

Vinnitsia National Technical University

(full name of a higher educational establishment)

Faculty of Computer Systems and Automation

(full name of institute, name of faculty (department))

Department of Control Computer Systems

(full name of the department (subject, cycle commission))

Explanatory note

to the master's qualification work
on the topic « Information system of intellectual analysis of natural-
language content of social networks. Part 3. The subsystem of
determination of sentimental characteristics of participants of social
networks on the basis of linguistic analysis of microblogging»

Fulfilled: 2nd year student of
the group 3ACIT-18m
specialty
151 Automation and computer-integrated

technologies

(cipher and name of the field of training,
specialty)

educational-professional program
Information systems and Internet
of things

(name of the educational-professional program)

Yaroslav Yu. Kotyk

(surname and initials)

Supervisor Oleg V. Bisikalo

(surname and initials)

Вінницький національний технічний університет
(повне найменування вищого навчального закладу)

Факультет комп'ютерних систем та автоматики
Кафедра автоматизації та інтелектуальних інформаційних технологій
Освітньо-кваліфікаційний рівень магістр
Спеціальність 151 Автоматизація та комп'ютерно-інтегровані технології
(шифр і назва)
Освітньо-професійна програма Інформаційні системи і Інтернет речей

ЗАТВЕРДЖУЮ
Завідувач кафедри
АПТ,
д.т.н., проф. Р. Н. Кветний

“” _____ 2019 року

**ЗАВДАННЯ
НА МАГІСТЕРСЬКУ КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ**

Котику Ярославу Юрійовичу
(прізвище, ім'я, по батькові)

1. Тема магістерської кваліфікаційної роботи «Інформаційна система інтелектуального аналізу природно-мовного контенту соціальних мереж. Частина 3. Підсистема визначення емоційних характеристик учасників соціальних мереж на основі лінгвістичного аналізу мікроблогів.»

Керівник магістерської кваліфікаційної роботи Бісикало Олег Володимирович, д.т.н., проф. АПТ
каф.

(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом вищого навчального закладу від “__” _____ 2019 року №__

2. Строк подання студентом магістерської кваліфікаційної роботи _____ року.

3. Вихідні дані до магістерської кваліфікаційної роботи: підтримка ОС: Windows, Linux та MAC; використання засобів для розпізнавання мови – так; препроцесинг текстового файлу – так; обробка природної мови – так; отримання результатів у вигляді структурованого файлу з часовими штампами для слів – так; мова графічного інтерфейсу – англійська.

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити): вступ, аналіз архітектурних рішень систем-аналогів, аналіз засобів розпізнавання емоційної забарвленості мови, розробка архітектури системи, розробка програмного забезпечення системи, тестування програмного забезпечення, висновки.

5. Перелік графічних матеріалів (з точним зазначенням обов'язкових креслень): алгоритм роботи додатку (обов'язковий), UML-діаграма варіантів використання (обов'язковий), UML-діаграма діяльності системи (обов'язковий), UML-діаграма класів (обов'язковий), екранна форма бібліотеки додатку, екранна форма режиму читання додатку.

6. Консультанти розділів магістерської кваліфікаційної роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		Завдання видав	Завдання прийняв
1-4	Бісікало О.В., проф. каф. АІТ		

7. Дата видачі завдання _____ 2019 року

Календарний план

№ з/п	Назва етапів дипломної роботи	Строк виконання етапів роботи	Примітка
1	Варіантний аналіз архітектурних рішень систем-аналогів		
2	Аналіз засобів визначення емоційного забарвлення тексту		
3	Розробка архітектури системи		
4	Розробка програмного забезпечення системи		
5	Розробка UML-діаграм системи		
6	Тестування програмного забезпечення		
7	Оформлення пояснювальної записки, графічного матеріалу і презентації		
8	Попередній захист		
9	Захист МКР		

Студент _____ Котик Я.Ю .
(підпис) (прізвище та ініціали)

Керівник магістерської кваліфікаційної роботи _____ Бісікало О.В.
(підпис) (прізвище та ініціали)

Vinnitsia National Technical University
(full name of a higher educational establishment)
Faculty of Computer Systems and Automation
Department of Automation and Intelligent Information Technology
Education level master's degree
Specialty 151 Automation and computer-integrated technologies
(cipher and name)
Educational-professional program Information systems and Internet of things

APPROVED

Head of the AIIS department

Dr. of Sci., professor Roman N. Kvetny

“ ” 2019 year

THE TASK
ON THE MASTER'S QUALIFICATION WORK STUDENT

Yaroslav Yu. Kotyk .

(fullname)

1. Theme of master's qualification work «Information system of intellectual analysis of natural-language content of social networks. Part 3. The subsystem of determination of sentimental characteristics of participants of social networks on the basis of linguistic analysis of microblogging»

Supervisor of master's qualification work Oleg V. Bisikalo ,

(full name, degree, academic rank)

approved by the order of the higher educational institution from № _____

2. The term of submission by the student of master's qualification work _____.

3. Output data for master's qualification work: OS support: Windows, Linux and MAC; the use of language recognition tools - yes; text file preprocessing - yes; natural language processing - yes; getting results as a structured time stamp file for words - yes; GUI language is English.

4. Contents of the settlement and explanatory note (list of issues to be developed): introduction, analysis of architectural solutions of analog systems, analysis of sentimental characteristics recognition tools, development of system architecture, development of system software, testing of software, conclusion.

5. List of graphic material (with exact indication of mandatory drawings): application algorithm (required), UML-diagram of use cases (required), UML-diagram of activity (required), UML-diagram of classes (required), screen view of the application library, screen view of the application reading mode

6. Consultants of sections of master's qualification work

Section	Full name and position of consultant	Signature, date	
		The task was issued	The task was accepted
1-4	Dr. of Sci., professor Oleg V. Bisikalo		

7. Date of issue of the assignment _____.

Calendar schedule

№	The name of the stages of work	The duration of the stages of work	Note
1	Variant analysis of architectural solutions of analog systems		
2	Analysis of language recognition tools		
3	Development of system architecture		
4	System software development		
5	Development of UML diagrams of the system		
6	Development of audio and text synchronization method		
7	Software testing		
8	Explanatory note, graphic materials and presentation design		
9	Previous protection		
10	Protection of MQW		

Student _____ Yaroslav Yu. Kotyk
(signature)(Full name)

Supervisor of master's qualification work _____ Oleg V. Bisikalo
(signature) (Full name)

CONTENTS

INTRODUCTION.....	8
1 ANALYSIS ARTIFICIAL INTELLIGENCE FROM THE POINT OF PROBLEMS COMPUTATIONAL LINGUISTICS	10
1.1 Analysis of approaches to the understanding of artificial intelligence... 10	
1.1.1 Turing Test and intuitive approach.....	10
1.1.2 Character Approach	12
1.1.3 Logical Approach.....	13
1.1.4 Agent-oriented and hybrid approach	13
1.2 Machine Intelligence.....	14
1.3 Modern methods of machine learning	15
1.4 Features of AI given by the task of Computational Linguistics	19
1.5 Relevance of the statistical analysis of big text data	21
1.6 Methods for analyzing large data.....	22
2 OVERVIEW AND RATIONALE TECHNOLOGICAL MEANS TO SOLVE PROBLEMS.....	24
2.1 Features establishing neural networks	24
2.2 Comparison of frameworks to support the machine learning.....	26
2.2.1 Theano.....	26
2.2.2 Caffe.....	27
2.2.3 Tensorflow	27
2.2.4 Torch	27
2.2.5 Microsoft Cognitive Toolkit	28
2.3 Overview Package DKPro Core	28

	7
2.4 Overview of resource WordNet.....	29
2.5 Word2Vec	29
2.6 Natural Language Toolkit, NLTK	30
3 DEVELOPMENT OF THE SOFTWARE	31
3.1. Key Data Features	31
3.2. Text Features.....	31
3.3 Selection of classification algorithm.....	33
3.3.1 Method of supporting vectors	33
3.3.2 Naive Bayes Classifier.....	35
3.4 Definition of key words in the text microblogging	36
3.4.1 Features approach to identify key words in the text.....	41
3.4.2 Algorithm for the keywords based on the proposed method.....	42
3.5 Input data.....	43
3.6 Analysis of the results	47
4 ECONOMIC ANALYSIS OF DEVELOPMENT.....	51
CONCLUSIONS.....	55
REFERENCES.....	56
ATTACHMENTS.....	61
ATTACHMENT A	62
ATTACHMENT B	65

INTRODUCTION

Relevance. The relevance of this work is associated with the growth of the Internet and the continuous increase in the amount of data in it. In this regard, it became necessary to automatically process textual information and determine the semantic characteristics of the text. Since existing methods and technologies do not achieve the required quality, new methods are needed to solve these kinds of problems. Examples of such tasks are the task of defining the keywords of the text and the task of constructing a lexical ontology of natural-language text. Effectively resolving these tasks allows you to deduce the overall theme and meaning of the text, which can later be used in areas such as SEO to determine the most relevant search results for users.

The purpose and objectives of the study. The purpose of the work is to solve the problem of computational linguistics on the basis of ready-made libraries, algorithms and tools of the NLTK linguistic software package, in particular in carrying out sentimentanalysis of natural-language texts on the basis of communication between the members of the sentence.

In order to achieve the above goals, you need to solve a number of the following tasks:

- a. analysis of areas of artificial intelligence in terms of the tasks of computational linguistics;
- b. overview and justification of the choice of technological means for solving the problem;
- c. development and testing of sentimenttext analysis software.

*Object of study.*The object of this study is the processes of linguistic analysis of English texts.

*Subject of study.*The subject of this study is the methods and means of solving the problem of sentimentanalysis of English texts using the NLTK software package.

*Practical implementation.*The implementation of the system will allow to solve the problem of sentimentanalysis of English-language texts based on the NLTK software package.

Research methods. In the process of research were applied: analysis, analogy, modeling, classification and experiment.

Novelty. To improve existing methods of sentiment analysis of natural-language texts on the basis of communication between the members of the sentence.

Approbation of the results. The research results were presented at the XLVIII Scientific and Technical Conference of the faculty, staff and students of VNTU (March 2018)[43].

1 ANALYSIS ARTIFICIAL INTELLIGENCE FROM THE POINT OF PROBLEMS COMPUTATIONAL LINGUISTICS

1.1 Analysis of approaches to the understanding of artificial intelligence

The only answer to the question what the artificial intelligence (AI) does not exist. Almost every author who wrote a book on artificial intelligence, it is repelled by any definition, considering his relation to achieving the science. The philosophy of unresolved questions about the nature and status of the human intellect. There is no exact criteria for achieving computers "reasonableness", although at the dawn of artificial intelligence has been proposed a number of hypotheses, such as the Turing test or hypothesis Newell-Simon [2]. Therefore, despite the existence of a set of approaches to understanding the problems of AI and the creation of intelligent information systems, there are two main approaches to the development of artificial intelligence:

- drop down (Top-Down AI) or semiotic - is the creation of expert systems, knowledge bases and inference systems that mimic human high-level mental processes, including thinking, reasoning, language, emotions, creativity;

- upward (Bottom-Up AI) or biology - the study of neural networks, evolutionary computation, modeling intelligent behavior based on biological elements, and creating appropriate systems such as neurocomputer or bio-computer [1,2].

The latter approach, strictly speaking, does not belong to the science of artificial intelligence in the sense given by John McCarthy, but they share common ultimate goal. The tasks of computational linguistics are the first semiotic approach.

1.1.1 Turing Test and intuitive approach

Empirical test was proposed by Alan Turing in his article "Computing machinery and intelligence» (Computing Machinery and Intelligence) published in 1950 in the

philosophical journal «Mind». The purpose of this test is to determine the possibility of artificial thinking close to human.

The standard interpretation of this test is as follows: "Man interacts with one computer and one (other) person. Based on the answers to the questions should determine first person to whom he speaks, with a person or computer program. The task of a computer program to introduce people into error, forcing make the wrong choice. " All test participants will not see each other.

The most common approach assumes that AI will be able to detect behavior (Figure 1.1), which is different from the human, and in normal situations. This idea is a generalization of the Turing test approach, which claims that the car will be reasonable when will be able to maintain a conversation with an ordinary man, and he can not understand that tells the machine (the conversation is by correspondence).

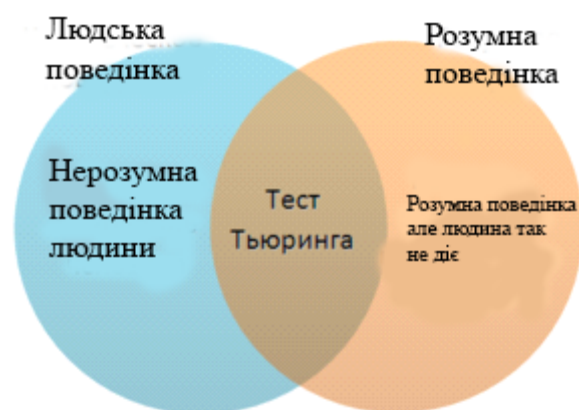


Figure 1.1–The principle of the Turing test

Fiction writers often suggest another approach: AI occurs when the machine will be able to feel and to create. Thus, the owner of the Andrew Martin "bicentennial man" begins to treat him as a man when he creates his own toy project. A Deyta of "Star Trek" by being able to communicate and training, wants to find emotion and intuition.

However, the latter approach is unlikely to hold water on closer inspection. For example, it is easy to create a mechanism that will evaluate some parameters of external or internal environment and to respond to their unfavorable value. On this system we can say that she is feeling ("pain" - a reaction to the shock sensor operation, "hunger" - a response to low battery power). A cluster created Kohonen maps, and many other products of "smart" systems can be roughly seen as a kind of creativity. [2]

It should be noted that the first time the emergence and the main criterion to date intelligence "Turing Test" of course refers to the problems of computational linguistics.

1.1.2 Character Approach

Historically, the approach was the first character in the era of digital machines, as it is after the creation of Lisp, the first language of symbolic computation in the author emerged virtually confidence in the ability to start implementing these means intelligence. Character approach allows to operate weakly-formalized concepts and their meanings. The ability to select only the relevant information depends on the efficiency and effectiveness of the assignment.

But the breadth of classes of problems that are solved effectively the human mind requires incredible flexibility in the methods of abstraction. This is not available in any engineering approach, in which the researcher chooses a solution methods based on the ability to quickly give effective solution to some nearby researcher this task. That is already under rules implemented as a single model abstraction and design entities. This results in significant cost resources for non-core tasks, ie the system of intelligence back to brute force in most tasks, but the essence of intelligence disappears from the project.

The main application of symbolic logic - is the solution of problems with making rules. Most studies on the impossibility stop it even mark new challenges that arise unexpectedly researcher, means chosen in previous phases of character, especially to solve them, especially to teach a computer to solve them or at least identified and out of these situations .

