyday problem: correlation between Project Plan and domain knowledge. This process definitely requires automation, as it takes a lot of time and effort. But also this task is very common and subsequently used often. To determine the belonging of the project plan to any of the domain areas requires the presence of an expert in a particular area, which is not always possible, as well as a correctly

written project plan. To solve the classification problem, consider the design documentation of a specific IT company. The moment of interest, namely the subject area, mention of words, or the general idea of belonging to any of the domains can be mentioned in any part of the project documentation. Consider this with the example of a Project Plan.

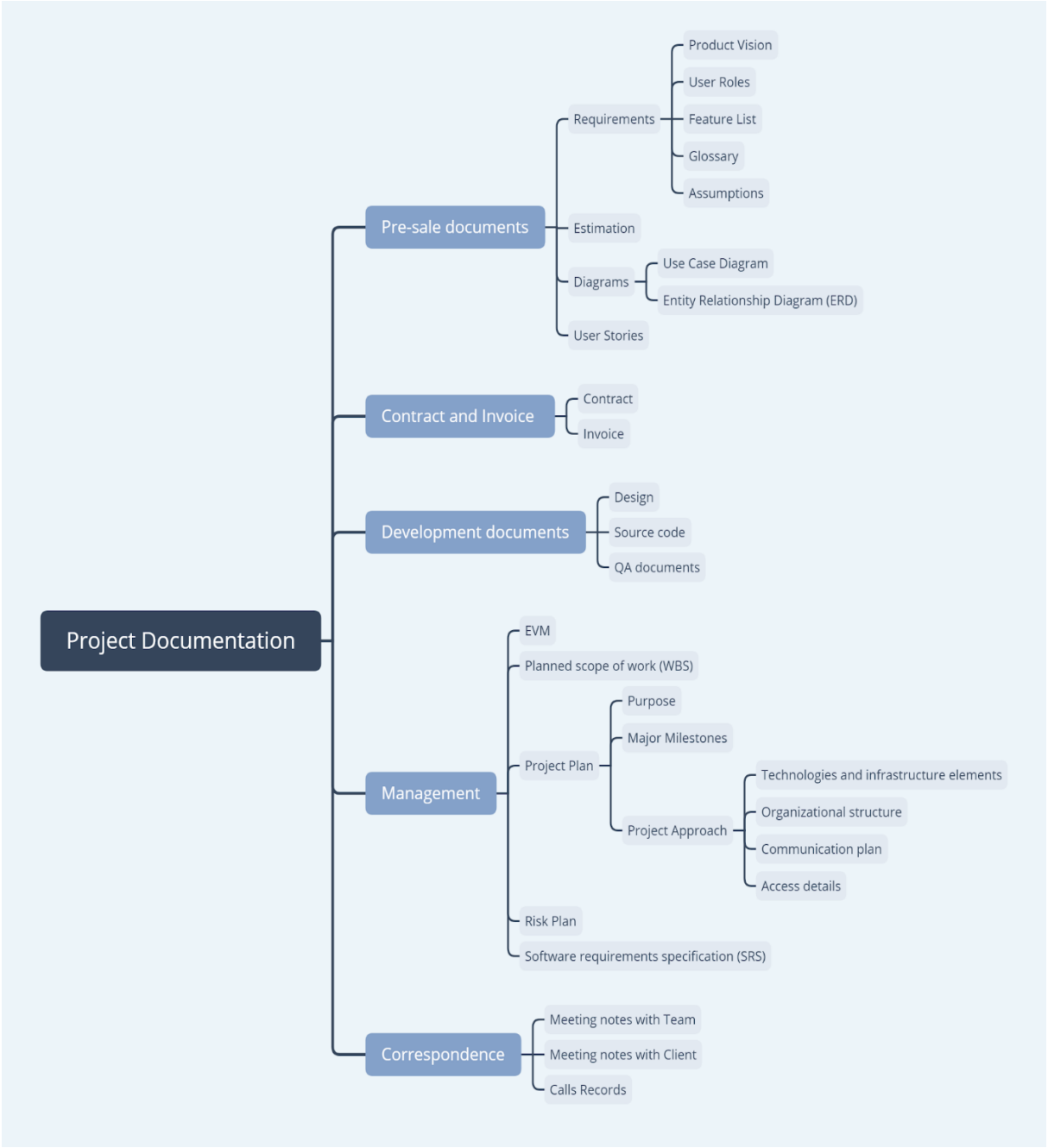

Figure 3.2 – Project Documentation of the IT Software

For a well-executed project plan, each company should have a certain structure of the document itself. As part of my diploma, a document with the following structure will be considered. The table 3.1presents in more detail its components.

Table 3.1 – Structure of the project plan

Section of the Project Plan		Content
Purpose		A brief description of the project and its main objectives.
Major Milestones		Prescribed milestones of the project, based on a preliminary assessment. Must be presented in the form of a list with the specified dates.
Project Approach	Technologies and infrastructure elements	Used technologies, technological approaches, architectural solutions. Data should be presented in tabular form.
	Organizational structure	A list of all project participants, presented in the table, including the team, client and all stakeholders interested in the project.

Continuation of the table 3.1

	Communication plan	The information is presented in the form of a table that contains the following fields: Communication channels, Email, Communication with client & Escalations, Internal communication with a team.
	Access details	Necessary access to the project, for example, access to the server or time tracking system.

This structure is common for the whole company, that is, all project plans, regardless of the type of project, should be stored in such a structure to simplify the classification algorithm. The structured view of the document not only facilitates the search for the necessary information manually, but also allows the algorithm to quickly find the necessary information and skip the one that does not matter in this task.

3.3 Programming language Java

Until a certain time, the program development process did not require the selection of specific programming technology. There was either no technology at all, or there were tools focused on a specific development. Currently, the situation has changed a bit. According to constant representations, the concept of «software development technology» includes a system or set of tools for the development of a particular software and a technological process that governs the organization and conduct of these works. All kinds of technologies and tools have appeared in the modern market, with similar characteristics and capabilities that support them.

