

Přírodovědecká fakulta

Klasifikace zdravotnických dat prostřednictvím neuronových sítí

Diplomová práceDiplomová práceDiplomová práce

Bc. Hoa Vu Thu

Vedoucí práce: RNDr. Martin Komenda, Ph.D.

Pracoviště

obor Matematická biologie

Brno 202020202020



Bibliografický záznam

Autor: Bc. Hoa Vu Thu  
Přírodovědecká fakulta  
Masarykova univerzita   
Pracoviště

Název práce: Klasifikace zdravotnických dat prostřednictvím neuronových sítí

Studijní program: Experimentální biologie

Studijní obor: Vyberte český název oboru

Vedoucí práce: RNDr. Martin Komenda, Ph.D.

Rok: 2019/20202019/20202019/2020

Počet stran: 3

Klíčová slova: [Napište 5–10 klíčových slov v češtině. Stejný seznam musí být vložen do archívu závěrečných pracích v informačním systému MU.]

Bibliographic record

Author: Bc. Hoa Vu Thu  
Faculty of Science  
Masaryk University  
[Vyberte anglický název katedry nebo ústavu]

Title of Thesis: Health data classification using neural networksHealth data classification using neural networksHealth data classification using neural networks

Degree Programme: Experimental Biology

Field of Study: Mathematical BiologyMathematical BiologyMathematical Biology

Supervisor: RNDr. Martin Komenda, Ph.D.

Year: 2019/20202019/20202019/2020

Number of Pages: 3

Keywords: [Napište 5–10 klíčových slov v angličtině. Stejný seznam musí být vložen do archívu závěrečných pracích v informačním systému MU.]

Abstrakt

[Napište abstrakt (500–600 znaků včetně mezer) v češtině. Shodný text abstraktu musí být vložen do archívu závěrečných pracích v informačním systému MU.]

Abstract

[Napište abstrakt (500–600 znaků včetně mezer) v angličtině. Shodný text abstraktu musí být vložen do archívu závěrečných pracích v informačním systému MU.]



Čestné prohlášení

Prohlašuji, že jsem svoji diplomovoudiplomovoudiplomovou práci vypracovalavypracovalavypracovala samostatně pod vedením vedoucího práce s využitím informačních zdrojů, které jsou v práci citovány.

V Brně 13. března 201913. března 201913. března 2019 ....................................... Bc. Hoa Vu Thu

Poděkování

[Zde můžete napsat poděkování (není povinné).]

Obsah

Lists of figures 15

Lists of tables 16

Seznam pojmů a zkratek 17

1 Introduction 19

1.1 Czech national health system and web presentation 19

1.2 Natural language processing in health care 21

1.3 Research questions 21

2 Methods 22

2.1 CRIPS-DM 22

2.2 Natural Language Processing approaches and algorithms 23

2.3 Models for Distributed representations of words 27

2.4 Technologies 31

3 Results 33

3.1 Data 33

3.2 Data preprocessing 34

3.3 Modeling 34

3.4 Evaluation 34

3.5 Deployment 34

3.6 Evaluation 34

4 Discussion 35

5 Conclusion 36

Použité zdroje 37

Příloha A [Název přílohy] 39

Rejstřík 40

List of figures

[Figure 1: Scheme of CRISP-DM 23](#_Toc31621783)

List of tables

Nenalezena položka seznamu obrázků.

List of terms and abbreviations

[Heslo] – [Definice]

# Introduction

The health care industry generates large amount of information on daily basis in form of reports from specialists, prescriptions, doctors’ notes, description new mutated diseases, lots of paperwork or even websites related with this industry. Lot of people can be drowned in this huge amount of data and it became harder to orientated because of its volume. As a consequence, it is hard to identify which information is relevant and which one is not.

In this thesis the main goal is to create and implement a model for measuring semantic similarity of two text documents and classify these documents into two class {“similar”, “not similar”}. This model should be easy to use for bigger projects of IHIS and Faculty of Medicine of Masaryk University for example as a part of a recommendation system based on the content on websites. It can help students or even public to find deeper understanding of some health-related question. Beside of that it can also raise awareness of IHIS’s and faculty’s websites that are credible sources of information, which can be very helpful too regarding to the amount of misinformation nowadays.