In computational linguistics problems symbol approach also showed its limitations at the effective application of context-independent grammars Chomsky classification, for example, to create a compiler compilers and programming languages. Once it comes to natural language semantics which is much broader than formal basic syntax of language modeling processes by means of context-sensitive grammars causes considerable difficulties [4].

1.1.3 Logical Approach

The logical approach to the creation of artificial intelligence aimed at creating expert systems with logical model of knowledge bases and using language predicates (most often - 2nd order).

Teaching model of artificial intelligence in the 1980s were adopted Language and logic programming Prolog system. Knowledge Base recorded in the language Prolog represent sets of facts and rules of inference, written in the language of predicate logic.

The logical model of knowledge bases can record not only the specific information and data in the form of facts language Prolog, but generalized information through the rules and procedures of inference and including logic rules definitions that express certain knowledge - both specific and generalized details and abstract concepts.

In general, the study of artificial intelligence as part of a logical approach to the design of knowledge bases and expert systems aimed at the creation, development and operation of intelligent information systems, including on training students and pupils, as well as training users and developers of intelligent information systems [3,13] .

Directly related to the problems of computational linguistics logical approach through ontologies, including natural language that is widely used for the development of "smart" version of the worldwide network WEB 3.0.

1.1.4 Agent-oriented and hybrid approach

The latter approach, developed since the early 1990s, called agent-based approach or an approach that is based on the use of intelligent (rational) agents. Under this approach, intelligence - a computing part (roughly speaking, planning) the ability to achieve our goals of intellectual machine. Most such vehicle is an intelligent agent that perceives the outside world through sensors, it can affect the objects in the environment (Figure 1.2) using actuators.

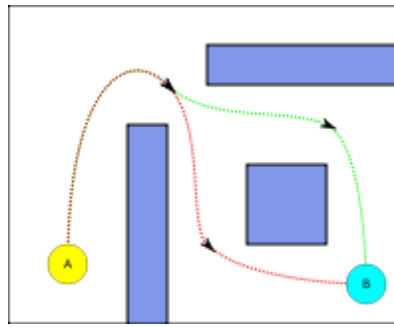


Figure 1.2–The principle of the agent-based approach

This approach focuses on the methods and algorithms, intelligent agents to help survive in the environment in the performance of its tasks. Yes, there is much more carefully studied the path search algorithms and decision making.

Hybrid approach assumes that only the synergistic combination of neural, and character models up to the full range of cognitive and computational capabilities. For example, expert opinions mental rules can be generated neural networks and generating rules obtained by statistical learning. Proponents of this approach believe that hybrid information systems will be much stronger than the sum of different concepts separately.

1.2 Machine Intelligence

From a historical point of view the term "intellectual" man describes itself. The presence of intelligence man differs from animals and plants. Today the word "intelligence" used to describe one of the distinguishing features of people; If someone called "highly intellectual" that by this means the person is particularly developed in mental terms.

From a conceptual point of view, no doubt, there is a universal form of intelligence. And the people and animals have a small share of this ability and show a concrete manifestation of the universal intelligence. Most likely, people have inherited most of this versatile abilities. But due to the fact that the basis of human intelligence is a biological nature, we lack some features of the universal intelligence (eg, completeness and neutrality).

Most researchers in computer science believe that biological intelligence is available to play, that intelligence is not exclusively human attribute. With this assumption essentially follows that of universal intelligence and can move the car. Thus, an artificial intelligence - is artificial intelligence counterpart produced by the human brain as a biological substance [5]. Obviously, the development of means results in different results than biological evolution, and this is the explanation for the fact that artificial intelligence is characterized by different properties compared to human intelligence (such perfection in all). Thus, the artificial intelligence is another manifestation of the universal intelligence.

It is not easy to define such abstract concepts as universal intelligence, especially in view of the fact that the number of truly distinct examples of its manifestation small. But you can try to make the definition of human intelligence.

In the absence of a better definition, we assume that intelligence - a set of skills that enable people to solve problems with limited resources. These skills include learning ability, abstract thinking, planning skills, imagination and creativity. They constitute the most important aspects of human intelligence.

Considering how wide palette of skills that make up intelligence, we can formulate a problem whose solution can be used as a test of intelligence. Thus, elements of intelligence even find animals: they can, for example, survive alone in the wild and manage your time. The colonies of insects are also able to adapt quickly to changes in habitat to protect their nests. Even popular IQ tests measure, rather, not so much intelligence as the level of fitness and ability to make such tests. In other words, these tests measure intelligence is only in its narrow sense [5].

1.3 Modern methods of machine learning

Machine learning has become quite common in our society in recent years. It has been used successfully in problems of search, computer vision, medical, unmanned drones and car control, as well as problems of computational linguistics. The core of many of these applications are methods and artificial intelligence, such as classification, localization and detection. Recent developments in the field of neural

networks (so-called "deep learning") significantly promoted the performance of visual recognition systems. Consider the existing methods of deep machine learning, neural networks architecture with a focus on the study models to solve these problems, including the problem of classification.

The most famous version of the algorithm neural network training - the so-called back propagation algorithm. There are advanced second-order algorithms such as conjugate gradient method and Levenberg-Marquardt method that many problems are significantly faster (sometimes on the order). Back propagation algorithm is the most easy to understand, and in some cases it has certain advantages. Also developed heuristic versions of the algorithm that work well for certain classes of problems - rapid expansion (Fahlman) and Delta-Delta with a yield (Jacobs) [5].

In back-propagation algorithm calculated gradient vector surface errors. This vector indicates the direction of the shortest descent to the surface from this point, so if we "just" moved on it, the error is reduced. The sequence of steps (slows down as it approaches the extremum) eventually lead to a minimum of a different type. During use raises several questions:

- a) you need to take long steps;
- b) how to deal with getting into local extremum point;
- c) provide a possible conversion network.

At great length step convergence will be faster, but there is a danger jump over the decision, or (if the surface of errors is particularly bizarre) go in the wrong direction. A classic example of this phenomenon in learning neural network is when the algorithm moves slowly along a narrow ravine with steep slopes, jumping from one side to another it (Fig. 2.1). Conversely, if a small step is likely to be caught right direction, but it required a lot of iterations.

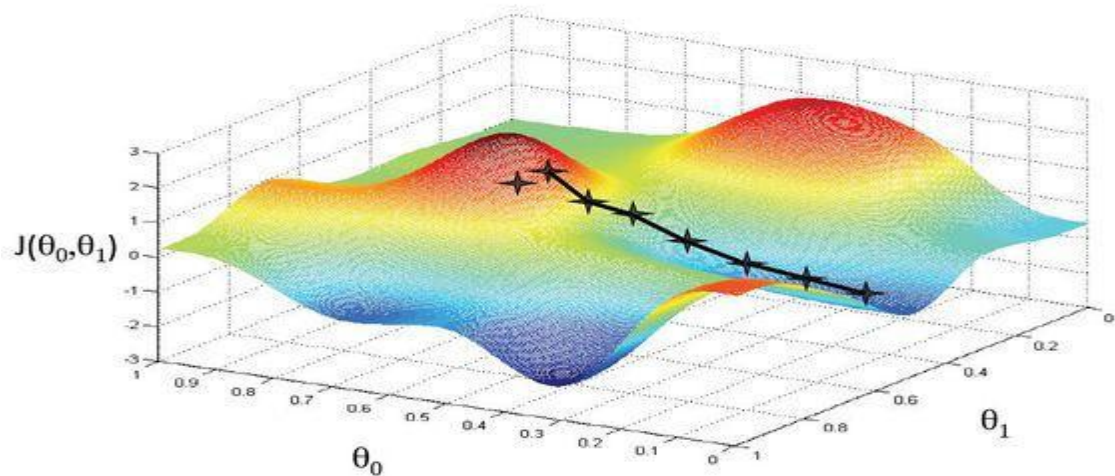


Figure 1.3 - Demonstration of the method of gradient descent

In practice, the value of the proportional step taken slopes (the algorithm slows the progress of near minimum) with some constant called speed training. The right choice depends on the speed of training and specific task usually done empirically; This constant can also depend on the time decreases as we move algorithm.

Usually this algorithm is modified so as to include the term pulse (or inertia). This member promotes a fixed direction, as if there were several steps in the same direction, then the algorithm "increase speed" that (sometimes) to avoid local minimum, and quickly pass flat areas.

Thus, the algorithm works iteratively, and his steps are called epochs. Each era of the input network in turn serves all educational observation network output values are compared with target values and calculated error. The value of the error and error surface gradient is used to adjust the weights, then all actions are repeated. Initial network configuration is chosen at random, and the learning process is terminated or when passed a number of periods or when some error reaches a certain level of smallness or if an error stops decreasing (the user can select a necessary condition for stopping) [5].

One of the most serious problems of the above approach is that the way we minimize the wrong error that actually have minimized - an error that you can expect from the network when it will be given a completely new observation. In other words, we would like neural network was able to generalize results to new observations. In

fact, the network is trained to minimize the error on the training set, because in the absence of perfect and infinite training set is not the same thing as to minimize the "real" error on the surface of errors in previously unknown effects model (Bishop).

The strongest difference is in this issue of retraining or too close fitting. This phenomenon is easier to demonstrate the non-neural network as an example of approximation using polynomials - while the essence of the phenomenon is absolutely the same.

Network weights simulating a large number of more advanced features and therefore prone to conversion. The network is with a small number of weights may be flexible enough to model the existing relationship. For example, a network without intermediate layers actually simulates the normal linear function.

How to choose the "right" level of complexity to the network? In most cases, more complex network gives smaller error, but it may not indicate good quality of the model, but the conversion.

The answer is to use a control mechanism for cross checking. We reserve the observation of teaching and not use them in training the algorithm back-propagation. Instead, in the course of the algorithm used for independent control results. At the beginning of the training network error and the control data set will be the same (if they differ significantly, it is likely that all observations splitting into two sets of data were heterogeneous). As soon as the network is trained, error learning naturally decreases, and while learning function reduces the actual error, error in the control set will also decrease. If the control error is no longer reduced or even began to grow, indicating that the network started too closely approximated by data and studies should be stopped. This phenomenon is very accurate approximation in training and retraining called. If this happens, it is usually advised to reduce the number of hidden items and / or layers, because the network is too powerful for the task. If the network rather taken enough rich to simulate existing relationship, then re likely will not happen, and both errors - training and checking - not achieved a sufficient level of smallness [5].

This problem of local minima and choosing the size of the network led to the fact that the practical work with neural networks usually have to experiment with a large number of different networks, sometimes teaching each of them several times (not to be

misled by local minima) and comparing the results. The main outcome measure of quality control there is an error. Thus, in accordance to the general scientific principle of "Occam's Razor", according to which ceteris characteristics should be preferred over simple model of two chains of about equal error control makes sense to choose one that is smaller.

The need for repeated experiments leads to the fact that the control sample begins to play a key role in choosing a model that becomes part of the learning process. Thus weakened its role as an independent criterion as a model - with a large number of experiments is a risk of selecting "successful" network that gives good results in the control sample. In order to provide the final model due reliability, often (at least when the amount of training data allows it) if it is so, another reserve - set of test cases. The final model is tested on data from the plurality to make sure that the results achieved in the training and control sets real and not artifacts of the learning process. Of course, in order to play its role well, the test set to be used only once:

1.4 Features of AI given by the task of Computational Linguistics

According to the second interpretation, Artificial Intelligence - is a set of technologies and techniques. The best definition of the term "artificial intelligence" in this interpretation an artificial intelligence - a scientific field that studies methods for solving using machines challenges, such methods are used to solve these problems people [4]. Usually, these methods are reduced to modeling biological characteristics of intelligence and the application of the models in the form of algorithms, playable on computers.

Artificial intelligence algorithms can be used to solve almost any application, not just for modeling human intelligence. For example, they can be used to control production line in industry or for pattern recognition in medicine. These techniques are part of the scientific field devoted to the study of artificial intelligence problems, due to the presence of the general characteristics of between biological and artificial intelligence (eg, learning ability or abstraction).

Artificial intelligence as a scientific discipline, is located at the intersection of many disciplines (eg, computer science, psychology and mathematics). Each, in turn, relies on a huge reservoir of human knowledge environment in nature, including ourselves. Given such a large number of issues related to modeling artificial intelligence is very difficult to say what relates to the field of scientific interests scientists involved in this topic and what is not. Research Interests specialists in artificial intelligence is constantly expanding, interfering in other areas of human knowledge; it shows the maturity of the scientific field and its compliance with scientific apparatus other natural science theories.

Historically, scientists dealing with artificial intelligence usually focus on fairly narrow problems and methods tied to a particular subject area. Such specialization makes the task of learning (as well as practical implementation) possible solutions relatively simple. These highly specialized solutions are called weak artificial intelligence, because they are very difficult to apply anywhere outside problem areas for which they are designed. [4]

This weakness artificial intelligence gradually become something of a boulder lying on the road that can neither circumvent or bypass. In many areas of artificial intelligence methods proved its best, but at the same time for their correct application in most cases requires a human expert. When attempts are made to collect these methods together to solve larger problems, it is clear that these methods are very specialized.

This is why high school need to train engineers specialized in artificial intelligence. If the theoretical foundations of artificial intelligence were perfect, the programmer does not be necessary. But before that point is at least several decades, and we are still at the development of technical systems need in people - experts in engineering.

Given the problem of computational linguisticsAI manifest weakness generally in the absence of universal programs to support natural language dialogue. Turing Test, as the results of the relevant annual program of free competition, yet inaccessible to the general subject area of dialogue. This known examples of information retrieval systems or systems support dialogue (bots) for narrow areas (such as ticketing or drugs) that demonstrate effective use of the results.

Outlines the problem of poor intelligence in Computational linguistics you can narrow down to a level ontologies - much easier to design and debug ontological model narrow domain of knowledge, rather than a universal ontology of the world. This ontology is unique for each person - a natural carrier of intelligence, and created and perfected in close connection with the development of language skills throughout the life of this man.

1.5 Relevance of the statistical analysis of big text data

Large data is intended for forecasting. Usually they are described as part of computer science called "artificial intelligence" (more precisely, its section "Machine learning"). The application of mathematical techniques to a large number of data for probability forecasting, for example, such that an email is spam, is considered. that instead of the word "koip" was supposed to be dialed "Copy"; that the trajectory and speed of the person passing the road in the unauthorized place, say that he will have time to pass the street in time and the car needs only a little lower speed. But most importantly, these systems work efficiently due to the large amount of data that they can build their predictions on. Moreover, the systems are designed in such a way to improve over time by tracking the most useful signals and models as new data arrives. As the defining characteristics for large data, note "three V": volume (English Volume, in terms of physical volume), velocity (English Velocity in the sense of both the rate of growth, and the need for high-speed processing and obtaining results), diversity (English Variety, in the sense of the possibility of simultaneous processing of various types of structured and semi-structured data)[1].

