SCIENTIFIC PUBLICATIONS OF THE STATE UNIVERSITY OF NOVI PAZAR
SER. A: APPL. MATH. INFORM. AND MECH. vol. 13, 1 2021, 17-25.

Creating a Stop Word Dictionary in Serbian

U. A. Marovac, A. M. Avdić, A. B. Ljajić

Abstract: By using natural language processing techniques, it is possible to get a lot of information from the extraction of document topics through mapping of document key words or content-based classification of documents, etc. To get this information, an important step is to separate words that carries informative value in a sentence from those words that do not affect its meaning. By using dictionaries of stop words specific to each natural language, the marking of words that do not carry meaning in the sentence is achieved. This paper presents creating a stop word dictionary in Serbian. The influence of stop words to the text processing is presented on three different data set. It is shown that by using proposed dictionary of Serbian stop words the data set dimension is reduced from 15% to 39%, while the quality of the obtained n-gram language models is improved.

Keywords: stop words, Serbian, text mining, natural language processing, normalization

1 Introduction

In the era of IT expansion, text mining is especially important in many fields because it can lead to new knowledge. By using text mining techniques, we get answers from search engines, companies based on our writings on social networks conclude about our needs and offer us products and services, etc. It is therefore of utmost importance to separate content which carries informative value in a sentence from those words that do not affect its meaning and serve to make meaningful connections between significant words. Before any processing of text documents, it is necessary to normalize the document, which reduces the documents to a form adapted for further processing. Marking stop words is one of the steps of normalizing textual data. Stop words usually are denoted by using available language-specific lexical resources. During this marking, care must be taken what is the goal of text document processing. In some situations, there are words from a set of stop words participate in sentence forms that are important for further processing, such as word negation. For example, the influence of negation processing on word sentiment analysis may be of

Manuscript received June 28, 2020. ; accepted November 12, 2020.

U. A. Marovac is with the State University of Novi Pazar, Novi Pazar, Serbia, umarovac@np.ac.rs; A. M. Avdić is with the State University of Novi Pazar, Novi Pazar, Serbia, apljaskovic@np.ac.rs; A. B. Ljajić is with the Faculy of Electronic Engineering, Niš, Serbia, adela.crnisanin@gmail.com;

great importance [8] so in this case stop words which are part of negation can't be omitted. Some stop words are abbreviations or expressions in some specific documents, so in these cases it is necessary to create domain-dependent stop word dictionaries. In the process of normalization, stop words are sometimes excluded from documents, while sometimes they are only marked because their presence helps further analysis of the document. All of this points to the importance of having stop word resources for a particular language. In Serbian there is no adequate stop word resource in electronic form suitable for further use, as well as a method for automatically extracting stop words from the corpus. Available sets of stop words are usually created by translating this resource from other languages and are incomplete in terms of number of stop words and their description. This paper describes a domain-independent, Serbian-language stop word resource that can be used for various applications. The experimental section provides a comparison of the created resource with existing resources and the impact of stop word removal on quality of obtained language models. The rest of this paper is structured as follows. Section 2. describes the current achievements and methods for removing stop words in different languages and especially for the Serbian language. Section 3. describes words that do not have meaning in the Serbian language and are therefore classified in the stop words dictionary. The data sets that will be used in the experiment are described in Section 4. A description of the language model to be used is given in Section 5. Section 6. contains the results of applying a language model to given data sets using different stop word dictionaries.

2 Related work

The significantly affect of processing a large amount of textual data have the noise that they contain. Stop words are one of the noise source in textual data. Many authors have dealt with improving the quality of text document processing by extracting a list of stop words [3]. The common approach is to manually assemble a list of stop words from list of words and it has been shown that it shows great accuracy and that can be applied adequately to different domain of texts [11, 12]. Automatic recognition of stop words is a task that the authors were among the first to deal with in [14]. They used statistical tests and cosine similarity. The automatic building of a stop word based on the information that the word carries is given in the paper [9] where the information of the word is calculated using the Kullback-Leibler divergence measure. They showed that comparable performance is obtained in relation to baseline methods for calculating informativeness (based on Zipf's law [13]) with less computational effort. These analyzes were done for the English language. Modern approaches of word embeddings and contextualized word embeddings almost never filter stop words before training of the model. Sometimes, filtering stop words when training word vectors using word embeddings can have a negative impact because they can provide some context. However, it has been shown that stop words have among the most context-specific representation in contextualized word embeddings [4]. This leads us to the conclusion that stop words have a specific behavior and that it is useful to have stop word resources, if not for filtering, then for their eventual special processing. For Serbian and related languages, to our knowledge, there are no approaches to automatically extract stop words, but they are created using available lexical and grammatical resources.