In the first chapter, there is introduction of IHIS and some of its project that have its own website. There is also reasons why it is important to use natural language processing (NLP) in health care domains, there is also mentioned some of the real work, when NLP was used in health care industry and some of the issues that every data scientist has to deal with when working with this specific type of data. At the end of the first chapter there are some research question that needed to be answered when processing this diploma thesis. In the second chapter there are described methods that are used in this thesis – CRIP-DM and neural networks their general description and specific models that are used in NLP for capturing semantic connection between words. The third chapter is devoted to results of models that were used. In the next chapter is discussion of results following conclusion.

## Czech national health system and web presentation

The Institute of Health Information and Statistics of the Czech Republic (IHIS) is an organisational component of the Czech Republic. It belongs to the Ministry of Health who delegated it to administrate the National Health Information System (NHIS). According to the Health Services Act the IHIS was specifically entrusted with the administration of:

* Data from the Program of statistical investigations of the Ministry of Health collected according to the Act on state statistical service
* National health registers including:
* the Czech National Cancer Registry
* the National register of hospitalisations
* the National register of reproduction health
* the National register of cardiovascular surgery and intervention
* the National register of joint replacement
* the National register of occupational diseases
* the National register of drug addict therapy
* the National register of injuries
* the National register of persons permanently excluded from blood donations
* the National register of autopsy and toxicological examination performed at forensic medicine departments
* the National register of health care providers
* the National register of health care professionals
* Data taken over from information systems on infectious diseases maintained according to Act on public health protection

All this data is used for monitoring health status of the population, the activity of health care providers, their economy and for obtaining the information about extend and quality of provided health services and for creation of health policy. NHIS is also designated for conducting and processing surveys on health status of population, on determinants, on the need and the consumption of health services, on its satisfaction and expenditure, for the needs of science and research in the field of the health.

The IHIS has currently several projects for example “Early detection of thyroid disease” in, or “Codification for Rare Disease”. For purpose of this thesis there will be introduced 2 of many projects and that is:

“**Methodological optimization and streamlining of the system of reimbursement of hospital care in the Czech Republic” -** this project was brought to the existence for the purpose of creating an information system of inpatient care reimbursement called CZ-DRG (diagnosis related group) and by its implementation there will be higher effectivity in management, controlling and optimization of distribution of financial flows in the health care.

The main goal is to build a long-term data, information and personnel base for the optimization and continuous cultivation of the system in the Czech Republic and to increase the predictive ability and effectiveness of reimbursement mechanisms for health care. For achieving the main goal a Competence Centre and a representative network of reference hospitals had to be built and now it consists 49 hospitals, methodologies for evaluation of hospital cases, electronic and database tools for the implementation of the system had to be developed and verified in a representative network of reference hospitals, a new classification of hospitalization procedures had to be formed, the DRG classification system has to be redefined and a new coder’s manual had to be designed. The redefined DRG classification system can be found on <https://drg.uzis.cz/klasifikace-pripadu/web/>.

“**Centre for the Development of the Technology Platform of National Health Information System Registers, Modernization of their Content Extraction and Extension of their Information Capacity**” - the target of this project is methodological development of NHIS, increase its information capacity and value and also development of necessary supporting software applications and components. It will also contribute to more efficient collection of data, better usage of information by the state administration and overall development of the system of departmental registries NHIS.

## Natural language processing in health care

Information that healthcare industry makes every day is usually stored in unstructured and non-standardized formats. They are composed of various grammatical structures, expressions. This is when NLP techniques stepped in. They help to give some structure to narrative information. They are able to capture unstructured information by extracting relevant information, analyze its grammatical structure, determine the meaning of the information and make it understood by computer’s language or in other words, they can interpret the meaning of unstructured text.

#priklady s clankami odcitovanymi

One of widely use of NLP is to create a chatbot. They can play an important role in medical care, they can assist clinicians during consultations, support consumers with behavior change challenges or assist patients in their living environment. This can lead to reduction of the workload of medical staffs. By extracting information from clinical notes, it is possible to identify cardiovascular risks factor. Analyzing responds of mental and physical health it is possible to predict suicide of patient after a psychiatric treatment.

There are still many challenges of NLP in health care. One of them is incompatible vocabulary meaning one concept can be expressed by multiple concepts or vice versa for abbreviations. For example, the abbreviation APC can refer to Activated Protein C, Advanced Pancreatic Cancer or Antibody Producing Cells. In some cases, tokenization can be a problem. Some words can contain hyphens (N-acetylcysteine) or slashes (10 mg/day). Text can also contain spelling errors like “hyprtension” which can refer to hypertension or hypotension, so it is often unclear. Those are reasons why it is important to capture semantic similarity in medical text.

