 doi:10.20944/preprints202204.0234.v1

Automatic detection of stop words for texts in the Uzbek language

Khabibulla Madatov,
Urgench state university, 14, Kh. Alimdjan str, Urgench city, 220100, Uzbekistan;
E-mail: habi1972@mail.ru

Shukurla Bekchanov,
Urgench state university, 14, Kh. Alimdjan str, Urgench city, 220100, Uzbekistan;
E-mail: shukurla15@gmail.com

Jernej Vičič
University of Primorska, UPFAMNIT, E-mail: jernej.vicic@upr.si
Research Centre of the Slovenian Academy of Sciences and Arts, The Fran Ramovš Institute;
E-mail: jernej.vicic@upr.si

Keywords: stop word detection, Uzbek language, agglutinative language, algorithm

Stop words are very important for information retrieval and text analysis investigation. This study aimed to automatically analyze and detect stop words in texts in the Uzbek language. Because of the limited availability of methods for automatic search of stop words of texts in Uzbek we analyzed a newly prepared corpus. The Uzbek language belongs to the family of agglutinative languages. As with all agglutinative languages, we can explain that the detection of stop words in Uzbek texts is a more complex process than in inflected languages: In inflected languages, words such as auxiliary words, articles, prepositions can be included in the stop words group. In agglutinative languages, the meanings of such words are hidden in the text. Therefore, it is not appropriate to apply all known methods of stop words detection in inflected languages directly to agglutinative languages. In this work, the "School corpus" which contains 731156 Uzbek words has been investigated. The bigram method of analysis was applied to the corpus. We proposed the collocation method of detecting stop words of the corpus. We proposed the method of automatically detecting stop words of texts in Uzbek. It is shown that the collocation method is 6 times better than the bigram method.

Povzetek: Mašila - stop besede - stopwords so zelo pomembne za iskanje informacij in preiskavo analize besedila. Namen te študije je bil samodejno analizirati in odkriti takšne besede v besedilih v uzbeškem jeziku. Zaradi omejene razpoložljivosti metod za samodejno iskanje stop besed v uzbeškem jeziku smo analizirali na novo pripravljen korpus. Uzbekistanski jezik spada v družino aglutinativnih jezikov. Tako kot pri ostalih aglutinativnih jezikih, je odkrivanje stop besed v uzbeških besedilih bolj zapleten proces kot pri pregibnih jezikih: v pregibnih jezikih so lahko besede, kot so pomožne besede, členki, predlogi, vključene v skupino stop besed. V aglutinativnih jezikih so pomeni takšnih besed skriti v besedilu. Zato priznane metode za odkrivanja stop besed v pregibnih jezikih ne moremo neposredno uporabiti v aglutinativnih jezikih. V tem delu je bil raziskan "šolski korpus", ki vsebuje 731156 besed v uzbekistanskem jeziku. Za korpus je bila uporabljena metoda analize, ki temelji na bigramih. Delo tudi predlaga kolokacijsko metodo odkrivanja stop besed korpusa in metodo samodejnega zaznavanja stop besedil v uzbekistanskem jeziku. Dokazano je, da je metoda kolokacije 6-krat boljša od metode na osnovi bigramov.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2022 doi:10.20944/preprints202204.0234.v1

1 Introduction

Uzbek language belongs to the Eastern Turkic or Karluk branch of the Turkic language family. External influences include Arabic, Persian and Russian. It belongs to the family of agglutinative languages. As with all agglutinative languages, detection of stop words in Uzbek texts is a more complex process than in inflected languages: in inflected languages, words such as auxiliary words, articles, prepositions form most of the stop words group. In agglutinative languages, the meanings of such words are hidden in the text. Therefore, it is not suitable to apply all known methods of stop words detection in inflected languages directly to agglutinative languages. The experimental results presented in this work that the use of a hybrid method (combining grammatical rules and statistical methods) yields best results in the task of detecting stop words for texts in Uzbek. As a result of this work we compare this method with bigram method (both methods are thoroughly presented in the paper). When someone works on a novel, a story, an article, or a text, this person uses semantic connection of words with artistic decoration in their own language to make it meaningful and interesting. Dealing with sentences, stop words which do not have an independent meaning or have little meaning are often used. As a result, the text size increases. As the volume of information increases, the process of data processing and analysis slows down and as the search space increases, the quality of results (searches) is potentially lowered. In such cases, removing unnecessary words from the text can reduce the amount of information and increase the efficiency of electronic data processing. It is also important to automate the generation of annotations and keywords from large volumes of text. The main purpose of identifying unimportant words is to facilitate automatic text analysis.