3 Words that do not carry as much meaning

Stop words are words that do not carry text meaning and are not important for the appropriate analyzes performed on the text. There are two type of stop words:

- domain-independent: words specific to the language in which the text is written,
- domain-dependent: words specific to the domain to which the text belongs.

The first group includes words that, regardless of the domain to which the text belongs, have no meaning and are specific to the appropriate language, they are: auxiliary verbs, pronouns, articles in English. They are defined by relevant experts and users cannot change the resources in which these words are stored.

When we talk about domain-specific stop words, these are words that appear frequently in documents of the same domain. For example, in the movie reviews, the words “actor” and ”film” appears frequently, or in the medical reports, the words that appear frequently are: ”patient”, ”report”, ”doctor”, etc. Users can also define this domain dependent stop word set as words that are not important. It should be noted that the meaning of a word is exclusively related to the analysis that is implemented over the text. Words which are marked as stop words in an analysis, can be marked as a keyword in another analysis. Thus, some stop words play a significant role in the accomplishing of negation in sentence construction, so they should be excluded from the set of stop words in the methods for processing negation. The word ”kontrola” (eng. control) in medical reports is a stop word in the process of detecting anamnesis with certain symptoms, while in the process of classification the anamnesis on the first and control examinations this word becomes a key word [1]. Stop words are also important in the Part-of-speech tagging process and in that case, they are not removed. There are ten types of words in the Serbian language [7]: nouns, adjectives, pronouns, numbers, verbs, adverbs, prepositions, conjunctions, particles, and exclamations. Nouns, adjectives, and verbs are the types of word that mostly carry the meaning in a sentence. Depending on the type of text analysis performed, different groups of words more or less influence the outcome of the corresponding analysis. That is why it is especially important to have a stop word dictionary with clearly defined word categories. The stop word dictionary contains words that are often used and do not carry significant information, they belong to the following types of words: auxiliary verbs, pronouns, adverbs, prepositions, conjunctions, exclamations, particles. Table 1 contains the examples for each group of stop words for the Serbian language.

There is no publicly available stop word dictionary for the Serbian language. There are some sets with a small number of stop words in Serbian, of which we create the stop dictionary (SW set) which contains 133 stop words, and it is available on [5]. In this paper is presented a new Serbian stop word dictionary (SSW dictionary) of 1241 different stop

Table 1. Table of examples for Serbian stop words

Type of words	Examples of stop words (Serbian)	Examples of stop words (English)
Verbs	jesam, nisam, bih, hoću, želim, mogu	I did, I didn't, I would, I will, I want, I can
Pronouns	on, ona, ono, njoj, njeno, svoj, vaš,..	he, she, it, her, her, your, your, ..
Adverbs	sada, nekada, sutra, dugo, davno, često,..	now, once, tomorrow, long, long ago, often, ..
Prepositions	do, duž, zbog, iz, iza, izvan, iznad,...	to, along, due to, from, behind, outside, above, ...
Conjunctions	i, a, ili, da, ako, dok, čim, jer, mada	and, a, or, yes, if, while, as soon as, because, though
Exclamations	Ha-ha, ho-ho, jao, avaj, kuku, lele, o, ah,...	Ha-ha, ho-ho, wow, alas, hook, wow, oh, ah, ...
Particles	baš, upravo, možda, valjda, ipak,...	just, just, maybe, I guess, still, ...

words for the Serbian language. It was manually created based on the grammar of the Serbian [7] as well as by comparing with available sets of stop words for the Serbian language and a set of stop words for the Croatian language [10]. Our stop word dictionary (SSW dictionary) for Serbian language contains words in different forms of their appearance. Each word is accompanied by a word type label. Figure 1 shows the percentage of different word types in the dictionary. The dictionary contains the largest number of pronouns that have no meaning without analyzing the context to which they refer. Verbs and adverbs have an important role in text analysis. Due to the frequent use of adverbs, they are included in SSW dictionary, but in some specific analyzes they can be excluded. Auxiliary and modal verbs that often occur do not carry significant information, so they are also added to SSW dictionary in all forms. Conjunctions, exclamations, prepositions have no meaning, but they play a role in the structure of the document, as well as some abbreviations that are often used, all these groups of words are included in the SSW dictionary.