## Research questions

When solving problem there is a lot of questions that need to be answered.

* What kind of data is needed for this kind of problem and where to obtain them?
* Is it necessary to preprocess data?
* What is a minimum of observations/documents for training model successfully and not overfitted it?
* What model is best suited for this type data and problem?
* Is it better to use a supervised model or an unsupervised model?
* Is it relevant to use pre-trained models or there has to be created new model?
* How to improve models?

All these questions were explored and in following chapters there will be explanation of the answer of every question.

# Methods

This chapter covers detailed description of used data, how these data were preprocessed, application of methods on data and their detailed description with some partial results. All this information will be introduced in following subchapters regarding to a data mining method called Cross Industry Standard process for Data Mining (CRISP-DM). The only exception is first step, to be precise business understanding step. This step is already characterized in previous chapter.

As it is illustrated in figure, this method is consisted of 6 steps:

1. **Business understanding –** it is important to understand what is the main goal that is needed to be accomplished, define the outputs, design the project plan, check availability of resources, list requirements and consider pros and cons.
2. **Data understanding –** in this phase data are described and explored to understand them and their structure, also their quality must be verified.
3. **Data preparation –** this is the most time-consuming stage when it has to be decided on what data are going to be used and they need to be edited (cleaning, transformation, creating new attributes etc.) for further use.
4. **Modelling –** different models are applied, validated and optimized on data
5. **Evaluation –** results are evaluated, their quality is checked, and based on them the best model or models are chosen regarding to the prime goal
6. **Deployment –** create deployment, monitoring and maintenance plan and present the results to customer



Figure 1: Scheme of CRISP-DM

Order of these steps does not need to be kept. Output of one step is the input for following step. All these steps are connected mutually, and it is possible to go step back to make changes. The whole process has cyclic character for the reason that according to achieved results there can be more specific or even new research questions.

## Data understanding

Traditional data analysis works with structured data that has definite forma, they are easy to store and analyse with hight degree of organization. A typical instance is a table. On the other hand, there are unstructured type of data that are not that easy to handle as structured data. They have not easily identifiable structure and they are not organised in a pre-defined manner, hence it cannot be stored in the form of rows and columns. For example, it can be a text, audio or video file. Semi-structured data are irregular data that may be incomplete and have a structure that changes rapidly or unpredictably however they does not conform to a fixes or explicit schema. An email or XML documents are one of most cases of semi-structured data.

Data that was obtained for this diploma were textual. They came from <https://drg.uzis.cz/klasifikace-pripadu/web/> and <https://reporting.uzis.cz/>. Those three websites are created by IHIS, where is information about the health care in the Czech Republic in Czech language.

The <https://drg.uzis.cz/klasifikace-pripadu/web/> is website representation of project “Methodological optimization and streamlining of the system of reimbursement of hospital care in the Czech Republic”. This website is divided into five sections:

* **Introduction** –there are information about the project, data sources and contacts
* **Structure of system CZ-DRG** –in this section is description of each taxonomic level of system CZ-DRG and its labels
* **Definition lists** –there can be found classification rules for every hospitalization case and characterization of taxonomic level
* **Interactive classifier (grouper)** –in this section it is possible to classify hospitalization cases according to classification rules of system CZ-DRG regarding to its latest version
* **Analyses and publications** –this section provides published analytical outputs related to the CZ-DRG classification system or other detailed views of its taxonomic units.

The latter website used in testing model was created as a part of “Centre for the Development of the Technology Platform of National Health Information System Registers, Modernization of their Content Extraction and Extension of their Information Capacity” project. Its purpose is to provide information on health care and health situation in the Czech Republic. At the homepage there is possibility to choose from three options:

* News for the Czech Republic
* News for regions
* News for city Brno

The structure of website is for all three options the same. It has five sections:

* **Introduction** –there is news, some information about the website, sources of data that are used in their analyses, contacts and some useful links
* **Summary reports** –there are outputs in a form of tables, graphs and maps that give a quick comprehensive information about:
  + demographic and sociodemographic indicators,
  + health status indicators,
  + economic indicators, health care infrastructure, physician and medical workers.
* **Statistical outputs** –there are more detailed outputs of summary reports. The outputs are also in a form of tables, graphs and maps
* **Analyses and publications** –there are described analyses and presentations of selected topics
* **Help**

Data were crawled from both website whereas from the first website were collected 1908 documents and from the other one XYZ documents.