This poses a problem for developers to choose a technology that is largely related to the choice of software development automation tools. That is why the choice of development tools should be approached with particular attention. To develop the components of the media content exchange service, it is necessary to choose technologies and tools that will be most effective and satisfy the points of the problem statement.

Technological as well as informational progress, especially in recent years, has taken a huge step forward. Modern humanity is rapidly turning into an information society, and this is especially happening in rapidly developing countries, which place great emphasis on the development of technology and information technology. That is why the programming language for developing a service must be multifunctional and have many features. One such multifunctional programming language is Java.

Java is an object-oriented programming language developed by Sun Microsystems since 1991 and officially released in 1995. Initially, the new programming language was called Oak and was developed for consumer electronics, but was later renamed Java and was used to write applets, applications, and server software.

A distinctive feature of Java in comparison with other general-purpose programming languages is the provision of high programming productivity, rather than the performance of the application or the efficiency of its use of memory (Figure 3.3).

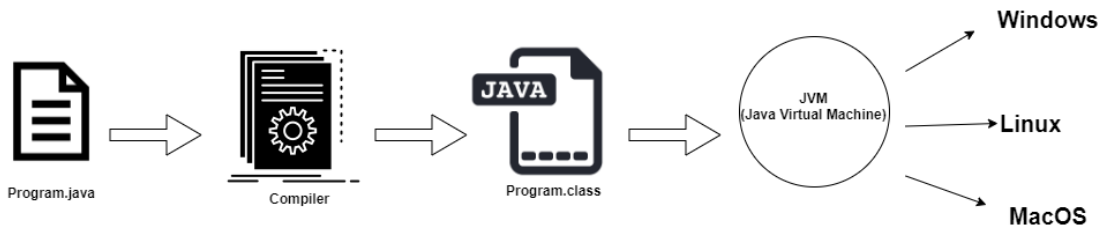

Figure 3.3 – The principle of the Java program

This programming language is characterized by complete independence of bytecode from the operating system and hardware, which allows you to run Java applications on any device for which there is a corresponding virtual machine. An important feature of Java technology is its flexible security system, in which the execution of the program is completely controlled by the virtual machine. Other advantages of this language are noted, such as:

- platform independence. Unlike many other languages, including C and C ++, Java, when it was created, did not compile on the platform of a particular machine, but on a platform independent platform. This byte code is distributed over the Internet and interpreted in the Java Virtual Machine (JVM), which it currently runs on;
- simplicity. Learning processes and introducing the Java programming language remain simple. It is enough to understand the basic concepts of object-oriented programming;
- security. Authentication methods are based on public key encryption;
- architectural neutral programming language. The compiler generates architecturally neutral objects of the file format, which makes the compiled code executable on many processors, with the presence of the Java Runtime system;
- sequential. Performs efforts to eliminate errors in various situations, focusing mainly on compilation time, error checking, and runtime checking;
- multithreading. Multithreading functions, you can write programs that can perform many tasks simultaneously. Introduction to Java of this design feature allows developers to create debugged interactive applications;
- interpretation. Java bytecode translates on the fly to machine instructions and is not stored anywhere. Making the process faster and more analytical, since binding occurs as additional with a small weight of the process;
- high performance. Just-in-time compiler, allows you to get high performance;
- prevalence. The presence of a large number of third-party libraries;
- dynamism. Java programming is considered more dynamic than C or

C ++, as it is designed to adapt to changing conditions. Programs can execute a vast amount during the processing of information, which can be used to verify and allow access to objects at runtime;

- the ability of the Java language to support different platforms and operating systems.

Java features, ease of use, platform independence and built-in security features make this programming language one of the best for creating various applications.

3.4 Rationale of choosing a Development Environment

To implement the classification algorithm, the IntelliJ IDEA development environment was chosen an intelligent integrated Java development environment focused on developer productivity, which provides a robust combination of advanced tools.

Advanced project navigation code structure simplifies the analysis of even large files, providing a convenient way to view them. The search recognizes language elements by showing results in an interactive navigation bar, helping to analyze all found files.

The first version of IntelliJ Idea appeared in January 2001 and quickly gained popularity as the first Java IDE with a wide range of integrated refactoring tools that allowed programmers to quickly reorganize their source code. The design of the environment is focused on the productivity of programmers, allowing them to concentrate on the development of functionality, while IntelliJ Idea takes on the execution of routine operations.

Starting with the sixth version of the product, IntelliJ Idea provides integrated tools for developing a graphical user interface. Starting with version 9.0, IntelliJ Idea is available in two versions: Community Edition and Ultimate Edition. Community Edition is a fully open-source version available under the Apache 2.0 license. It implements full support for JavaSE, Groovy, Scala, as well

as integration with the most popular version control systems.

The Ultimate Edition version supports JavaEE, UML diagrams, code coverage counting, as well as support for other version control systems, languages, and frameworks. Thus, the capabilities of this programming environment fully comply with the requirements when developing the algorithm.

4 SOFTWARE SOLUTION OF THE CLASSIFICATION PROBLEM WITH MULTINOMIAL NAIVE BAYES

4.1 Identify the prerequisites to train a Naive Bayes classifier

The algorithm that will be considered is called Multinomial Naive Bayes. I will analyze in more detail the algorithm applied to NLP. As shown earlier, the use of a Bayesian classifier to classify text is infinite. The only prerequisite is the existence of an existing set of examples for each domain area in which we want to place text fragments.

Some of the application of the Naive Bayes algorithm is:

- categorization of products. Assign product into some categories such as a Mobile app, Web-site, Plugin by description;
- tagging scope of work. Determining of the belonging to any category of the raw text, so we can assign it tags such as «development», «testing», «estimation»;
- mail Classification. Classify massages in the inbox category into folders such as Updates, Shopping, Work, Inportartant, etc.

In my thesis, I will use exactly tagging, so I can use the words within a document as «features» to help to recognise the classification of a document. In supervised methods of document classification, a classifier is trained on a manually tagged dataset of documents. The classifier can then predict any new document category and can also provide a confidence indicator [7]. The biggest factor affecting the quality of these predictions is the quality of the training data set.