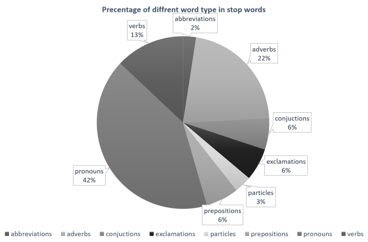

Fig. 1. Precentage of diffrent word type in stop words

4 Data set

We tested the obtained resource on three sets of data:

1. Data set of medical anamneses (electronic health records EHRs)- DS1_EHR
   - The set contains 2000 anamneses;
   - The set is composed of EHRs that are joined by 10 different diagnoses (B00, B01, B02, H10, H650, H66, J11, J18, N390, S60).
2. Data set of tweets – DS2_tweet
   - The set contains 6041 tweets;
   - The tweets were collected using twitter streaming API from a profile of 15 official mass media, 10 musicians, 5 public figures from the entertainment world, 5 athletes and 7 actors.
3. Data set of movie review—DS3_review
   - Set contains 1500 movie reviews;
   - A part the Serbian movie review data set [2].

It should be noted that these are three different types of documents. The first set contains formal specialized short texts. Anamnesis lengths range up to 70 words. The set contains a lot of professional terms. Sentences are often without subject and incomplete form. The second set contains informal short texts related to various topics. The maximum length of processed tweets is 32 words. The third set contains long texts from the same domain written in formal but still literary language. The length of movie reviews is up to 2074 words.

5 N-gram language model

Language models offer a way to assign a probability to a sentence or other sequence of words, and to predict a word from preceding words. The N-gram language model is the most widely used language modeling approach. An N-gram is usually written as an N-word phrase, with the first N-1 words as the history, and the last word predicted as a probability based on the N-1 previous words. N-grams models are Markov models that estimate words from a fixed window of previous words. N-gram probabilities can be estimated by counting in a corpus and normalizing it (the maximum likelihood estimate). N-gram language models are evaluated extrinsically in some task, or intrinsically using perplexity. The perplexity of a test set according to a language model is the geometric mean of the inverse test set probability computed by the model.The perplexity (sometimes called PP for short) of a language model on a test set is the inverse probability of the test set, normalized by the number of words [6]. The perplexity for test set $W = w_1w_2\dots w_n$, is given with Formula 1

$$PP(W) = \sqrt[n]{\prod_{i=1}^{n} \frac{1}{P(w_i|w_1w_2\dots w_{i-1})}} \tag{1}$$

There are problems of balance weight between infrequent grams and frequent grams. Also, items not seen in the training data will be given a probability of 0.0 without smoothing. To keep a language model from assigning zero probability to these unseen events, we will have to shave off a bit of probability mass from some more frequent events and give it to the events we've never seen. This modification is called smoothing or discounting. The simplest way to do smoothing is to add one to all the n-gram counts, before we normalize them into probabilities. All the counts that used to be zero will now have a count of 1, the counts of 1 will be 2, and so on. This algorithm is called Laplace smoothing [6].

6 Results and discussion

Removing stop words from a data set reduces the amount of data that will be used. We have shown the results of using the SSW dictionary on different data sets and compared it with the results of using the available set of stop words for the Serbian language. Table 2. presents the percentage of stop word from the SSW dictionary in relation to the SW set for different data sets. Our set of stop words contains significantly more words, and it reduces the size of the data set by 15% to 39%, which is significant for further data processing. We can notice the higher presence of stop words in informal communication and movie reviews than in the text form the specific domain, as in medical reports.

Table 2. The percentage of stop word in data sets

Data set	#words	% stop words from SSW dictionary	% stop words from SW set
DS1_EHR	17924	15%	2%
DS2_Tweet	99094	29%	5%
DS3_Review	840551	39%	8%

The percentage of different type of stop words (SSW dictionary) in data sets is shown on Figure 2. The most common occurrences are adverbs and preposition word types.

A N-gram language model predicts the probability of a given N-gram within any sequence of words in the language. It uses the previous N-1 words in a sequence to predict the next word. The impact of stop word removal on the quality of the language model using the method for model evaluation (perplexity) is examined. The N-gram language of the data set of anamneses and tweets was evaluated for: unigrams, bigrams, and trigrams. The Laplace smoothing algorithm was used to calculate the perplexity value. The perplexity value of the data sets and the data sets from which are removed stop words is calculated.