## Data preparation

In every data analysis is this step crucial and results of every analysis depends on it. If garbage data are input, then the analysis will also produce garbage results as output. This is a reason, why it is important to pay attention in preparing data. There are several techniques that are used when working with unstructured textual data.

**Filtering** –this is done when it is necessary to remove some words or characters. A common type of word filtering is to get rid of stop-words. This type of words appears in text frequently, but it is not very informative (like prepositions, conjunctions, pronouns or numerals). By removing then the volume of corpus will be reduced. In case of filtering characters, there is often needed to remove special characters (".", "," ,";", "-", "(", ")", "[", "]", '"\"', "/", " "), digits and roman numerals. If documents are XML documents, there is also HTML tags removal.

**Tokenization** – it is a form of text segmentation, when only alphabetic or alphanumeric characters are considered, and they are delimited by non-alphanumeric characters. In other words, it is identifying the smallest unit of a document called tokens. It can be single word, or a group of words called n-grams.

**Lemmatization** – methods that maps verb form to the infinite tense, nouns to the singular form, adjectives to singular masculine form, so the part of speech of every word has to be assigned. This process is time consuming and still error prone. In Czech language there are two tools for lemmatization - Majka and UDPipe.

UDPipe is a pipeline that preforms sentence segmentation, tokenization, POS tagging, lemmatization and dependency parsing. It can be also trained using data in CoNLL-U format.

**Stemming** – methods that try to obtain the basic form of words – stem, by removing suffixes and prefixes, thus every word is represented by its stem.

**Lowercase conversion** – uppercase and lowercase form of words are often assumed to have no difference, so all uppercase characters are converted to their lowercase form to reduce vocabulary dimension

Obtained data from websites were in json format, where were information if the website was crawled, the particular URL that was crawled, its parent URL and its specific XML document. Every XML document has structure and because of that it was possible to predefine paths where useful information was stored. This was done by using Xpath, which is used to select nodes from this type of documents, which leads to creating a corpus.

Corpus was pre-processed by removing all numbers, special characters and stop words in data and by transforming all letters in words into lower case except for abbreviations such as HIV. The corpus was then tokenized into words and lemmatized applying UDPipe. At this point, there was a lot of trouble due to inconsistent data. There were some parts of documents, where sentences were written in upper cases. UDPipe was able to detect some commonly used words and convert them to lower case, but some medical terms such as “laparoskopie” or names of drugs like “gamunex” had to be converter manually. Also, there was a lot of abbreviation that is often used in Czech language, for example “urč.”, “onem.” or “bakter.”

Corpus was

## Modelling

### Natural Language Processing approaches and algorithms

Natural Language Processing (NLP) is an interdisciplinary area of research and application that investigates how computers understand and process natural language text or speech. One of the techniques of NLP is to measure text similarity that is important in text related research and applications in tasks, for example information retrieval, text classification, checking plagiarism etc. Finding similarity between words is an important part of text similarity which is then used as a primary stage for sentence, paragraph and document similarities.There are 2 approaches how to measure text similarity:

* **Lexical similarity**
* **Semantic similarity**

**Lexical similarity** provides the similarity based on surface of text, so words are similar if they have a similar character sequence. To measure lexical similarity **string-based** algorithms are used which can be divided into **character-based** and **term-based** similarity measures.

**Semantic similarity** is similarity between texts is determined by their meaning, so words are similar semantically if they have the same thing, are opposite of each other, used in the same way, used in the same context and one is a type of another. Corpus-based and knowledge-based algorithms are used to measure semantic similarity. **Corpus-based similarity** computes similarity between words based on the information that is gained from large corpora. **Knowledge-based** similarity is used to determine the degree of similarity between words based on information derived from semantic networks. The most popular semantic network is WordNet which is a big lexical database of English language.

For the purpose of this diploma thesis there will be detailed description corpus-based similarity more specifically about word embeddings and deep learning methods on embeddings that use neural networks.

### Neural networks

Biological neuron

The nerve cell, also called neuron, is the basic building and functional unit of the nervous system for generation, processing and propagation signals. Thanks to these signals, information is transmitted from the internal and external environment. This is necessary condition for the organism’s ability to react and adapt to changes. The basic parts of neuron are the cell body, dendrites and axon.