In order to put the algorithm into practice it is necessary to define tags in advance, that is, classes of domain domains (Figure 4.1). Further, on the basis of existing documents to determine their belonging to classes. This is presented in the table 4.1.

Figure 4.1–A simple Illustration of Document Classification

In this case, we will consider the algorithm using the example of the IoT tag, which will show whether the project will belong to this domain based on the project plan.

Table 4.1 – Examples of the block description in the project plan

Plan name	Text	Tag
P1	Web application for streamlining document flow of real estate transactions. The project includes 2 parts: Web application and Admin panel. To create a Web application which allows uploading documentation concerning the estate transactions and deals. The web application includes authorization, profile/documents, etc. Creating Admin panel which allows browsing current documents, list of deals, user's permissions, etc.	Services

Continuation of the table 4.1

Plan name	Text	Tag
P2	IoT based smart system is regarded as IoT gadget focusing on Live Monitoring of Environmental data in terms of Temperature, Moisture and other types depending on the sensors integrated with it. The system provides the concept of «Plug & Sense» in which farmers can directly implement smart farming by as such putting the System on the field and getting Live Data feeds on various devices like Smart Phones, Tablets etc. and the data generated via sensors can be easily shared and viewed by agriculture consultants anywhere remotely via Cloud Computing technology integration. The system also enables an analysis of various sorts of data via Big Data Analytics from time to time.	IoT
P3	The target platform has to serve as a private exchange for the end-users (brokers) of trading agencies with a number of brokers 1000-10000. The private exchange should provide the ability to trade with equities, currencies, generally should be flexible on adding any new instrument for the brokers. The exchange should also be integrated with exchange using Ayers API and provide the possibility to integrate with other exchanges easily.	Banking and Finance

Continuation of the table 4.1

Plan name	Text	Tag
P4	App for iPad related to a medical centre that researches cancer diseases and monitors health condition. Doctors can use this App for researching patients as well as provide them with treatment. Also, this App will be connected with Apple Watch, where patients would be able to estimate their health.	Healthcare

Because of Naive Bayes is a probabilistic classifier, we need to calculate the probability that Project plan P1 is Services and the probability that it is not Banking and Finance, IoT and Healthcare. The first thing we need to do when creating a machine learning model is to decide what to use as functions. Fragments of information are features that we take from the text and pass to the algorithm so that it can work with them and take as an example.

For an example of the classification of project plans, we need a set of job lists that are known for being specially labeled to describe medical projects, a set of project plans that are known to describe projects in the service field. In this case, there are two classes of domain domains. With examples of project plans in each category, we can train the naive Bayesian classifier so that new plan files are automatically classified. The only prerequisite is the availability of an existing training kit (Figure 4.2).

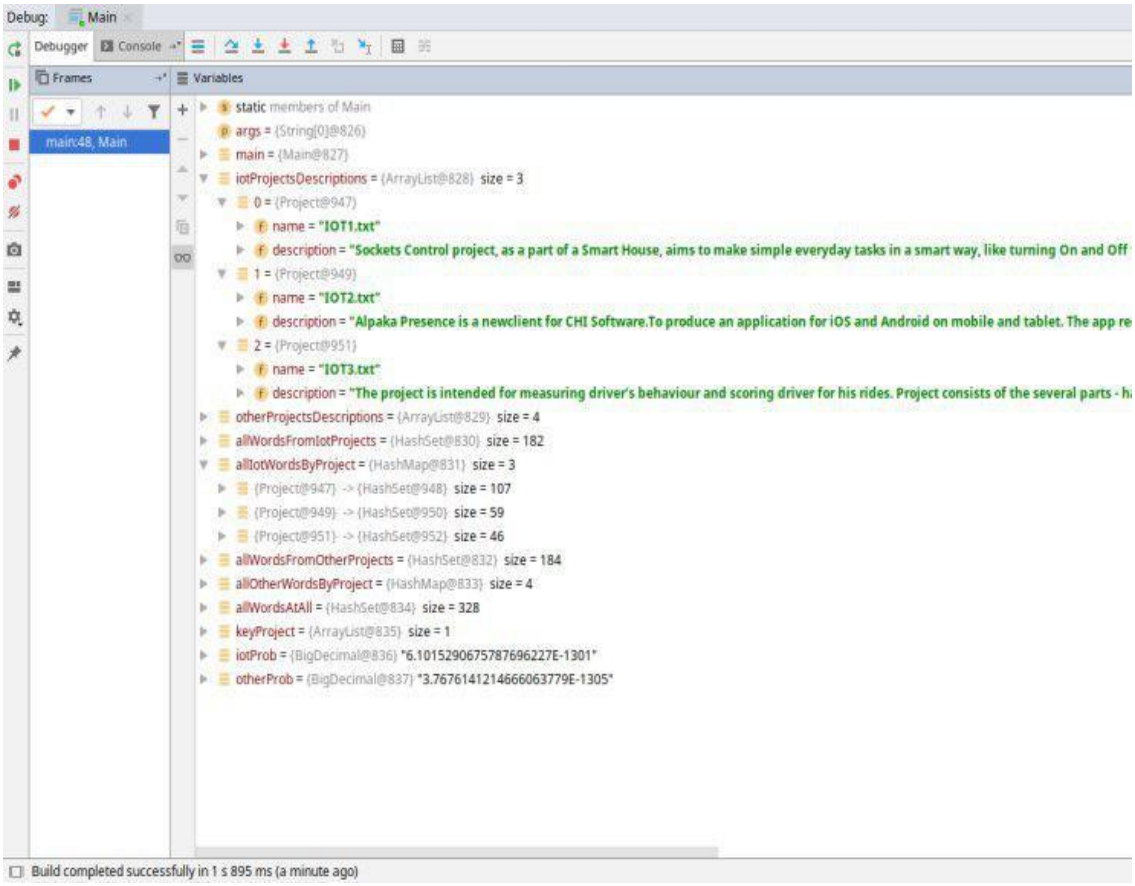

Figure 4.2 – Training kit

In this case, there are no numerical functions, there is only text. Thus, the input is multiple source files. The goal is to convert this text into numbers, from which calculations can be made further. To do this, it is necessary to use word frequencies, that is, we ignore the word order and sentence construction, considering each document as a set of words that it contains (Figure 4.3). Our features will be the counts of each of these words. Although this may seem like an oversimplified approach, despite this the method works surprisingly well.