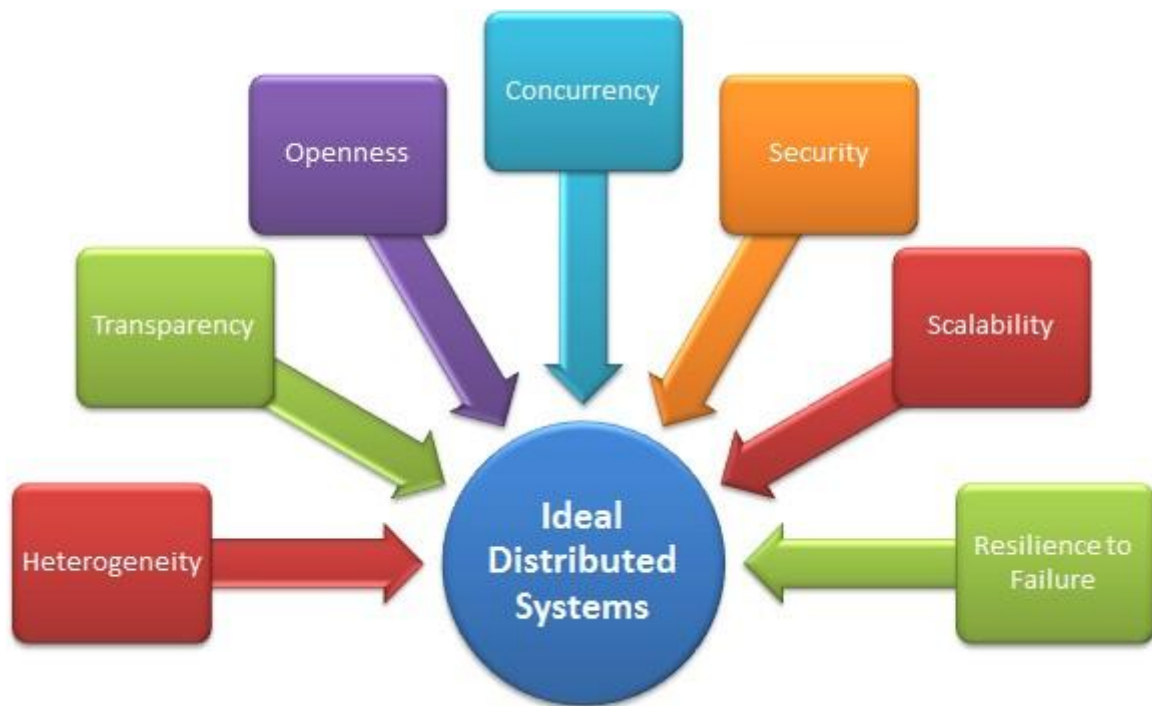


Figure 1.4 Ideal distributed system

1.6 Methods for analyzing large data

There are many different techniques for analyzing data arrays, which are based on a toolkit borrowed from statistics and computer science (for example, machine learning). Here are some of them:

- methods of the Data Mining class: Association rule learning, classification (methods for categorizing new data based on principles previously applied to existing data), cluster analysis, regression analysis;

- krautsorsing - categorization and enrichment of data by the forces of a broad, uncertain range of persons involved on the basis of a public offer, without entering into a labor relationship;

- data fusion and integration - a set of techniques that allow the integration of heterogeneous data from a variety of sources for the possibility of in-depth analysis; examples of such techniques, which are part of this class of methods, are digital signal processing and natural language processing (including tonal analysis);

- machine learning, including teaching with a teacher and without a teacher, and Ensemble learning - using models built on the basis of statistical analysis or machine learning for comprehensive forecasts based on basic models;
- artificial neural networks, network analysis, optimization, including genetic algorithms;
- forecasting analyst;
- simulation modeling;
- spatial analysis - a class of methods using topological, geometric and geographic information in the data;
- statistical analysis, as examples of methods given A / B testing and analysis of time series;
- visualization of analytical data - presentation of information in the form of drawings, diagrams, using interactive features and animation both for obtaining results, and for use as raw data for further analysis.

2 OVERVIEW AND RATIONALE TECHNOLOGICAL MEANS TO SOLVE PROBLEMS

Solving the problem requires a study of the use of modern technological tools of machine learning and computational linguistics. Consider the basic software systems and tools that potentially can be applied to the analysis of natural language texts.

2.1 Features establishing neural networks

The correct choice of network volume is important. Build a small and high-quality model is often impossible, and a large model will simply remember examples from the training set and produce an approximation, which naturally lead to incorrect operation of the classifier. There are two basic approaches to building networks - constructive and destructive. At first, the first network set minimum and then gradually increase it to achieve the required accuracy. At each step of re-teach. There is also the so-called cascade correlation, in which occurs after age adjustment network architecture to minimize errors. When destructive approach originally taken overvalued network volume, and then remove it from the nodes and links that have little impact on the decision. It is useful to remember the following rule: the number of examples in the training set must be greater than the number of elements of weight matrix. Otherwise, instead of summarizing data network simply remember and lose the ability to classification - the result is indefinite for example, are not included in the study sample. [8]

When choosing a network architecture usually tested several configurations with different number of elements. This key indicator is the amount of training set and generalization ability of the network. Commonly used algorithm of Back Propagation with supporting a plurality (validation set).

Fully-connected NN. Neural networks are modeled as sets of neurons connected as acyclic graph. In other words, some outputs neurons can become the inputs of other neurons. Cycles are not allowed, as this would mean an infinite loop in the forward

pass network. Instead amorphous clusters connected neurons, neural network models are often combined in some layers of neurons (Figure 2.2). For conventional neural networks, the most common type layer is fully-connected layer in which neurons between two adjacent layers fully connected in pairs, but neurons within one layer does not have relations.

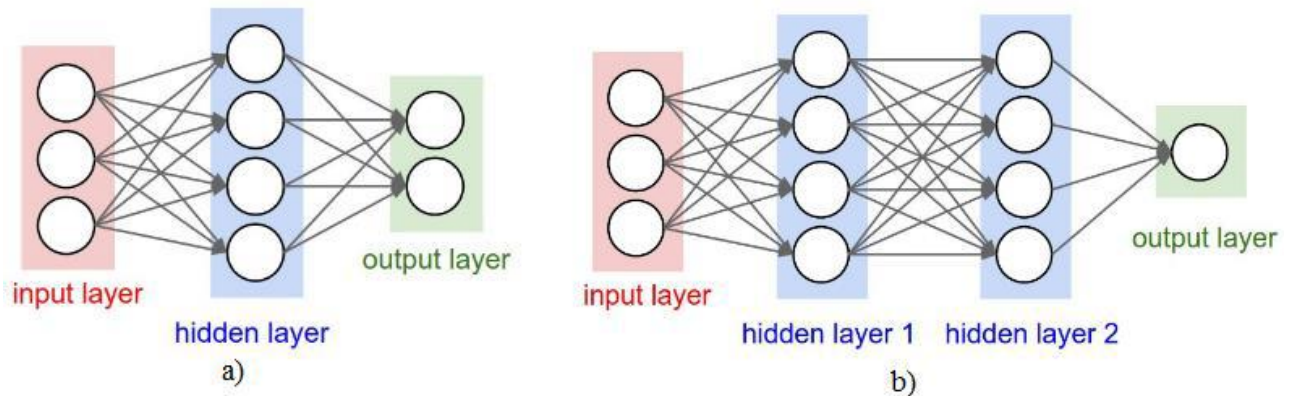


Figure 2.2 - 2-layer (a) and layer 3 (b) fully-connected NN

Above are two examples of neural network topology using stacked completely linked (fully-connected) layers:

a) 2-layer neural network (one hidden layer with 4 neurons and one output layer of 2 neurons), and three inputs;

b) 3-layer neural network with three inputs, two hidden layers of neurons 4 each and one output layer. Note that both cases have connections (synapses) between neurons between layers, but not within the layer.

2.2 Comparison of frameworks to support the machine learning

An important aspect in solving problems of machine learning is the choice of tools for easy work with neural networks. The table in Figure 2.3 presents a comparative analysis of popular frameworks that support work with machine learning.

Software	Open source	Platform	Interface	CUDA support	Has pretrained models	Recurrent nets	Convolutional nets	RBM/DBNs	Parallel execution (multi node)
Caffe	Yes	Linux, Mac OS X, Windows	Python, MATLAB	Yes	Yes	Yes	Yes	No	No
Microsoft Cognitive Toolkit	Yes	Windows, Linux (OSX via Docker)	Python, C++, Command line, BrainScript	Yes	Yes	Yes	Yes	No	Yes
TensorFlow	Yes	Linux, Mac OS X, Windows	Python, C/C++, Java, Go	Yes	Yes	Yes	Yes	Yes	Yes
Theano	Yes	Cross-platform	Python	Yes	Through Lasagne's models	Yes	Yes	Yes	Yes
Torch	Yes	Linux, Mac OS X, Windows, Android, iOS	Lua, LuaJIT, C, utility library for C++/OpenCL	Yes	Yes	Yes	Yes	Yes	Yes

Figure 2.3– Comparison of frameworks for machine learning

2.2.1 Theano

Theano is one of the most famous veterans and stable machine learning libraries. Getting deep learning libraries are the subject of dispute between Theano and Caffe. Theano - is a low-level library that follows Tensorflow style. It is used not only for deep learning how to optimize the numerical calculations such as automatic calculation of the gradient function, which together with Python interface and its integration with Numpy, made the library one of the most frequently used for general purposes in the field of Deep Learning. Today, it still exists and is used but the fact that it does not support multi-GPU computing gives reason to think about other choice framework.

2.2.2 Caffe

Caffe along with Theano is one of the first frameworks. It focuses only on the computer dawn, but performs its functions perfectly. Experimentally determined that the preparation of architectural CaffeNet won 5 times less than CaffeKeras (using backend Theano). Disadvantages that Theano is not flexible. If you must enter the new changes, you need to program in C ++ and CUDA, although it is possible to use interfaces Python or Matlab.

The disadvantages of the framework is the lack of good documentation and installation complexity, because it has many dependencies. However, as a tool for developers of industrial solutions, computer vision system Caffe is the undisputed leader. Since the theme of this work is not related to computer vision Caffe not suitable for solving the problems.

2.2.3 Tensorflow

TensorFlow defined as library software with open source Machine Intelligence, but a more precise definition is: TensorFlow is an open source software library for numerical calculation using data flow diagrams. This framework does not include himself in the frame Deep Learning as well as Teano. Tensorflow supports Python and C ++, allows distributed CPU and GPU computing and horizontal scaling using gRPC. As a result: Tensorflow the best frameworks for distributed work on machine learning and its functions for constructing graphs with experimental data help debug model, trained in a Tensorflow.

2.2.4 Torch

Torch Framework supports multiple ways (stacked layers or strata graph) to identify the network, but is essentially defined as a network graph layers. Because of

this level of detail Torch sometimes considered less flexible because new types of users layers must implement new network feedback and input gradient renovation.

However, unlike Caffe, define a new layer in the Torch is much easier, because you do not need to program in C++. In addition, the Torch difference between the new definition and the definition of minimum network. In the Caffe layers defined in C++ while a defined network through Protobuf.

Torch is more flexible than TensorFlow/Theano and that is extremely important, while TF / Theano declarative (that is, they need to initialize the computer graph). It simplifies some operations (for example, process flow) in the Torch compared to other frameworks.

2.2.5 Microsoft Cognitive Toolkit

CNTK as a system of deep learning better known in the community of NLP, than the total depth study of the community. In CNTK (as in TensorFlow and Teano), a network is defined as a symbolic Count vector operations such as matrix addition / multiplication or convolution. Layer - is a composition of these operations. Fine grain building blocks (operations) allows users to invent new types of complex layers without their realization in the language of low level (as in Caffe).

How to use CNTK like Caffe - you need to specify the configuration file and run the command line. CNTK has support for Python since version 2.0 and C# support in the process. Like Caffe, CNTK also a cross-platform C++ solution. Thus, the deployment should be easy in most cases. However, it does not work on ARM architecture, which limits its capabilities on mobile devices.

2.3 Overview Package DKPro Core

DKPro Core - a set of software components for natural language processing based on the Apache UIMA framework. Linguistic DKPro Core package was built to improve the productivity of researchers working with automatic language analysis [31].

DKPro Core approach is that researchers should be able to concentrate on their real scientific matters, not on technology development. Having developed a significant number of libraries with known models and algorithms in computational linguistics linguistic make DKPro Core package is the most convenient tool for solving research problems.

2.4 Overview of resource WordNet

WordNet Linguistic resource developed at Princeton University (USA). WordNet is classified as tools for lexical ontologies freely available online. On this basis were made thousands of experiments in the field of information retrieval. WordNet 2.1 includes approximately 155,000 different tokens and phrases organized into 117 thousand; the total number of couples "synset-value" has 200,000 [30]. Thesaurus Development was launched in 1984. In 1995, WordNet appeared on the Internet in the public domain and caused a surge of research into its use in various computer applications automatic text processing. The results of WordNet were not as clearly positive, but WordNet opened a new era of development of very large scale structured language resources

Originally established as WordNet model of human memory. Many decisions descriptions understanding words in WordNet motivated psycholinguistic experiments. However, it should be noted that WordNet caused a much greater interest in computer linguists than psycholinguistics. It is also important that research in an open environment WordNet actively continue and now - it leads to new and improvement of existing components of the popular linguistic resource.

2.5 Word2Vec

The technology is based on the representation of words in the form of vectors a given dimension, placing similar words close to each other. The distance between the vectors of words denoting similar things, for example, "Cat" and "dog" will be significantly less than between words, meanings which have little in common, for

example, “cat” and “plane”. This feature allows more flexible presentation of data that may be further used in the training of neural networks, various classifiers, etc. To create a word-vector correspondence base, the algorithm first scans the entire text issued to him, compiling a “dictionary”, which in subsequent iterations of the algorithm will be used to determine corresponding vectors. There are two main approaches: CBOW (Continuous Bag of Words) and Skip-gram. CBOW - “continuous bag with words” model architecture that predicts the current word based on from the surrounding context. Skip-gram type architecture works differently: she uses the current word to predict the words surrounding him. [6]

2.6 Natural Language Toolkit, NLTK

This free library for the Python programming language is one of the best for creating various software products on this language. It provides a large set of tools, text bodies, has provided wrappers for using other libraries within themselves. For example, to analyze the tonality of the text and markup sentences, there is the ability to connect the above Stanford CoreNLP product. Also For various classifications, the NLTK provided an interface for connections of classifiers from another library - Scikit learn, about which we will go further. And more information about use, device this can be found on the official website [13].