The cell body is consisted of a plasma membrane, a neuroplasm, a nucleus with a distinct nucleus and other cellular organelles. Due to the large amount of energy required to maintain and restrain the resting membrane potential, the neuron body contains a large number of mitochondria.

Branched extension of body cell is called dendrites. They propagate the electrochemical stimulation received from other neurons to the cell body. Because of their specialization, they contain a lot of receptors that take part in chemically controlled ion channels.

The propagation of the signal itself is in charge of the axon (neurite), which leads the information away from the body of the neuron. Axons can range in length from a fraction of a millimetre to a few meters, the branching itself is typically only in the terminal part. The signal (nerve impulse) is transferred electrically, when the signal from one cell to another is transmitted mostly chemical way. This process uses ion transfer across membranes, which causes a change in action potential (so-called depolarization) and conduction of information to dendrites another neuron. After the passage of ions, so-called repolarization occurs, during which the return occurs ions and return to their original potential. For the transport of information, it is important that it is higher the number of stimuli, since a single stimulus depolarization change is subliminal for a neuron. Thus, it does not create excitement. It is only thanks to the time and space summation that it is possible to use the accumulation of these signals, and thus the transmission of information.

### Artificial neuron

The origins of artificial neural networks date back to 1943, when Warren McCulloch and Walter Pitts were inspired by biological neuron and they constructed the first neuron model.

In general, neural network (NN) is set of units – neurons. Several units create a layer. There are 3 types of layers: input, output and hidden layer. Layers are connected each other and every connection between two layers has its own weight. In general, the input units receive a signal from the outside of the NN, they process this signal and sent it to the follow-up units – the hidden units, which process the received signals in the same way as the input unit. This process is repeated within the hidden layer until the signals is received to the output units. The input to each unit is compute as:

**Activation function** is a primitive function that transforms input of the unit into defined output and it decides whether a neuron will be activated or not (analogy with action potential in biological neuron). The purpose of the activation function is to give nonlinearity into the output of a neuron making it capable to learn and preform more complex tasks. There are several known activation functions:

Training the neural network include two steps that create one epoch:

* Feedforward step – predicting the outcome
* Backpropagation - updating weights

**Backpropagation** gradient descent – updating weights and biases minimize the cost function by adjusting network’s weights and biases

**Loss/cost function** is used to evaluate the goodness of the prediction from the trained neural network. By the process of backpropagation, the cost function is minimized. There are many loss functions that can be used.

The key in training neural network is to find the best set of weights and biases that minimizes the loss function, but it is also important to avoid overtraining the network.

### Neural network for NLP

**Recurrent neural network**

Recurrent neural network (RNN) is similar with feedforward neural network. The main difference between those two networks is that RNN has a feedback loop in their architecture which provides that RNN has “memory”. Mathematically it is written as:

Thanks to their ability to memorize they are used for sequential data, such as time series, genomes or text.

There are known two types of RNN:

* **Long short-term memory (LSTM)**
* **Gated recurrent unit (GRU)**

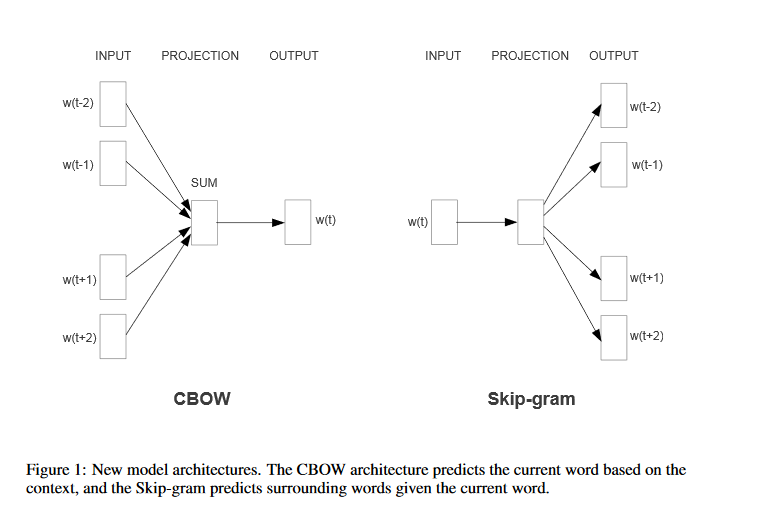
### Models for Distributed representations of words