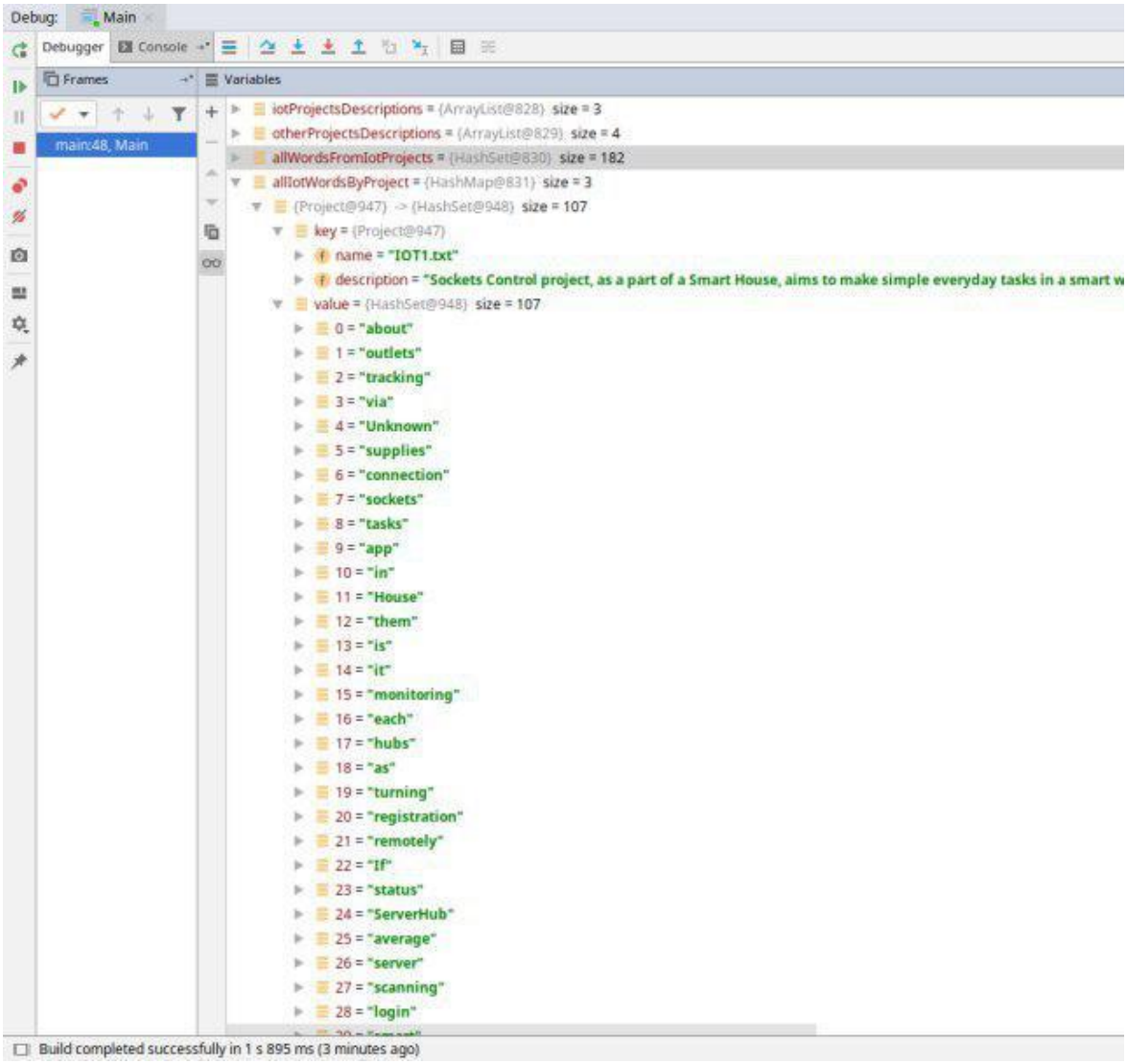

Figure 4.3 – Dividing a document into words

A terminology document matrix (TDM) consists of a list of word frequencies found in a set of documents. In this case, these are words belonging to various business and technical domain domains. The TDM matrix is a sparse rectangular matrix of n words and m documents. They say that it is rare because it contains mostly zeros. The TDM matrix entry (i, j) represents the frequency of the word «i» in the document «j». Regarding the number of examples in the class, the general rule is this: the more the better. However, in practice, several thousand messages give reasonable predictions.

From raw frequency count, you can infer that IoT Project Plans contain in high frequency terms such as «detection», «sensor» or «home automation»,

whereas Services Project Plans contain in high frequency terms such as «tool» or «plugin» (Figure 4.4). This is how the classifier will be able to tell apart one class from the other and exactly the same with other classes.