2 Related works

Stop words detection methods can be divided into two basic categories:

1. Based on grammar rules,

2. Statistical methods.

In this work, we use both categories for automatic detection of stop words in Uzbek texts.

2.1 Based on grammar rules

The sources mainly provide grammatical rules for finding stop words or a list of stop words for different languages [1], [2], [3], [4], [5],[6],[7],[8], [9]. The text is grammatically analyzed to identify stop words in Uzbek texts. According to the definition of stop words, words in the Uzbek language that are part of a rhyme, conjunctions, introductory words, adverbs, auxiliary words can be stop words. It is required to automatically separate them from the given text. Due to the lack of syntactic analysis programs in the Uzbek language, using a dictionary, a list of words that are supposed to be stop words will be given. In order to create the list of stop words from the dictionary we investigate and take into account the definition of stop words. In general pronouns, adverbs, connectors, Introductory words can be stop words in Uzbek texts.

2.1.1 Pronouns

Pronoun is a part of speech used instead of a noun, adjective, number. The meaning of pronouns and which word or words they substitute is defined by the context (intra or inter sentence). According to the meaning and grammatical features, the pronoun is divided into generalized - subject (pronouns - nouns: men (I), sen (You), u (he, she, it), kim (who), nima (what), hechkim (nobody), hechnima (nothing), generalized - nominal (pronouns-adjectives: bu (this), shu (this), o'sha (that), qaysi (which), allaqanday (somehow), hechqanday (no), generalized quantitative (pronouns-numerals: qancha (how much), necha (how many), shuncha (so many), o'shancha (so many). Pronouns differ from other parts of speech in polysemy, lack of word formation. Pronouns by meaning and grammatical features are divided into the following types: pronouns of the person –men (I), sen (you), u (he,she,it), biz (we), ular (they) used instead of persons, the proper pronoun - consist of a proper word denoting an object, strengthening its meaning, emphasizing it; indicative pronouns –bu(this), shu (this), o'sha (that), u (he, she, it) , ana (that),etc. indicate an object and its signs; interrogative pronouns indi-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2022 doi:10.20944/preprints202204.0234.v1

cate the questions as- kim? (who) nima? (what) qancha? (how many) of the subject, attribute and quantity; definitive-collective pronoun - indicates the generalization, generalization of the subject and its features in relation to hamma (all), bari (all), ba'zan (sometimes), har nima (anything), har qanday (every/any) indefinite pronoun - expresses the denial of meaning in relation to hechkim (noone), hechqanday(no), hechqanaqa (no), hechqaysi (none).

2.1.2 Adverbs

An adverb is one of the independent types of Parts of speech denotes a sign of an action and a state, as well as a sign of a sign. There are the following types of adverb meanings: state adverbs (tez(quick), sekin(slow), piyoda(on foot)); adverbs of a place (uzoqda(far), yaqinda(near), pastda(below); adverbs of time (hozir(now), kecha(yesterday), bugun(today); adverbs of quantity (ancha(much), sal(little), kam(few); adverbs of purpose ataylab(deliberately), jo'rttaga(willingly); adverbs of reason (e.g., noiloj(helplessly), ilojsiz(helplessly), chorasizlikdan(helplessly). All forms, except for the forms of time, place and purpose, according to the most general characteristics, can be attributed to one type and calledas status forms. An adverb as an independent phrase is characterized by the following morphological features:

- has a category of degree: tez (quick), ko'p(much) (oddiy daraja(simple degree)) — tezroq (quicker), ko'proq (more) (comparative degree) — eng tez(the quickest), judako'p(much more)(superlative degree);
- remains unchanged and is often associated with verbs: So'ridaqat-qatduxoba ko'rpachalar ustma-ust to'shalgan edi(The couch was covered with layers of velvet mattress);
- an adverb can also be associated with an adjective and a noun in some places. In such cases, the adverb does not indicate a sign of a sign or a sign of an object, but to the adjective to which it is attached, or to a sign of action understood from a noun: Kecha havo juda sovuq edi. (It was very cold last night).U hozir beqiyos va tasavvur qilib bo'lmas baxtiyor edi;(Now, he was incomparably, unimaginably happy);
- an adverb has suffixes: -cha, -ona, -larcha, -laband etc..
- Adverbs are formed in morphological and syntactic ways: (otlashish hollari bundan mustasno), for each time, including the moment, as in the moment (syntactic method).

By structure, adverbs are divided into simple (kamtarona (modest), vijdonan (conscientious), butunlay(whole)), compound (har dam (always), bir yo'la (together), oz muncha (much), har qachon (always)), paired (kecha-kunduz (day and night), qishin-yozin (winter and summer) and repeated (oz-oz (little by little), tez-tez (often), ko'p-ko'p(many-many). The modal form is considered to be such forms of verbs anchagina (much), juda (very), kam (little), kam-kam (little by little). Adverbs act in a sentence as case, determinant and cut. Adverbs are similar to adjectives in terms of the properties of the expression of a feature, but differ among themselves in grammatical properties: adjectives denote a feature of an object, an adverb that is a sign of an action or state; their function in the sentence, that is, syntactic, is also special.

2.1.3 Connectors

Connectors are auxiliary words that serve to link organized parts of a sentence and simple sentences in the structure of a combined sentence are called connecting words. Auxiliary words that connect two or more fragments of a sentence or sentence are called connectors.

Introductory words. Introductory words-words that are not syntactically related to the sentence. Expresses the speaker's attitude to the expressed thought ("baxtimga" (fortunately), "afsuski"(unfortunately)), the general assessment of the thought ("ehtimol", "albatta"), to whom it belongs ("menimcha"(to my mind), "aytishlaricha"(it is said)) or its connection with the preceding thought ("xullas"(so), "nihoyat"(finally)). Words used in a sentence in the function of an introductory word, expressing the speaker's attitude to the expressed thought, are called modal words. Modal words are not independent words, such as a thing, sign, action, etc., and cannot be

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2022 doi:10.20944/preprints202204.0234.v1

part of a sentence. Therefore, they are not syntactically related to the fragments of the sentence: “Demak, ishlasa bo‘ladi. Ehtimol, ketmon bilan yer ag‘darishga ham to‘g‘ri kelar.” (So, it can endure to work. Perhaps, you may even have to roll over with a hoe).

2.1.4 Usage

The rules described in previous subsections were used to detect stop words. The popular explanatory dictionary of the Uzbek language [1] with 80000 words and detected approximately 1100 stop words. These are by definition one-word stop words as they come from the dictionary.

2.1.5 Statistical method

Consider the statistical method of automatic detection of stop words in Uzbek texts. In this method, stop words are found based on the frequency of the word and the frequency of the inverse document Term Frequency – Inverse Document Frequency – TF-IDF [10]. The number of times of word occurrence in a text is defined by Term Frequency – TF. Inverse Document Frequency – IDF is defined as the number of texts (documents) being viewed and the presence of a given word in chosen texts (documents). TF-IDF is one of the popular methods of knowledge discovery. There are such words that are so common in the text, however they are almost insignificant in terms of meaning and conversely, there are words that are rare in the text, but they are very important in terms of the meaning of the text. In order to increase the impact of meaningful words and decrease the frequency of words that do not add up much to the meaning, we multiply TF to IDF. We see the statistical method is used as the basis for finding stop words of many languages. The sources mainly use the TF IDF method to analyze of the word of the text [2], [11], [12],[13],[14],[15],[16],[17],[18],[19],[20]. we see the statistical method is used as the basis for finding stop words of many languages. These sources mainly use the TF IDF method to analyze of the word of the text. Several methods to find stop words for Turkish are given in [16]. Comparing the current work with these sources, we bring scientific novelty of the article.