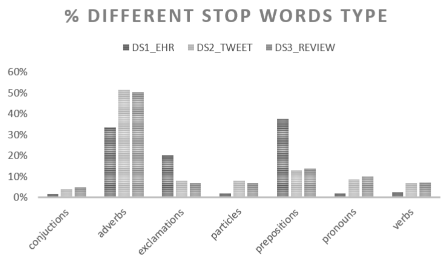

Fig. 2. Precentage of diffrent word type in stop words

The results for two data sets are presented, and the influence of two different set of stop words (SSW dictionary and SW set) is compared. Table 3 shows the perplexity values for the DS1-EHR n-gram data set. The table shows the estimate of the N-gram model on the anamneses data set, as well as on the anamneses data set which is cleaned of stop words by using: the available stop words set (SW set) and our created stop word dictionary (SSW dictionary). It can be concluded that removing the stop word increases the predictability of the data set. The value of ppl is also reduced by removing the stop word from the SW set, but it is better to use the SSW dictionary. Since these are short texts for specific domain, the number of stop words in these texts is limited, but we can notice that their removal does not affect the lost of significant information.

Table 3. The perplexity for n-gram language model of anamneses

Data set	Unigram PP	Bigram PP	Trigram PP
DS1_EHR	434.58	307.60	754.82
DS1_EHR/ SW set	330.99	281.57	735.02
DS1_EHR/ SSW dictionary	325.74	279.60	727.28

Tweets are also short texts, but they are informal and contain a larger number of stop words, so the impact of their removal is greater. Table 3 shows the ppl values on the DS2_tweet, as well as on the tweet data set which is cleaned of stop words from the corresponding dictionaries. For unigrams, we can notice that removing stop words from the SSW set increases the perplexity, which can be explained with the high frequency of words from the SSW dictionary in tweets. However, the prediction of the bigram and trigram is significantly improved by dropping the stop word from SSW dictionary, which indicates that dropping the stop word improves text analysis.

Table 4. The perplexity for n-gram language model of tweets

Data set	Unigram PP	Bigram PP	Trigram PP
DS1_EHR	1261.68	4260.08	7437.84
DS1_EHR/ SW set	1195.56	3830.79	7002.20
DS1_EHR/ SSW dictionary	1453.16	2721.64	5242.02

7 Conclusion

By using the created dictionary of stop words, the size of data sets is significantly reduced, which facilitates their further processing by NLP methods. It is shown that the created data set contains a much larger set of stop words as well as different forms of stop words, so they are better mapped. It is shown that the removing stop words by using SSW dictionary improved the performance of the N-gram language model. For text which are specialized as medical text this improvement is smaller so it can be concluded that it is necessary to create domain specific stop words. Hence, in the future, we will continue to work on creating a dictionary of stop words based on the text domain.

Acknowledgment

This work was partially funded by the Ministry of Education and Science of the Republic of Serbia after the project III-44 007.

References

[1] A. R. Avdić, U. A. Marovac and D. S. Janković, *Normalization of Health Records in the Serbian Language with the Aim of Smart Health Services Realization*, Facta Universitatis, Series: Mathematics and Informatics, (2020), 825-841.

[2] V. Batanović, *The Serbian Movie Review Dataset (SerbMR)*, https://vukbatanovic.github.io/project/serbmr/

[3] M. Choy, *Effective listings of function stop words for twitter*, arXiv preprint arXiv:1205.6396, (2012).

[4] K. Ethayarajh, *How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings*, arXiv preprint arXiv:1909.00512 ,(2019).

[5] https://github.com/Xangis/extra-stopwords/blob/master/serbian

[6] D. Jurafskyand J. Martin,*Speech and Language Processing: An Introduction to Natural Language Processing*, Computational Linguistics, and Speech Recognition.

[7] I. Klajn,*Gramatika srpskog jezika*, Zavod za udžbenike i nastavna sredstva,(2005).

[8] A. Ljajić and U. Marovac, *Improving sentiment analysis for twitter data by handling negation rules in the Serbian language*, Computer Science and Information Systems, Vol. 16(1), (2019), 289-311.

[9] R. T. W. Lo, B. He and I. Ounis, *Automatically building a stopword list for an information retrieval system*, In Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR),Vol. 5,(2005, January), 17-24.

[10] R. Lujo, *Locating similar logical units in textual documents in Croatian Language*, Master Thesis (in Croatian), Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia, (2010).

[11] C. Silva and B. Ribeiro, *The importance of stop word removal on recall values in text categorization*, In Proceedings of the International Joint Conference on Neural Networks, IEEE, Vol. 3, (2003, July), 1661-1666.

[12] M. P. SINKA and D. CORNE,*Evolving Better Stoplists for Document Clustering and Web Intelligence*, In HIS, (2003, January), 1015-1023.

[13] K. ZIPF, *Selected Studies and the Principle of Relative Frequency in Language*, Cambridge, MA; MIT Press, 1932.

[14] W. J. WILBUR and K. SIROTKIN, K.,*The automatic identification of stop words*, Journal of information science, 18(1),(1992), 45-55.