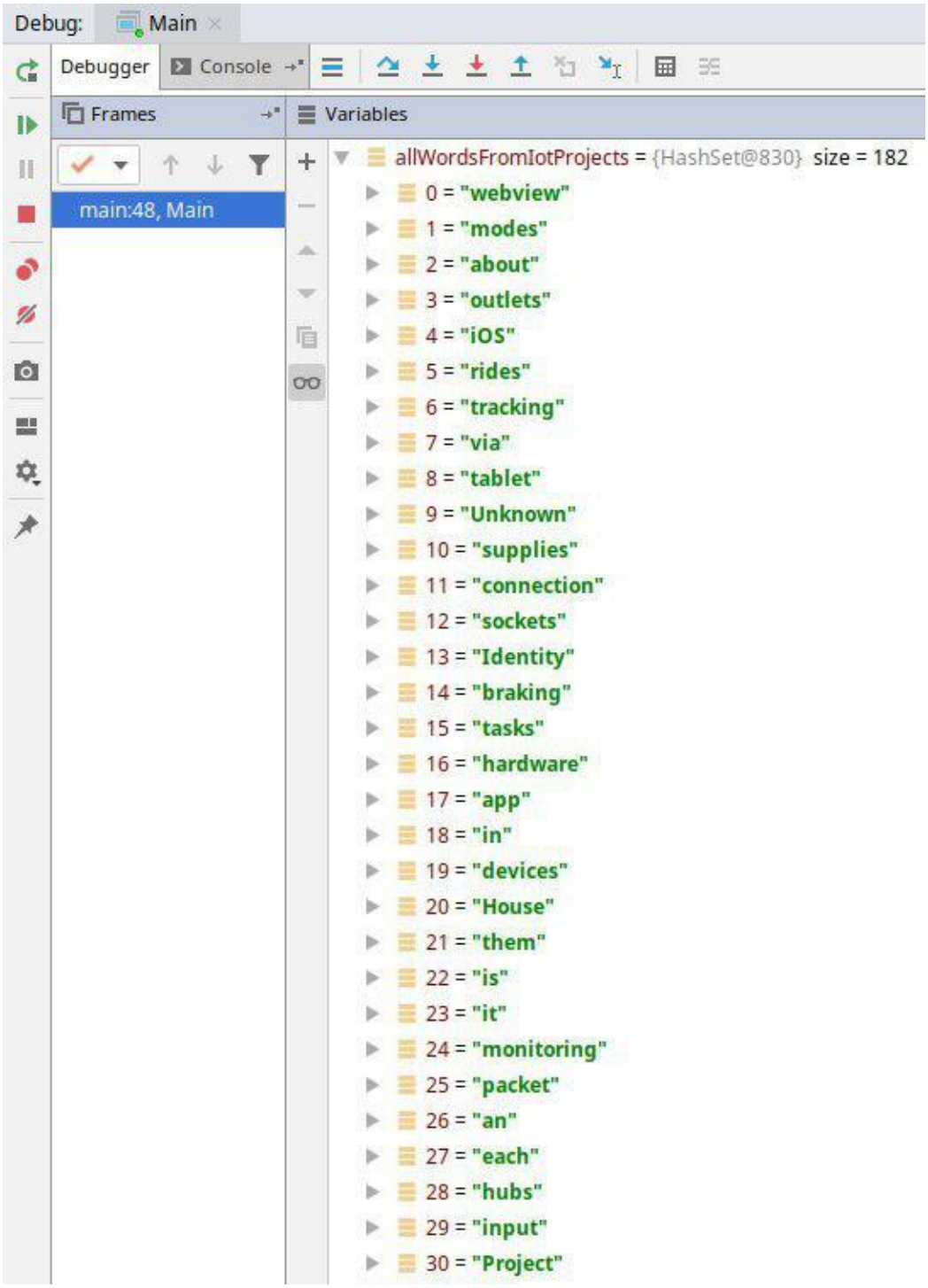

Figure 4.4 – The stage of combining words into a single array of IoT domain

4.2 Naïve Bayes Theorem

Next, you need to convert the probability that we want to calculate into something that can be calculated using word frequencies. For this purpose, at the previous stage of text processing, all the words contained in the documents were separated from each other in order to find out the exact list of a unique list of words for a particular subject area. The figure below shows the next step – combining all the texts to create a unique list of words. Thus, we can know the exact number of unique words in each of the domain domains (Figure 4.5).

Figure 4.5 – A unique list of words of IoT domain knowledge

To determine the frequencies of words, we will use some basic properties of probabilities and Bayes' theorem. Bayes' theorem is useful when dealing with conditional probabilities because it gives us a way to invert them.

Now we need to calculate the probability that the project plan «Smart Project» is «IoT». Then we take the largest. From a mathematical point of view, we want P(IoT | Smart Project) to be the probability that the offer tag is the Internet of things, given that the project plan is "Smart Project". Kespectively, we submit that plan to the input.

$$P(IoT \mid x) = P(x \mid IoT)\ x\ P(IoT) / P(x) \quad (4.1)$$

Where x is a feature vector which consist of words coming from different Project Plans:

$$x = [w_1, w_2, w_3, \ldots w_n] \quad (4.2)$$

Suppose each word in a sentence is independent of the others (Figure 4.6). This means that we no longer look at whole sentences but at individual words. Therefore, for our purposes, there is no value in the permutation of words.

This is naivety. The result is that «probability» is the result of an individual probability of seeing each word in the plan, regardless of which business area the project was completed in. So that we can write this:

$$P(Smart\ Project) = P(webview) \times P(models) \times \ldots P(w_n) \quad (4.3)$$

This makes this model work well with a small amount of data or data that may be incorrectly labelled(for example word «about» is not related to the IoT domain. The next step is to simply apply this to what we had before:

$$P(Smart\ Project|IoT) = \quad (4.4)$$
$$= P(webview|IoT) \times P(models|IoT) \times \ldots P(w_n|IoT)$$

All of these individual words actually show up several times in our training set and the next step is calculating it.