Nowadays distributed word representations, which will be explained in this chapter, have been widely used in NLP. It determined the similarity of words using words that occur around them (context). These models are based on the distributional hypothesis that says the meaning of a word is captured by the contexts on which it appears. This implies that both synonyms and antonyms are intended as similar words, since words with the opposite meaning are most likely to occur in the same context. In these models, each word in the dictionary is assigned a vector within the multidimensional vector space. The similarity of words is then determined as the similarity of their vectors. These vectors are obtained by corpus analysis. Thus, the quality of the word vectors directly depends on the amount and quality of word vectors directly depends on the amount and quality of data they were trained on.

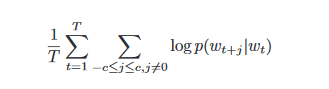
### Word2Vec

Word2Vec are several models used in natural language processing (NLP) to produce a word embedding. In general, word2vec is a two-layer neural network that is used to recreate linguistic context of words. As its name refer, word2vec transforms a corpus of text into a vector space with a big dimension and every word is assigned to a vector in the space. These vectors are called word vectors. Two words vector are close to each other when they share a similar context or meaning. Basically, it treats each words in corpus like an atomic entity and generate a vector for each word. The input and output layer contain word vectors that are being trained. Input layer vectors are used as model output vectors. The input layer is a matrix of size V × N, output N × V where V is the number of words in the dictionary and N is the size of the vector. The vectors in both matrices are not identical. The hidden layer has a size of N. All weights are initially initialized to random values and then adjusted during training due to back propagation. Word2Vec use two architecture: **continuous bag-of-word (CBOW) and continuous skip-gram**. It is a classifier into n classes, where n is the dictionary size. Ideally, the probability distribution is determined by softmax regression, that the word belongs to that class (word). For simplicity, consider one word as context. Using the Stochastic Gradient Descent, the algorithm tries to minimize the difference between the output distribution and the target probability distribution. The target distribution is determined from the context of the word and takes only one value in the current word, zero elsewhere. Because softmax regression is computationally demanding for large dictionaries, Word2Vec uses its approximations, which are less computationally computational – hierarchical softmax.

**Continuous bag-of-words** predicts the current word from the context. The order of context words does not affect prediction. In contrast, the continuous skip-gram uses the current word to predict the surrounding words and the weighs of words near to the current word are greater than the weights more distant from the current word. Continuous skip-gram is slower that CBOW but it has a better performance in case of infrequent words.



**Skip-gram model** – this model search for word representations that help to predict the surrounding words in a sentence or document. Assume that we have a given sequence of training words *w1, w2, w3 … wT*. The skip-gram model maximizes the average log probability



Where *c* is the size of the training context (which can be a function of the centre word *wt*). With the larger *c,* the higher accuracy can be obtained but the training time can be also increased.

**Hierarchical softmax**

If we calculate the probability of two words occurring together with normal softmax functions, the training would be very slow:

For this reason, word2vec uses hierarchical softmax that is more effective that compute softmax by word representation with a binary tree. Word2Vec uses Huffman's binary tree, in which the most common words are assigned the shortest code. Each word in the dictionary must be a leaf of a tree and a unique path from the root leads to each leaf. Use this path to estimate the probability of a word represented by a leaf. Each inner node has determined probabilities for passing to the left or right subtree. The product of these probabilities along the path to the sheet is the resulting probability.

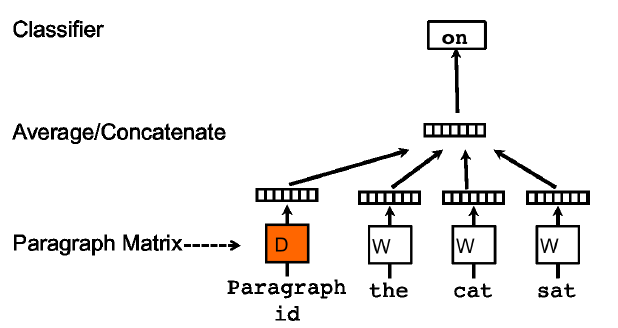
**Negative Sampling**

Negative sampling is another way to determine probability without having to scroll through all the words in the dictionary. Based on the probability distribution based on the frequency of occurrences of the word in the corpus, it is selected for other words to be considered.