3 DEVELOPMENT OF THE SOFTWARE

3.1. Key Data Features

Microblogs are, first of all, services to simplify publication and perception user data. Typically, microblog posts consist of one or a pair sentences, and for Twitter there is a strict limit on the length of the tweet - 140 characters. IN 140 characters platform users need to fit the context, their attitude to subject and, possibly, a link to a photo, Internet resource or other media object. Often the context is restored from the outside world, that is, the user writes that worries the Internet at this moment, and people, owning this information, compare the statement with real events. It's so easy for a computer to keep abreast of topics discussed it turns out, so you can't count on restoring the context.

Microblogging platforms are also social networks where users can interact with each other. On Twitter, for example, other than social graphs, you can observe the graphs in which the messages themselves are built: users can respond to tweets, as well as host other users tweets, in platform terminology this is called "retweeting". Social Interaction Information can refine the classification results, for example, there is an intuitive assumption, that the answer to a negatively colored message will also fall into the negative class.

3.2. Text Features

To express emotions in the text, users put emoticons. A smile is a set of characters that conditionally illustrates the expression on the face of the author, or rather his mood. All smiles can be divided into eastern and western by geography of their use, the latter are shown in table 4 with labels corresponding to theirsentimentalcoloring. In case of in short texts there is no easier way to mark your attitude to the topic than to put a smile, but not all users do this, so markup messages with their help in the general case will not work. There are more complex constructions

of brackets, colons and other symbols, but they are not used often and usually mean more than just attitude, and some actions or objects, that is, emotional coloring do not carry.

smile	mark	smile	mark	smile	mark	smile	mark	smile	mark
:~)	+	:~)	+	:o)	+	:~]	+	:3	+
:c)	+	:>	+	=~]	+	8)	+	=)	+
:~]	+	:^)	+	:>]	+	:~D	+	:D	+
8-D	+	8D	+	x-D	+	xD	+	X-D	+
XD	+	=-D	+	=D	+	=3	+	=3	+
B^D	+	:~))	+	:>[-	:~(-	:(-
:~c	-	:c	-	:<	-	:>C	-	:<	-
:~[-	:[-	:{	-	;(-	:	-
:~@	-	:>(-	:~(-	:~(-	:~)	+
:~)	+	D:<	-	D:	-	D8	-	D;	-
D=	-	DX	-	v.v	-	D-:	-	:*	+
:~^*	+	(+	}}}	+)	+	:~)	+
:~)	+	*~)	+	*)	+	:~]	+	:~]	+
:~D	+	:^)	+	:~,	+	:>P	+	:~P	+
:~P	+	X-P	+	x-p	+	xp	+	XP	+
:~p	+	:p	+	=p	+	:~b	+	:b	+
:~b	+	:~b	+	:~b	+	:b	+	d:	+
:>~\	-	:>/	-	:~/	-	:~.	-	:/	-
:~\	-	=/	-	=\	-	:~L	-	=L	-
:~S	-	:><	-	:~	-	:~	-	:~\$	-
O:~)	+	O:~3	+	O:3	+	O:~)	+	O:~)	+
O:~^)	+	O_O	-	ø/	+	<3	+	<3	-

Figure 3.1–Emotional coloring of smiles

In addition to ASCII smiles, there are also graphic ones - these are pictures that are inserted into text. Modern web services and mobile applications use graphic Emoji language for recording words, emotions and actions. Figure 3.2 shows some well-known graphic emoticons that are used on the Facebook social network. Usually for each of them there is an ASCII analogue, and not one. When typing a message on the keyboard of a computer or laptop, it's more convenient to put a colon with a bracket, but smartphones and tablets provide all the amenities for inserting smiling pictures: along with Russian and an English keyboard, for example, you can connect a graphic keyboard to them Emoji language.

Since emoticons are a kind of markup of messages by users themselves, it should be used in the analysis of emotional coloring. In this work, we will consider the use of symbolic and graphic smiles for collecting the body of tweets and for data preprocessing immediately before classification.



Figure 3.2–Some of graphic emoticons that are used on the Facebook

3.3 Selection of classification algorithm

In a number of studies to determine the polarity of the text, high teaching methods with a teacher showed effectiveness. These methods used as in early work to determine the polarity of a document on average, and in modern works, where the proposals are analyzed and short text messages.

To solve this problem, 2 algorithms were chosen, proven to be most effective in determining polarity short text messages [10,12].

3.3.1 Method of supporting vectors

The support vector method belongs to the family of linear classifiers. The purpose of linear classification is to search hyperplanes in the space of signs dividing all objects into two class.

The basic idea of the support vector method is to search separating hyperplane as far as possible from the nearest to it points in feature space. In the case of a linearly separable sample, the search for a hyperplane can be write as optimization task:

$$\begin{aligned} \frac{1}{2} \|\omega\|^2 &\rightarrow \min_{\omega, b} \\ y_i(\omega^T x_i + b) &\geq 1, j = 1, \dots, m \end{aligned} \quad (1)$$

Where $\frac{1}{\|\omega\|}$ - the gap between the hyperplane and the nearest to it points of both the first and second class. $y_i(\omega^T x_i + b)$ - product values of the class of the point and its position relative to the hyperplane.

For the more general case of a linearly inseparable sample, the algorithm may make mistakes at training facilities. New optimization challenge includes the requirement to minimize error:

$$\begin{aligned} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^l e_i &\rightarrow \min_{\omega, b} \\ y_i(\omega^T x_i + b) &\geq 1 - e_i, i = 1, \dots, k \\ e_i &\geq 0, \quad i = 1, \dots, k \end{aligned} \quad (2)$$

Variables e_i characterize the magnitude of the error in the sample of k elements. Constant C allows you to find a compromise between the maximizing of the gap and minimizing the total error by training sample.

The realization of the Method of supporting vectors shown at the Figure 3.3:

```

from sklearn import linear_model
df = pd.read_csv('linear_regression_df.csv')
df.columns = ['X', 'Y']
df.head()
sns.set_context("notebook", font_scale=1.1)
sns.set_style("ticks")
sns.lmplot('X', 'Y', data=df)
plt.ylabel('Response')
plt.xlabel('Explanatory')
linear = linear_model.LinearRegression()
trainX = np.asarray(df.X[20:len(df.X)]).reshape(-1, 1)
trainY = np.asarray(df.Y[20:len(df.Y)]).reshape(-1, 1)
testX = np.asarray(df.X[:20]).reshape(-1, 1)
testY = np.asarray(df.Y[:20]).reshape(-1, 1)
linear.fit(trainX, trainY)
linear.score(trainX, trainY)
print('Coefficient: \n', linear.coef_)
print('Intercept: \n', linear.intercept_)
print('R2 Value: \n', linear.score(trainX, trainY))
predicted = linear.predict(testX)

```

Figure 3.3–The realization of supporting vectors method

3.3.2 Naive Bayes Classifier

Naive Bayes Classifier – Probabilistic classifier based on Bayes theorem and the (naive) assumption of statistical independence of random variables (3).

$$p(C | F_1, \dots, F_2) = \frac{p(C)p(F_1, \dots, F_2 | C)}{p(F_1, \dots, F_2)} \quad (3)$$

The main advantage of this classifier is the low computational complexity, as well as optimality, provided true independence of signs.

The realization of Naïve Bayes Classifier shown at the Figure 3.4:

```

from __future__ import division
from collections import defaultdict
from math import log

def train(samples):
    classes, freq = defaultdict(lambda:0), defaultdict(lambda:0)
    for feats, label in samples:
        classes[label] += 1           # count classes frequencies
        for feat in feats:
            freq[label, feat] += 1   # count features frequencies

    for label, feat in freq:         # normalize features frequencies
        freq[label, feat] /= classes[label]
    for c in classes:               # normalize classes frequencies
        classes[c] /= len(samples)

    return classes, freq           # return P(C) and P(O|C)

def classify(classifier, feats):
    classes, prob = classifier
    return min(classes.keys(),      # calculate argmin(-log(C/O))
               key = lambda cl: -log(classes[cl]) + \
               sum(-log(prob.get((cl,feat), 10**(-7)))) for feat in feats))

```

Figure 3.4–The realization of Naïve Bayes Classifier

You can see the UML diagram of the Naïve Bayes algorithm at the Attachment B

3.4 Definition of key words in the text microblogging

In general, the approach to identify key words, given the linkages between members of the proposal consists of stages:

1. create multi-text markup;
2. use markup syntax that takes into account the complex relationships between pairs lems;
3. verbal noise reduction;
4. choice of the first n words with the most connections, where n - the number of required keywords.

Creating a multi markup text and markup syntax that takes into account the complex relationships between pairs of Lemma invited to perform with sophisticated language packages.

Filtering verbal noise invited to provide with the following operations: replacing the relevant pronouns nouns them; noise removal relations; noise removal words; removal of stop words.

Consider the conflict in finding keywords, which is especially important for small text that can be small blogs or posts on social networks.

Today one of the most important and visible areas of Web, which is a key principle of user participation in the website are weblogs, or web logs, abbreviated name - blogs. Conceptual development blogs due to their broad socialization is microblogging, which have a number of distinctive features: limited length messages, high frequency of publication varied subjects, different ways of delivering messages, etc.

The first and most famous microblogging service Twitter was launched in October 2006 by Obvious San Francisco. Obviously, the automatic selection of the most significant terms (words) from the stream of messages generated by the community of Twitter, has practical importance for determining the interests of different groups of users, and to build a personal profile of each.

However, it should be noted that the classical statistical methods for extracting key terms based on analysis of collections of documents, are ineffective in this case.

This is an extremely small length messages (up to 140 characters), their various subjects and lack of logical connection between them, and a lot of seldom-used acronyms, abbreviations and specific elements of micro syntax.

Since Twitter does not exist simple and convenient way to group "tweets" on the subject of different users, user community enjoys its own solution: use hashtag.

They are similar to other examples of the use of tags (e.g. for annotating entries in blogs usual) and can add "tweets" in a certain category.

Hashtag symbol starting with "#" followed by any combination permitted in Twitter characters without spaces; this is often a word or phrase in which the first letter of each word capitalized provided.

They can occur in any part of the "tweets", often users simply add the symbol "#" in front of any word. When you add a hashtag message it will be displayed in the search in a stream of messages for Twitter hashtag it.

By unofficial but generally accepted rules of use hashtag refers choice for these terms are relevant to the subject, and add only a small number of them in one message. This allows us to consider them as terms that a sufficient degree of probability reflecting the overall theme of the message.

One of the tasks of text mining is a selection of key terms with some degree of reliability reflecting themes of the document. Automatically extract key terms can be defined as the automatic selection of important thematic terms in the document. It is one of the more common tasks subtasks - automatic generation of key terms for which the color key terms need not be present in this paper [36].

In recent years, there are many approaches that allow the analysis of sets of documents of different sizes and extract key terms consisting of one, two or more words were created.

The most important step is extracting key terms of the calculation of the weights in the sample document to assess their importance relative to each other in this context. To solve this problem, there are many approaches, which can be divided into 2 groups: require training and require no training. During training refers to the need for pre-processing the original corpus of texts to extract information about the frequency of occurrence of terms throughout the body. In other words, to determine the significance

of the term in this document should first analyze the entire collection of documents to which he belongs. An alternative approach is to use linguistic ontologies that are more or less approximate model of the existing set of words specified language. Based on both approaches were set up systems for automatic extraction of key terms

The best quality is achieved linguistic text processing methods or combinations at the statistics, because the system automatically determine the key phrases from the text natural language should be developed using morphological dictionary (vocabulary) and syntax.

Resultsofparsingnatural language with modern language packs allow you to program level available to operate the syntactic relationships between words separate proposal [41].

Applying the approach to the definition of keywords based on the use of additional information about the complex relationships between members of the English sentence, you find keywords in the text messages and microblogging compare them with the hashtag messages given by the author.

In computer science and cryptography hash collision - is equal hash values on two different blocks of data.

Competing in finding keywords - is equal frequency values for two or more candidates for keywords, and select keywords as are necessary only part. In most cases this problem is relevant for small texts such as abstract or records microblogging.

Consider an example - when determining keywords English text for text "Variability management in software product lines using adaptive object and reflection" [42] obtained and shown in Table 3.1 list of keywords sorted by number of links (frequency) descending.

Table 3.1 - Keywords and number of links to them

Word	Number of connections (frequency)	Word	Number of connections (frequency)	Word	Number of connections (frequency)	Word	Number of connections (frequency)
model	9	plan	4	challenge	2	savings	2
line	7	vehicle	4	creation	2	support	2
product	7	define	3	derive	2	variant	2
reuse	7	design	3	development	2	activity	1
variability	7	diagram	3	domain	2	adaptive	1
approach	6	extract	3	feature	2	architecture	1
base	6	identify	3	implement	2	brazilian	1
software	6	mechanism	3	key	2	finally	1
increase	5	offer	3	launcher	2	hypothetic	1
management	5	reflection	3	productivity	2	large-scale	1
process	5	step	3	propose	2	object	1
aim	4	study	3	quality	2	space	1
issue	4	benefit	2	satellite	2	specific	1

Conflict arises when you need to select from a list of potential keywords only first N words with the greatest frequency, which will be considered as keywords. For this example the text if you want 10 keywords, select the first eight easily. These will be the words: model, line, product, reuse, variability, approach, base, software. Then the last two words need to select between three words five with the same frequency: increase, management, process. So the small text solution of the conflict is an urgent task.

To reduce collisions can use the following approaches:

- Sort words with the same frequency by the frequency of their occurrence in a particular case. The relative importance of terms in the context of the sample is determined using data on the frequency of their use as a key in the online encyclopedia Wikipedia. The algorithm is based on calculating the "informative" of each term, like assessing the likelihood that he may be elected a key text [35]. This approach is quite accurate, but requires a preliminary analysis of the shell.
- Sort words with the same frequency by frequency of their occurrence in the frequency of word dictionary for the text. This approach generalizes word to word

forms, requires no pretreatment body of texts and easy to implement, but its accuracy is low.

- Check your keyword list connectivity, that even take into account the dependencies for different types of sentences [43]. Choose new key words that have a greater total number of connections with these words that formerly were in the keyword list. This approach does not require a housing and can be implemented by means of linguistic package.

- Choose first nouns, then verbs, and other parts of speech. As head of the sentence are usually nouns and verbs, then it will get out the words that may belong to the set key. Also, for the first noun can choose their own names, because one of the questions that must comply keywords: which names of organizations, people, and more geographical areas associated article [9].

Combining the last two approaches to reduce collisions can improve the results of the keyword for smaller sized texts.

It is proposed for a combined approach, first check the keyword with the same frequency on connectivity. In the second stage, if the remaining block potential keywords with the same frequency, selected first nouns, then verbs, and other parts of speech. Subsequent experimental studies have confirmed the feasibility of using this approach.

The combined approach for reducing collisions can be used as an optional module that improves the results of the keyword method for determining keywords English text based tools package DKPro Core, developed, and other algorithms to find key words.

Verbal noise or stop words - a term from the theory of information search by keyword. These are words that do not have meaning, because their benefit and the role of search is not significant [34].