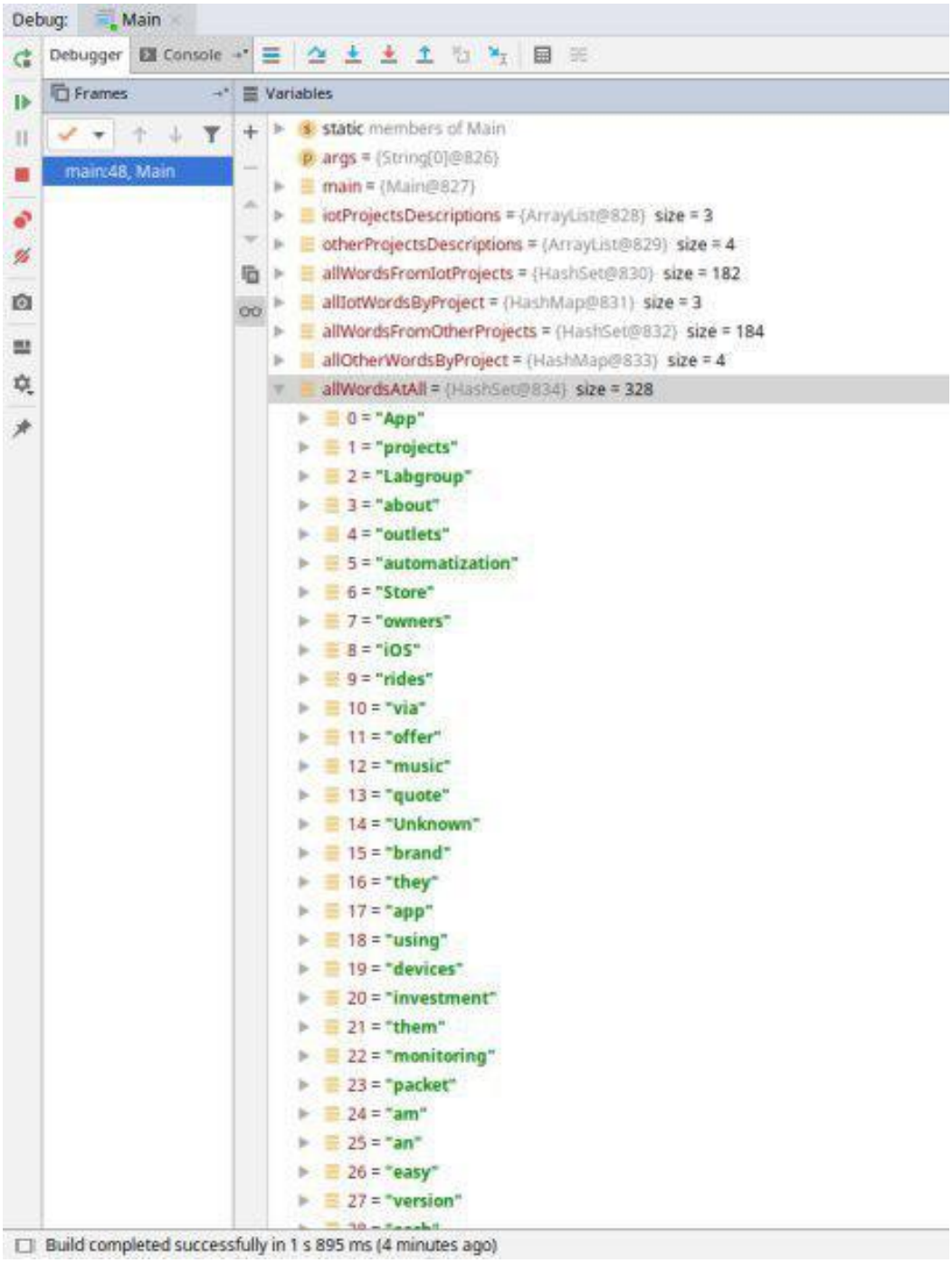

Figure 4.6 – A list of words of all domain knowledge

4.3 Calculating probabilities

Given that a database of probabilities for words appearing in IoT and Services project plans were computed we can proceed to the last step of the Naive Bayes Classifier, which is the classification.

The formal decision rule is:

$$\hat{y} = \underset{k \in \{IoT, Services\}}{argmax} p(C_k) \prod_{i=1}^{n} p(x_i|C_k), \tag{4.5}$$

What it means is that for every project plan or another type of the documentation in these domains we have to compute the probability of such of it being IoT and Services and our final veridic will be given by the largest probability. Consider the following Project Plan as an example is on the figure 4.7.

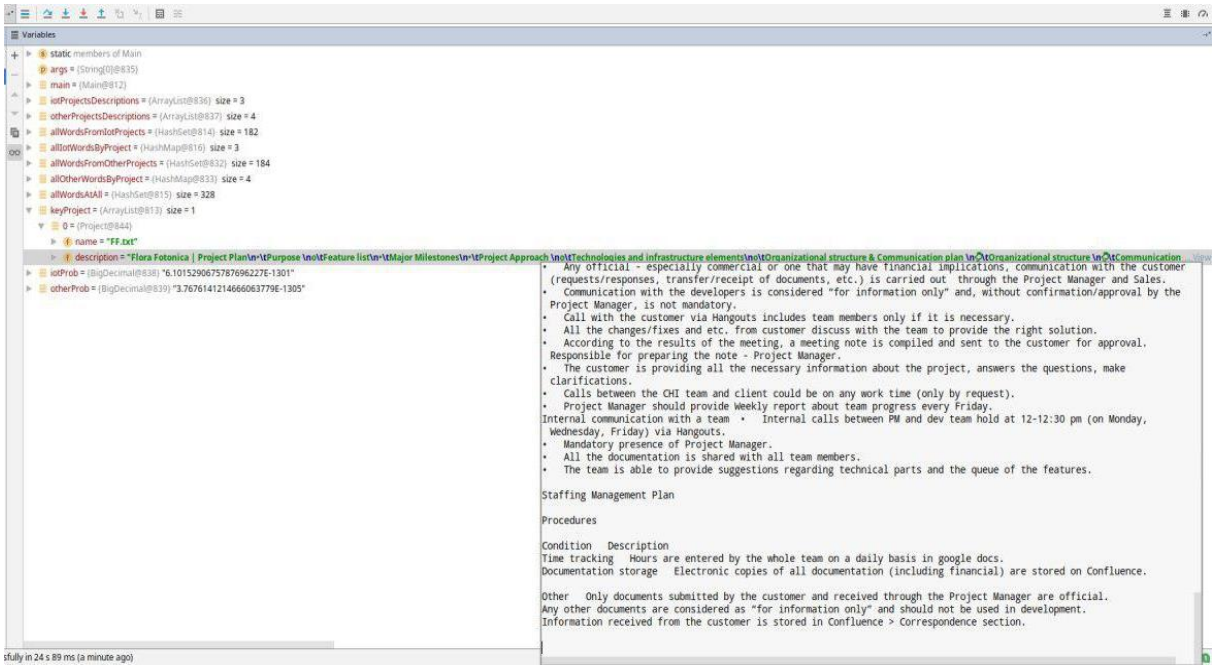

Figure 4.7 – Test project plan

But in spite of all this, there is one caveat, a case when there is no specific word in our training set. This means that $P(w_1|IoT) = 0$. This is rather inconvenient, since we are going to multiply it by other probabilities, so in the end we get zero. This is equal to 0, since when multiplying, if one of the members is equal to zero, the entire calculation is reset. Such actions simply do not give us

any information, so we must find a way around this. To do this, you need to use Laplace smoothing: we add 1 to each account so that it never vanishes. To balance this, we add the number of possible words to the divider so that the division never exceeds 1.