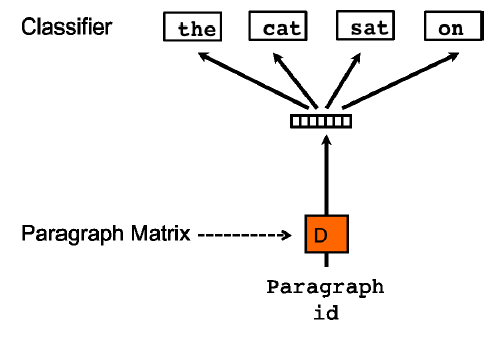
### Doc2Vec

Paragraph vector is an unsupervised framework, it learns continuous distributed vector representation for pieces of texts (sentences or even documents). Doc2Vec is based on Word2Vec. The vector representation is trained to predict words in a paragraph. The paragraph vector is concatenated with several word vectors from a paragraph and predict the following word in the given context. Words vectors and paragraph vectors, which are unique among paragraphs in contrast with word vectors, which are shared, are trained by stochastic gradient descent and backpropagation. At prediction time, the paragraph vectors are inferred by fixing the word vectors and training the new paragraph vector until convergence. Doc2Vec has two approaches: **A distributed memory model of Paragraph Vectors (PV-DM)** and Distributed **Bag of Words of Paragraph Vector (PV-DBOW)**.

In the PV-DM every paragraph is mapped to a unique paragraph vector that is represented by a column in matrix D. Words are also mapped into a matrix W, where one word is represented by a column in matrix W. The paragraph vector and word vectors are averaged or concatenated to predict the next word in a context. Both vectors are trained using stochastic gradient descent. The paragraph token can be thought as another word and it acts as a memory which remembers what is missing from the current context.



The PV-BDOB approach the context words in the input is ignored. Instead of that the model predicts words randomly sampled from the paragraph in the output. This model requires to store less data in contrast to PV-DM. It is very similar to Skip-gram model in Word2Vec.



### FastText

FastText uses a hierarchical classifier which reduces the time complexities of training and testing (from linear to logarithmic with respect to the number of classes). It also exploits the fact that class are imbalanced by using Huffman algorithm.

Essentially it is an extension of word2vec but instead of considering whole words it considers sub-words. It treats each word as composed character of n-grams. The word vector is made of a sum of this character n-grams. It generates better word embeddings for rare words. It can construct the vector for a word from its character n-grams even if word does not appear in training corpus

## T-SNE visualization

## Neural networks used for text comparison

For comparing similarity of two documents there were chosen two approaches:

* MaLSTM
* Doc2vec

Because of MaLSTM requires a pre-trained word embedding, data from OPTIMED were used for this purpose. Embeddings were created using Word2Vec and FastText. In both cases the skip-gram and CBOW architectures were applied and all of the methods were compared by visualizing the most similar word of a concrete word. The following figures shows the result of a word “sval”.

One of models that is able to capture similarity between text documents and quantify it is by Siamese adaptation of the Long Short-Term Memory. That means there are two identically weighted LSTM neural network. Siamese networks perform well on similarity tasks and they are used in sentence similarity, recognizing signature or … . The LSTM is used for accepting variable-length inputs, which is documents, they do not have a problem of vanishing gradient, and they are able to learn long range dependencies thanks to its memory cell unit. This helps to capture semantic differences during training.

Two documents are encoded into two word-vector by pre-trained word2vec model, in this case it was trained on data from OPTIMED. Each document, represented as a sequence of word-vectors) is passed into the LSTM layer, where the output of this layer are two vectors that contain all information from documents. The similarity of vectors is computed by cosine similarity and it is rescaled into range from 0 to 1, where 0 indicates that documents are completely different and 1 means that documents are identical regarding to semantic.

## Technologies

UDPipe, which is trainable pipeline for tokenization, lemmatization and dependency parsing of CoNLL-U files developed by Institute of Formal and Applied Linguistics at Charles University.

**Software**

Python 3., anaconda,

**Libraries?**

# Results

# Discussion

# Conclusion

Použité zdroje

URL 1: <https://www.uzis.cz/> The Institute of Health Information and Statistics of the Czech republic [online] [cit. 23.November 2019]

URL 2: <https://hecc.ubc.ca/quantitative-textual-analysis/qta-theory/preprocessing-texts/> The university of British Columbia [online] [cit. 30.March 2020]

1. [Název přílohy]

Rejstřík

Nebyly nalezeny položky rejstříku.