In the processing carried exclusion from the studied words of the text, which by definition can not be so significant that constitute "noise". Unlike these key words are called neutral or Stop (stop words). These are words that belong to service parts of speech, and pronoun [3].

Reducing noise words can be achieved by using the approach: replacing the relevant pronouns nouns them; extract combinations of types of relationships that do not have significant semantic load; removal of words that relate to the often useless parts of speech; extracting words that refer to the list of stop words.

3.4.1 Features approach to identify key words in the text

The texts in any language consist of two parts that have to be expressed by the laws of the language, and that reflects the specific subject of the text and style of the author. These components are called the thematic neutral, label and thematic vocabulary. Marking both groups of vocabulary - a step towards determining the content relatedness of the text. This process allows both to approach the content of the text and make a definite opinion about the originality of the vocabulary of the author and his language.

When processing a text file raises a number of difficulties, because there are certain requirements to source code.

Most of these requirements are defined printing standards and user considerations:

- is that the text is built correctly (for example, two places can not go in a row);
- hyphen immediately after the word at the end of the line says to postpone words, after a row can not go any characters, including spaces;
- Dash is a single space after the symbol and the right of it must also be a space;
- Dash next the symbol next letter in the presence of the letter in the same line, said the phrase and never marked by spaces.

The adoption of these rules facilitates word processing, but makes it difficult to set on the computer, because it requires constant monitoring of material introduced [14].

One of the tasks of text mining is a selection of key terms, which to some extent reflect the themes of authenticity of the document. Automatically extract key terms can be defined as the automatic selection of important thematic terms in the document. It is

one of the more common tasks subtasks - automatic generation of key terms for which the color key terms need not be present herein.

In recent years, many approaches that allow the analysis of sets of documents of different sizes and extract key terms consisting of one, two or more words were created.

The most important step is extracting key terms of the calculation of the weights in the sample document to assess their importance relative to each other in this context.

The laws of statistics does not work on small documents (such as annotations) as they all frequencies singular and verbose key terms are approximately equal and strive for a single entry within the context of the document. But thanks pronouns replace nouns corresponding to the proposed method before using the laws of statistics may increase the frequency of potential keywords.

3.4.2 Algorithm for the keywords based on the proposed method

Algorithm for determining keywords:

- a) Create multi-text markup;
- b) markup syntax that takes into account the complex relationships between pairs lems;
- c) exclusion often useless to analyze the types of connections. Removed words that do not belong to separate parts of speech;
- d) replacing pronouns in pairs obtained in accordance with these nouns;
- e) receipt of pairs of keywords. At this step in the list of terms obtained in the previous step, the possible construction of a semantic graph. The semantic graph is a weighted graph whose vertices are the terms of the document, the presence of edges between two vertices means the fact that the terms are semantically related, rib weight is the numerical value of semantic proximity of two terms which connects this edge;
- f) couples splitting into separate words and determine the number of connections. This step increases the chances of getting the right keywords, as many of them are replaced in the following sentences, pronouns to avoid repetition;

g) n the first choice words with the most connections, where n - the number of required keywords.

3.5 Input data

As a input data for our experiment we used the database with 6 thousand tweets. You should create a Twitter Developer Account for getting Twitter API data about users and their posts, this is could be done on the official Twitter Developers site. Then you need to create an API application on Twitter (Twitter Developers). After that you must specify all the necessary parameters and APIs that will be used - in our case it is to get a list of users and their posts. We need to describe in detail the reasons why we need it and whether we will process data derived from their APIs. In our case it is processing posts of users for information about them, their comments. Once we have access to the Twitter API, we can retrieve information about users and their posts.

An example of a request for getting a list of users with the name:

GET <https://api.twitter.com/1.1/users/search.json?q=alex&page=1>

You also need to submit OAuth 1.0 authorization tokens when prompted. For this we send to Consumer Key, Consumer Secret Key, Access token, Access Token Secret.

As a result of the request, we get the following answer:

```

{
  "id": 7388362,
  "id_str": "7388362",
  "name": "alex (PVD)",
  "screen_name": "alex",
  "location": "San Francisco and Providence",
  "description": "Journalist. @Crunchbasenews. Co-host: Equity with @TechCrunch. Just married @lizaswagner. alex@crunchbase.com",
  "url": "https://t.co/lmg5XYqw3T",
  "entities": {
    "url": {
      "urls": [
        {
          "url": "https://t.co/lmg5XYqw3T",
          "expanded_url": "http://alexwilhelm.com",
          "display_url": "alexwilhelm.com",
          "indices": [
            0,
            23
          ]
        }
      ]
    },
    "description": {
      "urls": []
    }
  },
  "protected": false,
  "followers_count": 66155,
  "friends_count": 1492,
  "listed_count": 3553,
  "created_at": "Tue Jul 10 18:47:05 +0000 2007",
  "favourites_count": 116721,
  "utc_offset": null,
  "time_zone": null,
  "geo_enabled": true,
  "verified": true,
  "statuses_count": 43705,
  "lang": null,
  "status": {
    "created_at": "Tue Nov 26 15:38:00 +0000 2019",
    "id": 1199351604756938760,
    "id_str": "1199351604756938760",
    "text": "Reading this https://t.co/vA5IRU60Er",
    "truncated": false,
    "entities": {
      "hashtags": [],
      "symbols": [],
      "user_mentions": [],
      "urls": [
        {

```

Figure 3.5—The result of the request by name

Similarly, we get a list of user records by ID:

GET

https://api.twitter.com/1.1/statuses/user_timeline.json?user_id=134&page=1

The `user_id` parameter is the user ID you want to retrieve.

Page parameter - defines the page number of the data (on one page returns 20 entries). Similar to the previous request, authorization tokens should be sent.

As a result of the request, we get the following answer:

```

{
  "created_at": "Thu Nov 21 11:28:13 +0000 2019",
  "id": 1197476807672508417,
  "id_str": "1197476807672508417",
  "text": "coin-operated рой",
  "truncated": false,
  "entities": {
    "hashtags": [],
    "symbols": [],
    "user_mentions": [],
    "urls": []
  },
  "source": "<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android</a>",
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null,
  "in_reply_to_user_id_str": null,
  "in_reply_to_screen_name": null,
  "user": {
    "id": 1342892378,
    "id_str": "1342892378",
    "name": "jon slow",
    "screen_name": "dillinger59",
    "location": "dp",
    "description": "дзен-луддизм",
    "url": null,
    "entities": {
      "description": {
        "urls": []
      }
    },
    "protected": false,
    "followers_count": 41,
    "friends_count": 67,
    "listed_count": 0,
    "created_at": "Wed Apr 10 21:19:12 +0000 2013",
    "favourites_count": 779,
    "utc_offset": null,
    "time_zone": null,
    "geo_enabled": false,
    "verified": false,
    "statuses_count": 634,
    "lang": null,
    "contributors_enabled": false,
    "is_translator": false,
    "is_translation_enabled": false,
    "profile_background_color": "FFCC4D",
    "profile_background_image_url": "http://abs.twimg.com/images/themes/theme1/bg.png",
    "profile_background_image_url_https": "https://abs.twimg.com/images/themes/theme1/bg.png",
    "profile_background_tile": true,
    "profile_image_url": "http://pbs.twimg.com/profile_images/1170260880854081543/oeMwDTJ_normal.jpg",
    "profile_image_url_https": "https://pbs.twimg.com/profile_images/1170260880854081543/oeMwDTJ_normal.jpg",
  }
}

```

Figure 3.6—The result of the request by ID

These APIs have created and filled a database of 2000 users and 6,000 posts users. Posts were only taken into account with the text (post length should be greater than 0, images in posts are ignored).

The essence of "User" includes the following named characteristics: Numeric ID, ribbon ID, name, login, city, city description, or protected, number of subscribers, number of friends, number of user communities, date and time account creation, number of likes, time zone, or geolocation enabled, or verified account, number of posts, language.

The essence of "Post" includes the following named characteristics: Numerical ID, ribbon ID, date and time the post was created, text, the ID of the user who owns the post. So we filled the database for it analyze .

You can see system architecture diagram at the Figure 3.7

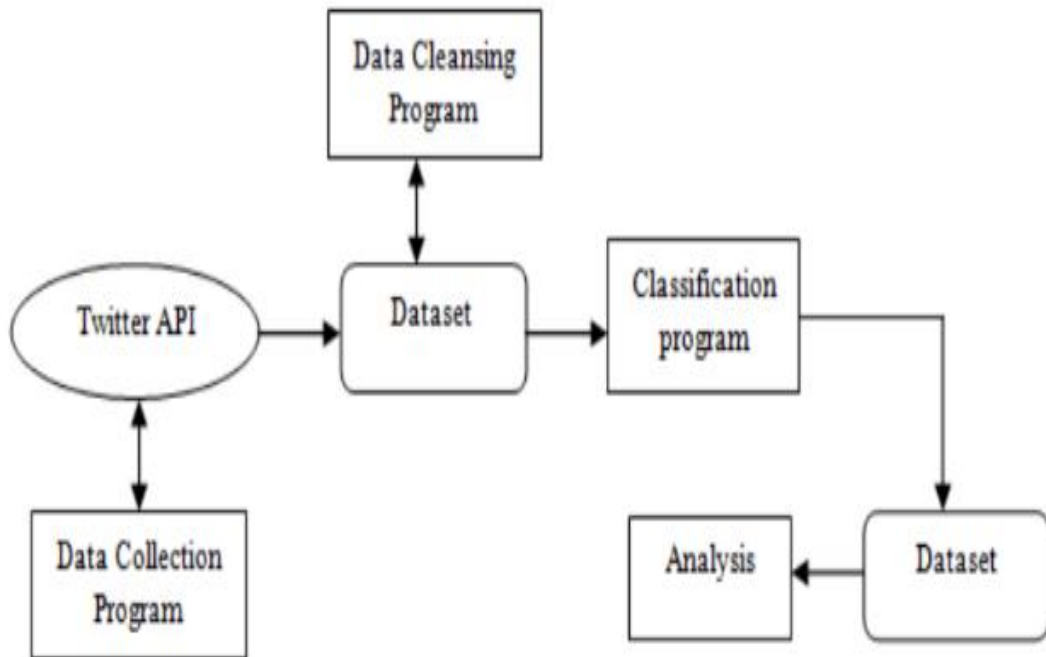


Figure 3.7–System architecture diagram

3.6 Analysis of the results

For this experiment database «mathling» and table «posts» were used:

The screenshot shows the phpMyAdmin interface for the 'mathling' database, specifically the 'posts' table. The table structure is as follows:

id	id_str	created_at	text	user_id
1000002552073736193	1000002552073736193	Fri May 25 13:16:02 +0000 2018	Heading into the long weekend like...#MemorialDay ...	38441438
1000003558937382912	1000003558937382912	Fri May 25 13:20:02 +0000 2018	RT @keisi_keisi: господи как же хочется простого ч...	3412622103
1000005316422008832	1000005316422008832	Fri May 25 13:27:01 +0000 2018	RT @ScolNational: It may seem like a repeat of a 2...	236786367
1000007621091512320	1000007621091512320	Fri May 25 13:36:10 +0000 2018	City with their unlimited spend can have their rec...	221289012
1000016663977316352	1000016663977316352	Fri May 25 14:12:06 +0000 2018	Mes sœurs d'engagement pour la cause de la santé d...	29499388
1000017864747823104	1000017864747823104	Fri May 25 14:16:53 +0000 2018	Good to meet the MG Environmental Trust today on s...	2245759171
1000019861916082183	1000019861916082183	Fri May 25 14:24:49 +0000 2018	RT @BuzzFeed: "You're damn right it's content" htt...	335534204
1000021066100756480	1000021066100756480	Fri May 25 14:29:36 +0000 2018	When you're negotiating who's turn it is to crutch...	393028587
1000022597445169152	1000022597445169152	Fri May 25 14:35:41 +0000 2018	RT @ASCAP: A moment between #ASCAP President @IMPa...	29572254
1000022947120037888	1000022947120037888	Fri May 25 14:37:04 +0000 2018	RT @cinematic_void: Tonight: HOCUS POCUS in 35mm a...	29572254
1000023067727351809	1000023067727351809	Fri May 25 14:37:33 +0000 2018	RT @JennyBicks: @CDanieChelliah @GreatestShowman ...	29572254
1000027987675402240	1000027987675402240	Fri May 25 14:57:06 +0000 2018	RT @natefreiman: There's been a lot written recent...	226769484
1000029081956765696	1000029081956765696	Fri May 25 15:01:27 +0000 2018	Freaking out on the @UAudio Eden Bass Amp plugin -...	29655016
1000029510941773826	1000029510941773826	Fri May 25 15:03:09 +0000 2018	JP day!!!!!! congrats lads! https://t.co/G5JoA1rX...	29655016
1000036297149440002	1000036297149440002	Fri May 25 15:30:07 +0000 2018	We're kicking off #gimletfest on June 16th at 12:3...	195738329
1000045202751479809	1000045202751479809	Fri May 25 16:05:30 +0000 2018	@MegPrik Love this!	252274830
1000045659213283329	1000045659213283329	Fri May 25 16:07:19 +0000 2018	@01FirstSecond @verabee Bought this for my girls, ...	252274830
1000046098990252032	1000046098990252032	Fri May 25 16:09:04 +0000 2018	@bensears Plus Man late night doodle from last nig...	252274830
1000049531151699971	1000049531151699971	Fri May 25 16:22:42 +0000 2018	Yess!!!!!! https://t.co/oubH4TnZ	29655016
1000049748882681600	1000049748882681600	Fri May 25 16:23:34 +0000 2018	SAT WORLDWIDE is excited to announce "We are growi...	120872742
1000056393129431040	1000056393129431040	Fri May 25 16:49:58 +0000 2018	Last night's Dialogue on women in politics was ama...	32056006
1000064570373877761	1000064570373877761	Fri May 25 17:22:28 +0000 2018	RT @DollSkinUK: HAPPY BIRTHDAY @punk_plant!! You'...	2460019718
1000064712834150144	1000064712834150144	Fri May 25 17:23:02 +0000 2018	@DollSkinUK Your car karaoke meme edits are legend...	2460019718
1000064986402697222	1000064986402697222	Fri May 25 17:24:07 +0000 2018	All I want for my birthday is for @MaxPortnoy... to ...	2460019718
1000065309120876544	1000065309120876544	Fri May 25 17:25:24 +0000 2018	Fan signing in Australia 🇺🇸 gotta love the... http...	85538853

Figure 3.8–Database «mathling»

Table posts using fields:

- id - unique identifier Tweet;
- text - text tweet.