3 Methodology

Scientific novelty of the article. First, a collocation method is proposed for automatically finding stop words of the Uzbek corpus, consisting of 731156 words and comparing with bigram method its advantage is shown. Second, stop words detecting algorithm is proposed for Uzbek texts.

This section is dedicated to the method of automatic detection of stop words in Uzbek texts. The following procedure was used for detection of stop words.

3.1 Corpus

A corpus named “School corpus” was created using freely available school books such as “Reading book”, “Mother tongue” and “Literature”. The texts were downloaded form Eduportal [1]. Total number of documents is 25. The motivation behind the selection of the texts for the corpus was the following:

- everyone enriches personal language dictionary knowledge during the school period,
- free availability of the texts,
- school textbooks are thoroughly checked for errors,
- marginally big enough selection of documents and length of the documents (taken into account low availability of Uzbek texts in digital form in general).

Some basic data about the corpus:

- name: *School corpus*,
- total number of words: 731156,
- number of unique words: 47165.

The investigation on finding stop words in the Uzbek language has shown that in most stop words that are collocations, each single word is not a stop word when viewed as individual word, but when considered as a collocation word, they become stop words. A few examples that further confirm our claim are presented in the Examples 3.1 and 3.2 where the meaning of the

[1] Eduportal: https://eduportal.uz/Eduportal/Barchasi/33

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2022 doi:10.20944/preprints202204.0234.v1

sentences is the same. When viewed as individual words the words *bir* and *martalik* are not stop words, but if they are observed as a collocation, they become stop word(s).

Example 3.1. *Xalqimiz bir martalik shprits vositasida emlanadi.* – (Our people are vaccinated with a disposable syringe.)

Example 3.2. *Xalqimiz shprits vositasida emlanadi* – (Our people are vaccinated with syringes.)

Thus, there is a need to expand the problem of finding stop words which consist of one word. A collocation is considered if there are 2 or more words. Only a two-word collocation is considered in this article and the motivation behind this is that three or more word collocations that act as stop words are not that common, but we still believe that a further work needs to be done in this direction. The proposed methodology does not change for longer collocations.

3.2 Bigram method

For the purpose of this article, the following definition will be used: bigrams are pairs of consequent words appearing in the text. Let's consider the use of the bigram method of finding stop words for the corpus. Algorithm 1 presents the implementation of the bigram method.

The Algorithm 1 applied to the "School corpus" produced 4548 pairs of words as collocation stop words. A few examples are presented in Figure 1 bigram.

3.3 The collocation method's algorithm

The following definition of collocation will be used throughout the article: an occurrence of consecutive words in a corpus. In our case only two-word collocations will be observed. A (two word) collocation and bigram represent essentially the same starting set of word pairs, but bigrams are limited to the most probable pair, collocations take in consideration all pairs.

A collocation is considered for 2 or more words. In this article only a two word collocation is considered. The presented method and derived results are limited to two word collocations for