Now we just multiply all the probabilities, and see which one is bigger. The probabilities for each class, assuming an arbitrarily small probability for words that are not in our training set, are equal. So figure show us two probabilities marked with green color (Figure 4.8).

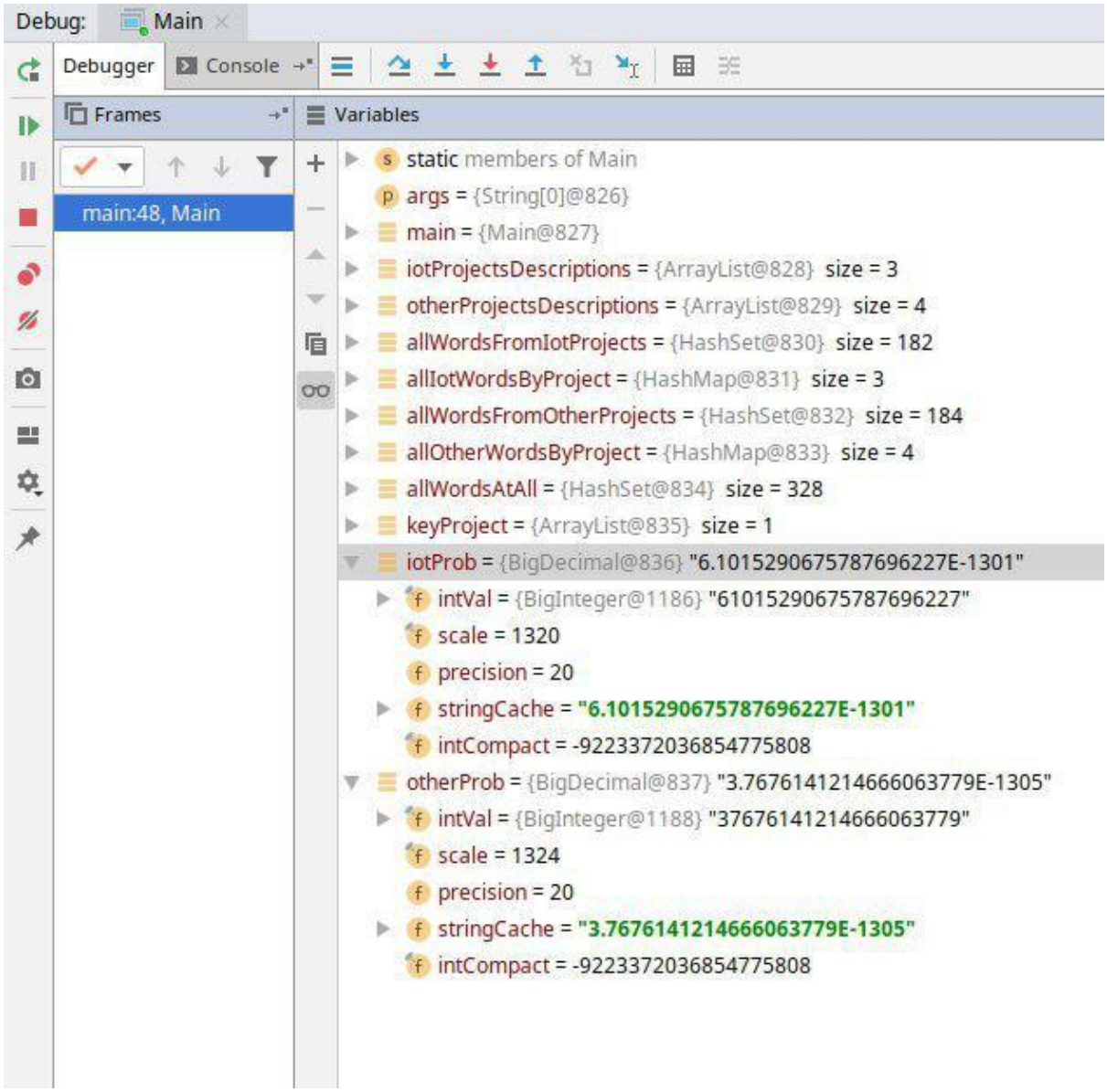

Figure 4.8 – Result of the classification algorithm

We present this result in a more comfortable form and get the numbers presented below (Figure 4.9). Summing up classifier gives «Smart Project» project the IoT tag and it is correct. Detailed information about source code could be orverviewed in Attachment E.

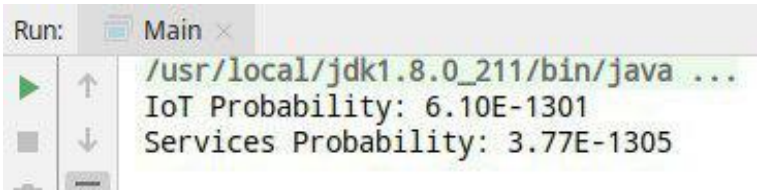

Figure 4.9–Final result of the Multinomial Naive Bayes

4.4 Future research opportunities

In the course of work on the diploma project, an algorithm was developed to classify the project documentation, which should become a convenient tool for its daily use, experts who do not have the opportunity to independently develop such a system using Text Mining. The program gives the user the ability to easily define the business domain of the project and get ready solutions. This approach to the analysis of project documentation using classification can have a large number of development options since there are still many unsolved problems.

Intermediate forms of text analysis with varying degrees of complexity can be used for various data mining purposes. To gain new knowledge in a particular domain area, it is necessary to perform semantic analysis. This will help to get a sufficiently broad view to capture the relationship between the objects or concepts described in the documents. However, semantic analysis methods are computationally expensive and often work in the order of several words per second. It remains only to decide how to speed up the algorithm and make semantic analysis much more efficient and scalable for very large textual cases.