As a result, it was determined keywords to 6 thousand tweets. Each output file has a name that has the prefix "output_without_stop_words_frequency[" , and then tweets id:

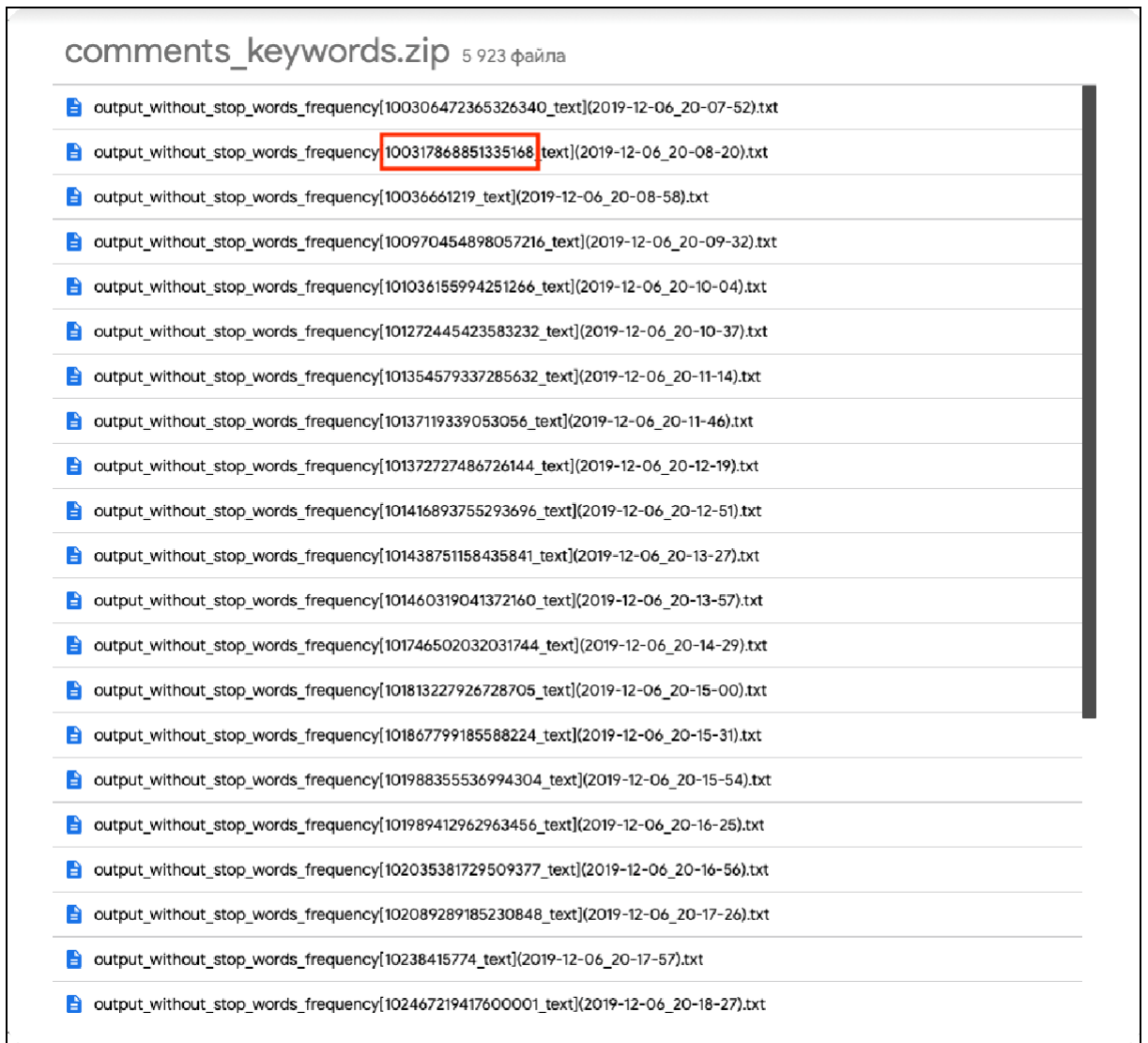


Figure 3.9—The folder with output files

Output file contains keywords and the number of connections for each word. The word that has the largest number of connections are in the first place:

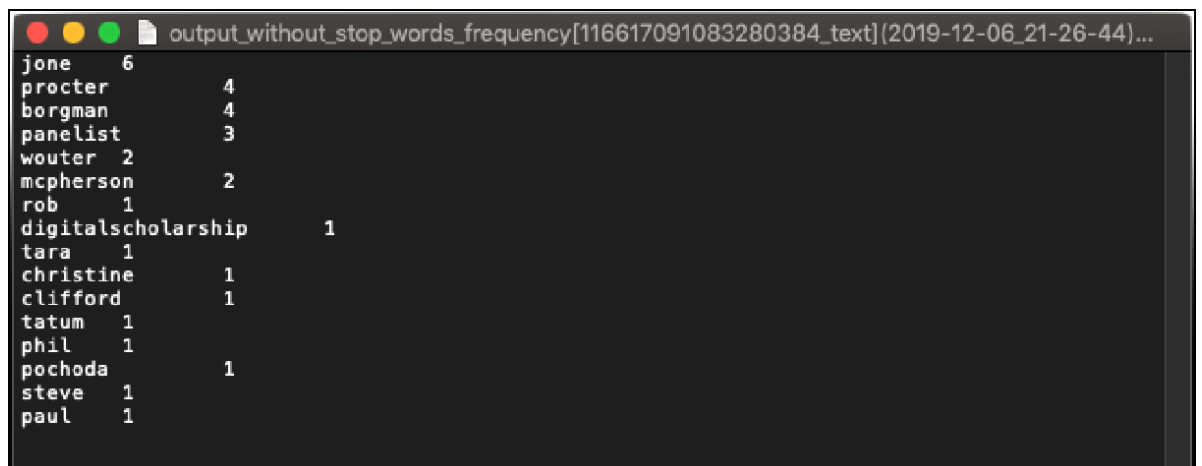


Figure 3.10—One of the output files

As a result of our work we built some graphs to show the distribution of common negative and positive words.

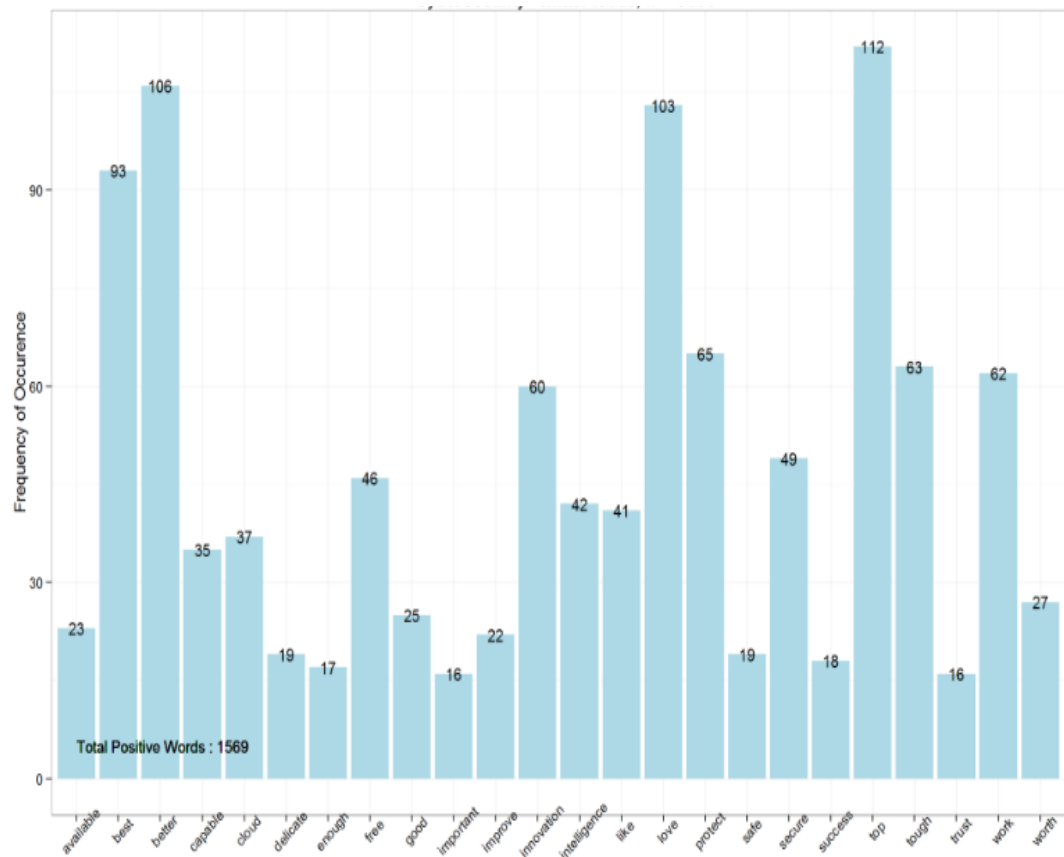


Figure 3.11–Major positive words

Similarly, we derive negative words and their frequency. There are 2063 negative words in 5,000 tweets .The result we can see at the Figure 3.12

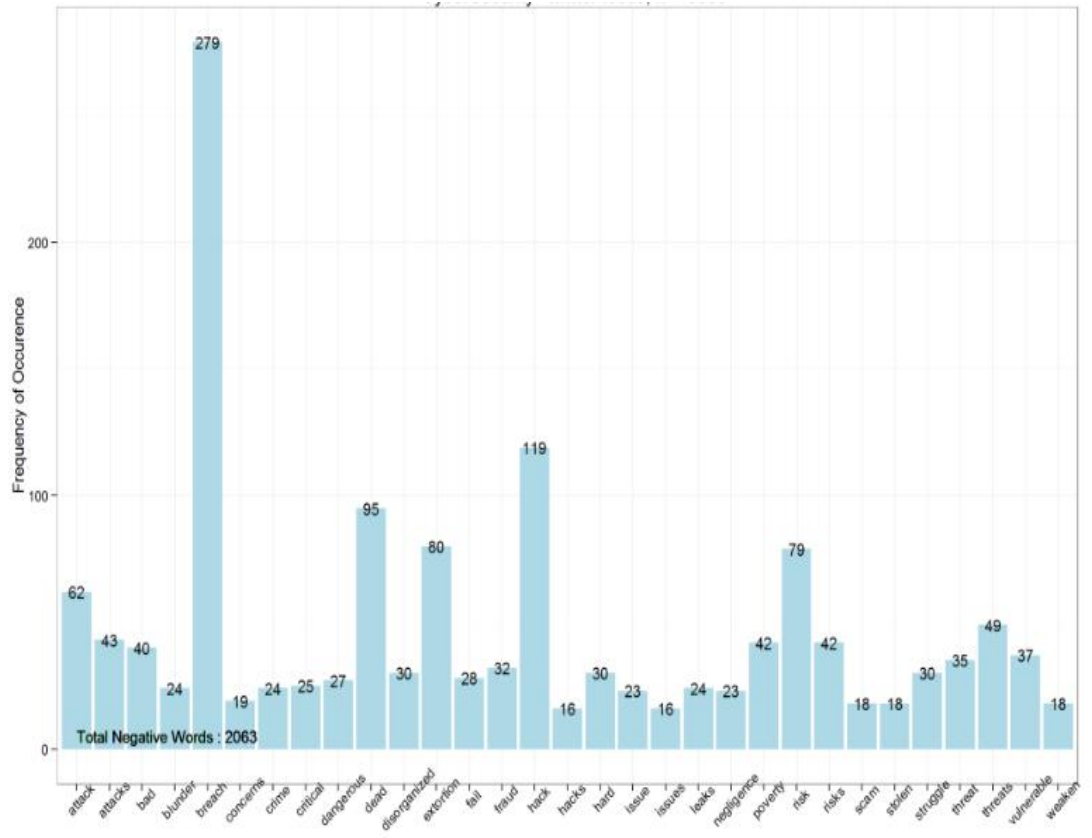


Figure 3.12–Major negative words

4ECONOMIC ANALYSIS OF DEVELOPMENT

A feature of modern business processes in every sector of social activity is to automate the collection and processing of information for management decisions. However, automation is not possible without the use of software products. The solution of any problem of information associated with the use of not only the system software, but also a variety of software– applications.

Software development projects associated with significant resource costs (labor, material, financial). In this regard, the creation and implementation of each of the software project needs corresponding feasibility (FS).

It should be stressed that the feasibility study related to the assessment of both technical and economic parameters of the future product. It is not justified is the fact, when the feasibility study is only their economic evaluation. In other words, economic calculation "superimposed" on some technical solutions, and every effort to prove the cost-effectiveness of these solutions, while technical evaluation of their absence.

We should start with the technical evaluation of projects and on the basis of enumeration of several alternative technological options to allocate the best, to be confirmed by economic calculations.

Lack of technical feasibility studies assessment impoverishes and makes it one-sided. During creating of software can be used by multiple technologies and select one of them as the most effective for a particular case is a creative author of the project task.

In today's market there are a variety of software tools for developing software. The specialist must be able to from a plurality of embodiments of hardware and software to select those which will create a reliable and high-performance system according to the user requirements, but having a lower value of compared with other possible embodiments.

The project is worth pursuing if it offers certain advantages in comparison with known advanced counterparts or, in extreme cases, by compared with the existing practice. Therefore, before proceeding with the development of software project, the specialist should use appropriate methods to find the most efficient solution providing

the high technical level of the program and giving a significant saving of resources both in the development of the project in the scientific and technical organization (the developer) and in its implementation from the user (buyer, customer).

The aim is to develop software, which solve the problem of sentiment analysis of English-language texts based on the NLTK software package.

To accomplish these goals requires 1 programmer and computer where there is a certain development environment. For writing programmer software took about six weeks at work for four hours a day.

In this project, the development of the student performs at the expense of its free personal time, he does not receive a salary and therefore This project on the cost of software development no programmer. However, there are other implicit costs of this project such costs as the cost of electricity, internet payment, payment food and al. programmer for activity.

Since the programmer will not need to program the interaction of the developed software with hardware PC levels (this will be the development environment), it will completely focused on the implementation of the program directly.

Also for development of software the programmer must have a personal computer, which will conducted software development. Ideal for such purposes for a small laptop or netbook. Advantages of this type of computer equipment is mobility. No need to be tied to a specific place, it is also possible to demonstrate what is already implemented part of the developed software. Another advantage is that there is no need to buy expensive monitors, laptops, he has built.

The cost of the computer on which this was being developed software, is 670 dollars. This laptop has the following characteristics:

Type	A laptop
Operating system	Windows 7 Professional x64
CPU	Intel Core i7 3520M
Number of Cores	2
Memory	6 GB DDR3
Number of slots	2
Screen	15.6-inch, 1366x768, widescreen
Screen type	Glossy
Backlight	LED
Video card type	Discrete and integrated
Integrated video card	Intel HD Graphics 4000
Discrete video card	Nvidia GeForce GT 645M
Type of GPU memory	GDDR 5
Video card memory	2GB
Optical drive	DVD-RW, internal
The capacity of HDD	1 TB
Battery type	Li-Ion
The positioning device	Touchpad
Built-in microphone	Present
Dimensions	375x263x25.4 mm
Weight	2.2 kg

Figure 4.1–Laptop Characteristics

This notebook is a mid-segment and has sufficient computing power for development. Also this laptop has a certain margin of performance in the future. This allows be sure that in 2-3 years this laptop is not yet time to mentally obsolete, and its performance is enough for the modern development in the future of software. As this refers to notebook the middle class, then the application is developed it will be low or average requirements for computers. This will allow software developed on this laptop is much wider, as it will be covered by the park, as the slower machines, and middle.