While data mining in most cases does not depend on the language, text mining uses a significant part of it. Therefore, to improve the recognition of text,

there is a variant of developing an additional algorithm for refining text based on NLP, which will update the multilingual text of documents and then complement the existing intermediate forms. Thus, mining from documents, and not just canning and reading, allows you to access previously unused information and offers new opportunities.

Knowledge of the subject area, which is not taken into account by any modern tools for text analysis, can play an important role in text mining. In particular, knowledge of the subject area can be used already at the stage of refinement of the text. It’s interesting to learn how domain information can be used to improve parsing, efficiency and get a more compact intermediate form. This problem can be solved using Clustering. Thus, saving time on the formation of training sets to recognize new sets for indefinite classes, it is possible to apply one of the clustering algorithms.

Knowledge of the subject area can also play a role in the dissemination of knowledge, namely, the initialization of the structure of knowledge and the creation of previously discovered.

Another option for the development of this system is to establish links between elements of an intermediate form. For example, the search for the dependencies of the selected methodology, project scope and type of contract, or the relationship between the domain area of the project and specific software solutions. In other words, in the short term, this algorithm can be developed in different areas and can be used as a large system.

CONCLUSION

During the pre-certification practice, a visual area was considered. Based on this goal, the objectives of the study were clearly defined. I have examined in detail the Data mining area, which is related to identifying patterns in text data that are often very important (such as business critical ones). Unlike numeric data, text is often amorphous and difficult to deal with. The extraction of texts usually consists of multiple analysis of the text of the document by removing key phrases and concepts, text preparation.

Text analysis is an interdisciplinary field that relies on information retrieval, data mining, machine learning, statistics, and computer envistics. As most of the information is currently stored as text, text mining is considered to be of high commercial value. Growing interest is being given to multilingual data mining: the ability to receive information by language and to cluster similar objects from different linguistic sources according to their meaning. The extraction of useful data is the extraction of features, that is, the identification of terms and concepts most commonly used in input documents. The second goal is usually to identify any associations between the features (for example, linkages between the subject areas according to the development methodology). So, the first step to developing word processing algorithms is usually to "encode" the input information. As a second step–getting new information–you can use different methods, such as Algorithms for Association Rules, Classifications, and Clustering.

Having studied the modern methods of storing information online, I came to the conclusion that the Text Mining industry is a convenient way to solve many problems. In addition, the main analogues of the developed algorithm were identified and their comparative analysis was performed. In any case, regardless of its scope, the analysis of textual data is becoming more and more relevant every day. The Bayes algorithm solves the problem of classifying project documents into different domain domains. This classification helps the experts to automate

their activities and thereby enables them to concentrate on solving more complex issues. In addition, this classification can be used on any project documents and help to find similar previous experience.

Thus, summarizing the data obtained during the pre-certification practice, based on the current level of development of communication tools, we can conclude that the prospect of using methods and algorithms Text Mining for large corporations that store information in text form in any field, as well as in domestic practice to solve other human problems.

REFERENCES

1. Hulko D. Real world of data mining techniques //. Science, research, development №22 : materials of scientific-practical. conf. from the international. participation (London, 31th of Ocober 2019). Kharkiv, 2019. p. 45–48.

2. Text Mining. Predictive Methods for Analyzing Unstructured Information / Sholom M.Weiss, Nitin Indurkhya, Tong Zhang, Fred J. Damerau., 2005. p. 243.

3. Ronen Feldman, James Sanger. The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data. 2006. p. 423.

4. Min Song, Yi-fang Brook Wu. Handbook of Research on Text and Web Mining Technologie. New York, Hershey, 2009, p.189.

5. Michael W. Berry, Jacab Kogan. Text Mining. Application and Theory. 2010. p. 232.

6. Ashok Shivastava, Mehran Sahami. Text Mining. Classification, Clustering and Applications. 2009. p. 308.

7. Anne Kao, Stephen R.Poteet. Natural Language Processing. 2007. p. 272.

8. Natural Language Processing with R. Steven Bird, Ewan Klein, Edward Loper [and others]. 2009. p. 504.

9. Stylianos Hatzipanagos, Steven Warburton. Handbook of Research on Social Software and Developing Community Ontologies. New York, Hershey, 2009. p. 633.

10. S.Staab, R. Studer. Handbook on Ontologies. 2009. p. 832.

11. Ontologies. A Handbook of Principles, Concepts and Applications in Information Systems / Raj Sharman, Rajiy Kishore, Ram Ramesh [and others]. 2007. p. 929.

12. Використання технології контент-моніторингу для вирішення задач онтологічного інжинірингу / Н.В.Рябова, М.В. Білоіваненко, Ю.В. Сидоренко // Восточно-Европейский журнал передовых технологий. 2012.

№ 5. С. 53– 66.

13. Witold P. Knowladge-Based Clustering. Hoboken, Wiley-Interscince. 2005. p. 188-190.

14. Brian Henderson-Sellers. On the Mathematics of Modeling, Metamodeling, Ontologies and Modeling Languages. Sydney, Springer. 2012. p. 577.

15. Giovanni Sartor, Pompeu Casanovas, Maria Angela Biassiotti. Approaches to Legal Ontologies. Therories, Domain, Methodalogies. London, Springer. 2010. p. 645.

16. Gilbert Paquette. Visual Knowladge Modeling for Semantic Web Technologies:Models and Ontologies. New York, Harshey. 2011. p. 358.

17. J.D. Melier. Getting Results the Agile Way. A Personal Results System for Work and Life. Bellevue, Innovation Playhouse. 2010. p. 102.

18. David Allen. Getting Things Done the art of stress-free productivity. New York, Penguin Books. 2015. p. 345.

19. Frederick P. Brooks. The Mythical Man-Month: Essays on Software Engineering. Boston, Addison-Wesley. 2010. p. 885.

20. Roy Osherove. Notes to a Software Team Leader: Growing Self Organizing Teams. New York, Team Agile Publishing. 2013. p. 66.