The laptop has a comfortable enough in size diagonally screen - 15.6 'with enough for a resolution of 1366x768 pixels.

For software development, the following free software or trial versions:

- Python IDE;
- Photoshop CS6 Trial;
- Google Chrome.

For the development of the software used free or trial (Trial) version, so their acquisition doesn't cost anything.

The price of developing the software includes the cost of staff salaries, overhead expenses, contributions to the extra-budgetary funds and injuries, the costs associated with the use of computer time. All of the above costs are not relevant in this thesis project, as the development is a student programmer. Development price amounted to 670 dollars. This is a price for a buying of the laptop for programmer. Everything subsequent developments and projects will be carried out without this article expenses. Therefore, all subsequent projects will already have economic benefit for the programmer.

CONCLUSIONS

The work was brought analysis approaches and directions of development of artificial intelligence due to the complex problems of computer linguistics. It was made popular a detailed analysis of existing frameworks that allow to solve the problem of computational linguistics.

At the third part of this work developed and successfully tested an algorithm based tools software packageNLTK andsuccessfully solves the task due to its flexibility. In the future, the algorithm can be upgraded for other languages by filling according to the relevant files.The method of automatic analysis is developed sentimentalcolored posts on the social network Twitter. An experimental evaluation of the effectiveness of the methods was carried out.

REFERENCES

1. Бісікало О.В. Формальні методи образного аналізу та синтезу природно-мовних конструкцій : монографія / О. В. Бісікало. – Вінниця : ВНТУ, 2013. – 316 с. – ISBN 978-966-641-528-1.
2. Havasi, C. ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge / Catherine Havasi, Robert Speer, Jason B. Alonso // Proceedings of Recent Advances in Natural Languages Processing, – Borovets, Bulgaria, 2007.
3. Chi-Yung Wang Knowledge-based Sense Pruning using the HowNet: an Alternative to Word Sense Disambiguation : A Thesis of Master of Philosophy In Computer Science / Wang Chi-Yung. – Hong Kong, 2002. – 111 p.
4. Брушлинский А.В. Чиможливий штучний інтелект? / А.В.Брушлинский – М., 2008. – С 180-201.
5. Вінер Н. - М.: Наука, електронна версія / Н. Вінер , 1998.
6. Vanderwende L. MindNet: an automatically-created lexical resource / Lucy Vanderwende, Gary Kacmarcik, Hisami Suzuki, Arul Menezes // Proceedings of HLT/EMNLP Demonstration Abstracts, – Vancouver, 2005. – pp. 8–9.
7. FrameNetII: Extended Theory and Practice / J. Ruppenhofer, M. Ellsworth, Miriam R.L. Petruck, Cristopher R. Johnson, Jan Scheffczyk // – 2006. – 166 pp.
8. Beth Levin. English Verb Classes and Alternations: A Preliminary Investigation / Beth Levin. – Chicago: The University of Chicago Press, 1993, – 348 pp.
9. Kipper K. Building a class-based verb lexicon using TAGs. /Karin Kipper, Hoa Trang Dang, William Schuler, Martha Palmer // TAG+5. Fifth International Workshop on Tree Adjoining Grammars and Related Formalisms, – Paris, France, 2000.
10. WordNet: An Electronic Lexical Database / [Reem Al-Halimi, Robert C. Berwick, J. F. M. Burg etc.]; Edited by Christiane Fellbaum. – Cambridge, MA: MIT Press; 1998. – 422 pp.

11. Лесников С. В. Интерактивное конструирование модели семантической системы лексики. // Проблеми інформатики і моделювання. Тезиси дванадцятої міжнародної науково-технічної конференції. – Харків: НТУ "ХПІ", 2012.
12. Рогушина Ю. В., Гладун А. Я., Штонда В.Н. Розробка онтологічних терміносистем інформаційних ресурсів інтернет та їх когнітивних моделей у наукових дослідженнях. // Проблеми програмування № 2–3. Спеціальний випуск – 2010.
13. big-library [Електронний ресурс]: Тезауруси в описі інформації видавничої діяльності. – Режим доступу: http://www.big-library.com.ua/book/77_Informaciine_zabezpechennya_vidavnicnoi_diyalnosti/7192_31_Tezayrysi_v_opisi_informacii_vidavnicnoi_diyalnosti.
14. Building a multilingual database with wordnets for several European languages. [Електронний ресурс]. / Режим доступу: <http://www.ilc.uva.nl/EuroWordNet/>
15. BALKANET: Design and Development of a Multilingual Balkan WordNet. [Електронний ресурс]. / Режим доступу: <http://www.ceid.upatras.gr/Balkanet/>
16. Вердиева З. Н. Семантические поля в современном английском языке / З.Н. Вердиева. – М.: Высшая школа, 1986. - 115 с.
17. Апресян Ю. Д. Лексическая семантика. Синонимические средства языка / Ю.Д. Апресян. – М.: Наука, 1974. – 368 с.
18. Нижегородцева-Кириченко Л.А. Лексико-семантическое поле «ИНТЕЛЛЕКТУАЛЬНАЯ ДЕЯТЕЛЬНОСТЬ»: опыт концептуального анализа (на материале существительных современного английского языка: дис. канд. филол. наук: 10.02.04 / Лариса Алексеевна Нижегородцева-Кириченко. – К., 2000. – 257 с.
21. Langacker R. W. Foundations of cognitive grammar. V. 1: Theoretical prerequisites / Langacker R.W. – Stanford, CA: Stanford University Press. 1987. – xi, 516 p.

22. Langacker R. W. Cognitive grammar. A basic introduction /R.W. Lang'acker. –NewYork: OxfordUniversityPress, 2008. – 562 p.
23. Жаботинская С. А. Концепт / домен: матричная и сетевая модели / С. А. Жаботинская // Культура народов Причерноморья. – 2009. – № 168, Т. 1. – С. 254-259.
24. Ламб С. М. С какими данными работает нейрокогнитивная лингвистика / С. М. Ламб; пер. с англ. Н. Р. Бухарева // Компьютеры, мозг, познание: успехи когнитивных наук. – М., 2008. – С 180-201.
25. Скороходько Э. Ф. Семантические сети и автоматическая обработка текста / Э.Ф. Скороходько. – К.: Наукова думка, 1983. – 218 с.
26. Капра Ф. Паутина жизни. Новое научное понимание живых систем / Ф. Капра; пер. с англ.; под ред. В.Г. Трилиса. – М.: ИД «Гелиос», 2002. – 336 с.
27. Жаботинская С.А. Лексическое значение: принципы построения концептуальной сети // InPstyga A. (Red). Slovo z perspektywyjezykoznavcy s ilumacza–Gdansk: WydawnictwoUniversvtetuGdanskiego, 2005. – P. 53-62.
28. Fillmore Ch. J. Frame semantics / Ch. J. Fillmore // Linguistics in the Morning Calm. Seoul, 1982. –P. 111-137.
29. Жаботинская С. А. Онтологии для словарей тезаурусов лингвокогнитивный подход. Філологічнітрактати. – 2009 – № 2, Том 1.
30. WordNet: An Electronic Lexical Database / [Reem Al-Halimi, Robert C. Berwick, J. F. M. Burg etc.]; Edited by Christiane Fellbaum. – Cambridge, MA: MIT Press; 1998. – 422 pp.
31. Кунда Н.Т. Методинауковихдосліджень / Н.Т. Кунда, О.М. Куницька. - К.: НТУ, 2007. - 83 с.
32. Теор Г.А. Информационные технологии: уч. - М.: Юнити, 2000. - 335 с
34. AppStore: VoiceDreamReader– [Электронный ресурс] – Режим доступа до ресурсу: <https://apps.apple.com/ru/app/voice-dream-reader/id496177674>

35. GitHub - westonruter/html5-audio-read-along: HTML5 AudioRead-AlongSector – [Електронний ресурс] – Режим доступу до ресурсу: <https://github.com/westonruter/html5-audio-read-along>
36. ESVAPI | ESV.org – [Електронний ресурс] – Режим доступу до ресурсу: <https://api.esv.org/>
37. What is Whispersync for Voice? – [Електроннийресурс] – Режимдоступудоресурсу: https://audible.custhelp.com/app/answers/detail/a_id/7371/~/~what-is-whispersync-for-voice%3F
38. AmazonTranscribe – [Електронний ресурс] – Режим доступу до ресурсу: <https://aws.amazon.com/ru/transcribe/>
39. How to Use Amazon’s Whispersync for Voice and Immersion Reading Features – [Електроннийресурс] – Режимдоступудоресурсу: <https://blog.the-ebook-reader.com/2016/04/02/how-to-use-amazons-whispersync-for-voice-and-immersion-reading-features/>
40. Burgareli, LA (2009, Jul.-Dec.). Variability management in software product lines using adaptive object and reflection. Journal of Aerospace Technology and Management, V. 1, № 2. [electronic resource]. - Access: http://www.jatm.com.br/papers/vol1_n2/JATMv1n2_thesis_abstracts.pdf. - Name of the screen.
41. O. Bisikalo. Formal methods imagery analysis and synthesis of natural language constructions: Monograph / O.V. Bisikalo. - Ball, NTB, 2013. - 316 p.
42. Hraschenko LA O model stop-dictionary / LA Hraschenko // Proceedings of the Academy of Sciences of the Republic of Tajikistan. Oddzial physical and mathematical, chemical, geological and of technical sciences - 2013. - № 1 (150). - P. 40-46.
43. Котик Я.Ю. Розв’язання задачі емоційного аналізу тексту на основі лінгвістичного пакету DKProCore[Електронний ресурс] / О.БісікалоЯ.Котик. – Режим доступу до ресурсу: <https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2019/paper/view/7148/6177> - Назва з екрану.

ATTACHMENTS

ATTACHMENT A

(Required)

Ministry of Education and Science of Ukraine

Vinnytsia National Technical University

Faculty of Computer Systems and Automation

Approved

Head of the AIIS Department, Dr. of Sci., Prof.

_____ Roman N. Kvetny

“ ____ ” _____ 2019 year

Approved

Head of the course work, Dr. of Sci., professor

_____ Oleg V. Bisikalo

“ ____ ” _____ 2019 year

DEVELOPED BY

Student of the 3ACIT-18m group

Yaroslav Yu. Kotyk

“ ____ ” _____ 2019 year

TECHNICAL TASK

for master's qualification work

on the topic

« INFORMATION SYSTEM OF INTELLECTUAL ANALYSIS OF
NATURAL-LANGUAGE CONTENT OF SOCIAL NETWORKS. PART 3. THE
SUBSYSTEM OF DETERMINATION OF SENTIMENTAL CHARACTERISTICS
OF PARTICIPANTS OF SOCIAL NETWORKS ON THE BASIS OF LINGUISTIC
ANALYSIS OF MICROBLOGGING »

Designated product name: Intelligent System for sentiment analysis of participants of social networks on basis of linguistic analysis of microblogging.

Product use area: Work results can be used at other scientific research . The achievements in sentiment analysis can be used in other industries.

Reason for product development:

1. Curriculum of specialty 151 – Automation and computer-integrated technologies.
2. Individual assignment for master's qualification work

Main functions of the program:

1. Collecting information for the database
2. Sentiment analysis of the information
3. Output data to the database.

Requirements for a software product:

1. Operating system – WINDOWS/Linux/macOS.
2. Programming environment – arbitrary.
3. Required memory – 10 GB of disk space
4. Software should be tested and made available on its suitability for use.

All software and accompanying technical documentation must comply with the following GOSTs:

GOST 19.701-90

ISO 5807-85 - GOST for development of software documents, schemes of program algorithms, data and systems.

GOST 19.781-74 - requirements for software development.

GOST 19.401-78 – The text of the program. Requirements for content and design.

GOST 19.106-78 – Requirements for software documentation.

GOST 7.1.-84 and DSTU 3008-95 – development of technical documentation.

Development stages:

1. Classification of the type of task
2. Setting the task
3. The choice of methods for solving the problem
4. Software implementation
5. Testing
6. Conclusions

The order of control and acceptance of master's qualification work:

Obtaining a task for qualification work – «__» _____ 2019 year.

The term of the term paper is checked – till «__» _____ 2019 year.

Term of protection of qualification work – till «__» _____ 2019 year.

ATTACHMENT B

(Required)

APPROVED

Head of the AIIT department,
Dr. of Sci., Prof. Roman N. Kvetny

“ _____ ” _____ 2019 year

LIST
OF GRAPHIC MATERIALSfor the protection of master's qualification work
on the topic

« INFORMATION SYSTEM OF INTELLECTUAL ANALYSIS OF
NATURAL-LANGUAGE CONTENT OF SOCIAL NETWORKS. PART 3. THE
SUBSYSTEM OF DETERMINATION OF SENTIMENTAL CHARACTERISTICS
OF PARTICIPANTS OF SOCIAL NETWORKS ON THE BASIS OF LINGUISTIC
ANALYSIS OF MICROBLOGGING »

1. Definition of key words in the text microblogging algorithm
2. UML-diagram of Naïve Bayes decision tree algorithm
3. Screen view of the output file
4. Diagram of the major positive words
5. Diagram of the major negative words
6. Screen view of the database «mathling»

Developed by Yaroslav Yu. Kotyk

(signature)_____
(date)

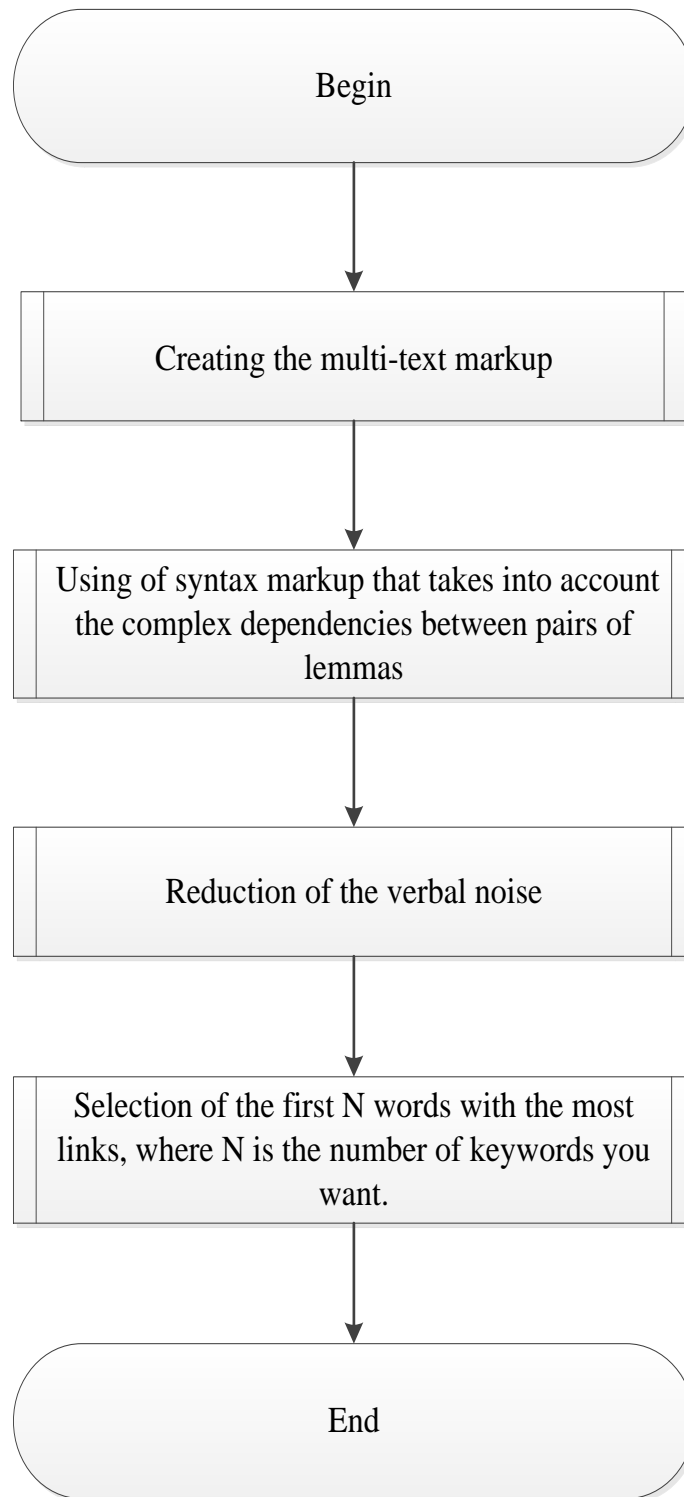
Checked by Oleg V. Bisikalo

(signature)_____
(date)

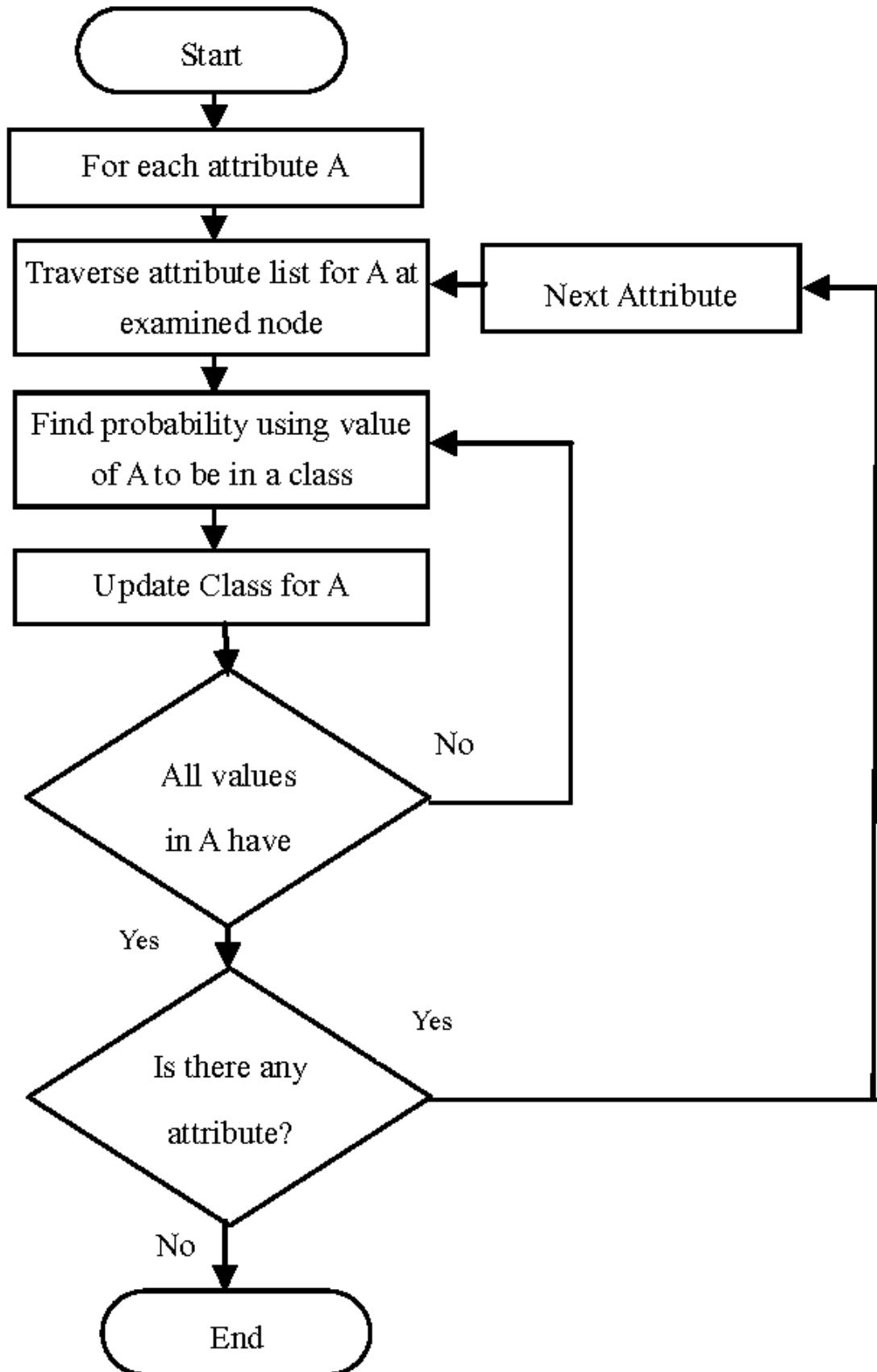
Reviewed by Vyacheslav V. Kovtun

(signature)_____
(date)

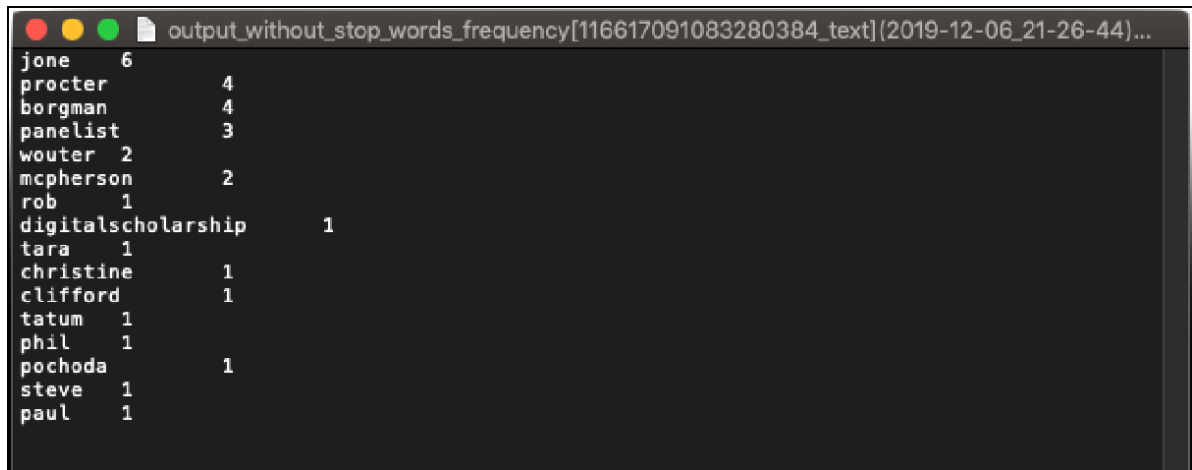
Vinnytsia 2019



B1–Definition of key words in the text microblogging

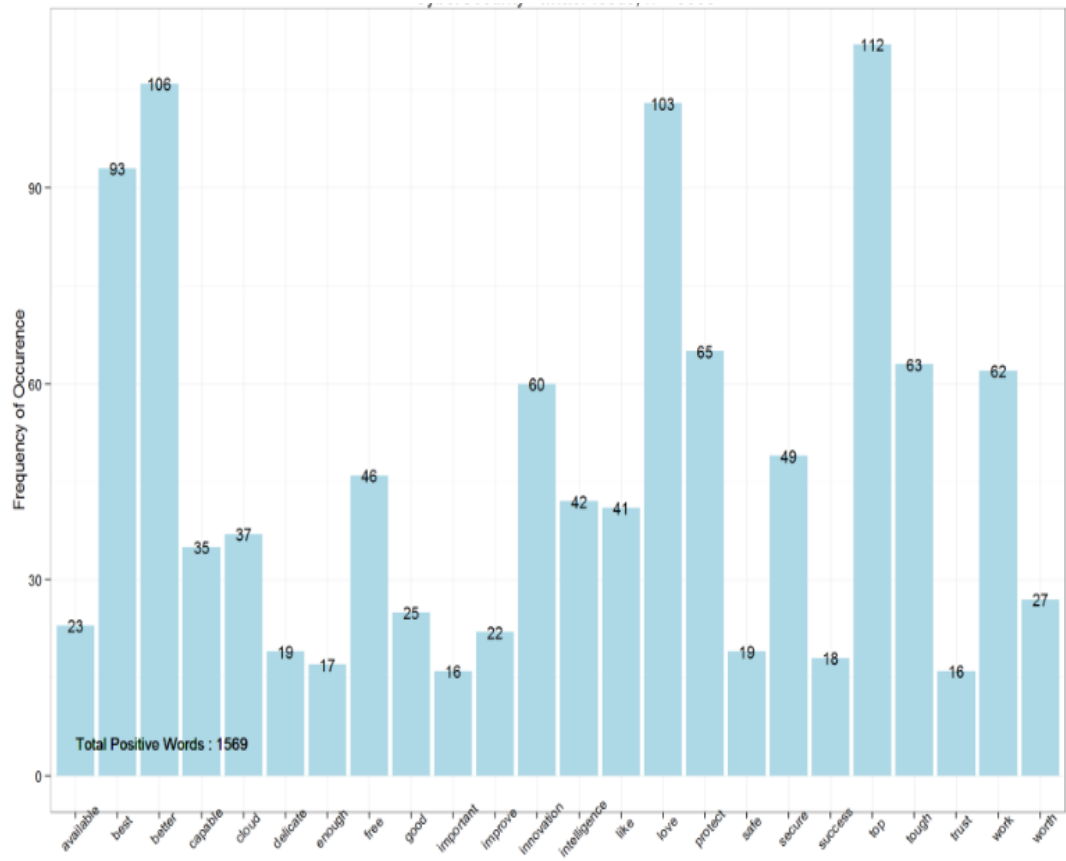


B2 –UML-diagram of Naïve Bayes decision tree algorithm

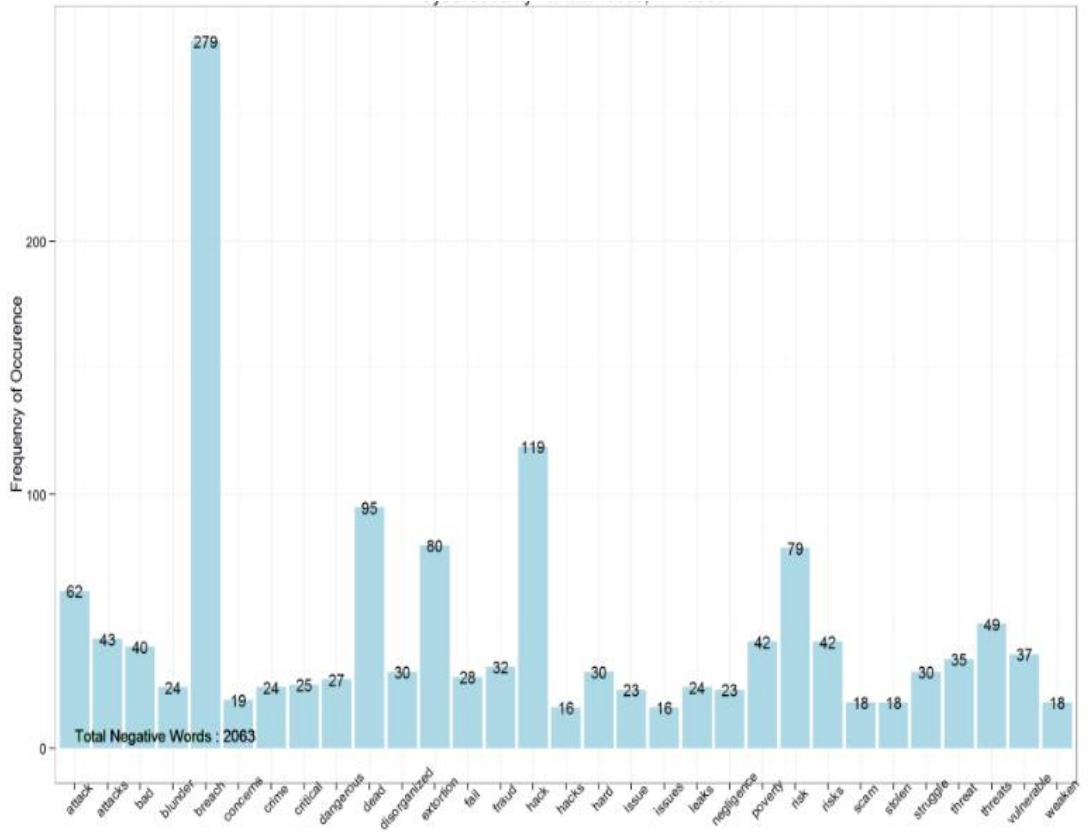


```
output_without_stop_words_frequency[116617091083280384_text] (2019-12-06_21-26-44)...
jone 6
procter 4
borgman 4
panelist 3
wouter 2
mcpherson 2
rob 1
digitalscholarship 1
tara 1
christine 1
clifford 1
tatum 1
phil 1
pochoda 1
steve 1
paul 1
```

B3 –Screen view of the output file



B4 –Major positive words



B5 –Major negative words

The screenshot displays the phpMyAdmin interface for a MySQL database named 'mathling'. The 'posts' table is selected in the left sidebar. The main area shows a table with the following columns: id, id_str, created_at, text, and user_id. The 'text' column header is circled in red. The table contains 24 rows of data, each with a set of action icons (Edit, Copy, Delete) to its left. The first row of data is:

id	id_str	created_at	text	user_id
1000002552073736193	1000002552073736193	Fri May 25 13:16:02 +0000 2018	Heading into the long weekend like... #MemorialDay ...	38441438

B6–Database «mathling»