Algorithm 1: The bigram method

1. Consider total occurrences a_i, a_{i+1} of collocation words in a corpus. Construct a list of unique pairs $UP1$. In our corpus example, the number of such collocations was 731155. Among them 489857 unique pairs.
2. Consider the list $UP1$, for each pair a_i, a_{i+1} from the list take a_i and find the word with the biggest bigram probability in the corpus for the next word a'_{i+1}. There were 90959 (a_i, a'_{i+1}) unique pairs $UP2$.
3. Calculate term frequency (TF) of unique pairs $UP2$, for each document in the corpus. In our example corpus that meant for each of the 25 documents. We denote it as $DjTF(a_i, a'_{i+1}), j = 1.\overline{25}$.
4. $DjTF(a_i, a'_{i+1}) = k_j/h_j$, where h_j is the number of occurrences of the pair words in the document j. k_j is the number of unique pairs in document j.
5. $IDF(a_i, a'_{i+1}) = ln(n/m); n = 25$. m is the number of documents which include unique pairs (a_i, a'_{i+1}), in our example among 25 documents.
6. $W_{ij}(a_i, a'_{i+1}) = \frac{1}{25}\sum_{j=1}^{25} IDF(a_i, a'_{i+1}) * D_jTF(a_i, a'_{i+1})$
7. $W_{ij}(a_i, a'_{i+1})$ – weights of unique pairs .
8. We got 5% of the 90957 unique pairs, which $W_{ij}(a_i, a'_{i+1})$ is close to zero and declare them as stop words.

```
1. chop etildi(published)
2. har bir(each)
3. kitob jamgarmasi(book fund)
4. nima uchun(what for)
5. o'rta talim(secondary education)
6. men ham(me too)
7. bilan birga(along with)
8. yaxshi muqova(good cover)
9. oz vaqtida(It's on time)
10. ham bir(also a)
11. bir necha(a few)
12. barcha varaqlari(all sheets)
13. o'zi ham(himself)
14. bu yerda(here)
15. bo'lib qoldi(has become)
16. u ham(he too)
17. uchun ham(for both)
18. uning bu(its this)
19. butun darslikning(of the whole textbook)
20. yangi darslikning(new textbook)
.....................................
4529. Velosiped baxtiga(Luckily for the bike)
4530. Vodiy daralariga(To the gorges of the valley)
4531. Voqealarga aralashadi(Interferes with events)
4532. Xarakteri amallari(Character actions)
4533. Xarakterini izohlang(Explain the character)
4534. xonimning uylariga(to the lady's house)
4535. xoqonning hayoti(the life of a hawk)
4536. xotirasini abadiylashtirish(immortalize the memory)
4537. xudoyor davron(godly era)
4538. xushxabar ammo(The good news, however)
4539. yapon arab(Japanese arab)
4540. yasagan qayiqlarni(made boats)
4541. yasalgan fe'llar(made verbs)
4542. yaxshilar ahbob(good fellow)
4543. yig'isi alomatning(crying symptom)
4544. yig'lagan bolasini(crying baby)
4545. yig'och chog'liq(wood chips)
4546. yodlang islom(Remember Islam)
4547. yo'lakda bir(one in the hallway)
4548. yo'llardan biri(one of the ways)
```

Figure 1: Examples selected from the list of all stopwords generated by the bigram Algorithm 1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2022 doi:10.20944/preprints202204.0234.v1

the sake of simplicity, but the method can be abstracted to any length. The method should be used before the single stop word detection method.

To find collocation stop words, we use the following Algorithm 2:

Algorithm 2: The collocation method

1. Consider all occurrences of collocations in a corpus. In our case the total number of such collocations was 731155. Among them 489857 collocation words are unique collocation words.
2. $DjTF(a_i, a_{i+1}) = k_j/h_j$, where h_j is the number of occurrences of the pair words in the document j. k_j is the number of unique pairs in document j.
3. $IDF(a_i, a_{i+1}) = ln(n/m); n = 25$. m is the number of documents which include unique pairs, in our example among 25 documents.
4. $Wij(a_i, a_{i+1}) = \frac{1}{25}\sum_{j=1}^{25} IDF(a_i, a_{i+1}) * D_jTF(a_i, a_{i+1})$
5. $Wij(a_i, a_{i+1})$ – denotes weight of a collocation – $(a_i a_{i+1})$.
6. 5 % of all unique collocations had the weigh $W_{ij}(a_i, a_{i+1})$ close to zero and were declared as stop words.

The Algorithm 2 applied to the "School corpus" produced 24490 pairs of words as collocation stop words. A few examples are presented in Figure 2.

```
1. har bir(each)
2. nima uchun(what for)
3. bir kuni(one day)
4. o'rta talim(secondary education)
5. uchun darslik(textbook for)
6. chop etildi(published)
7. kitob jamg'armasi(book fund)
8. abad ham(forever)
9. abadiy kuchidan(from eternal power)
10. abadiy manziliga(to the eternal address)
11. abadiy muhrlanib(sealed forever)
12. abadligi hamda(   eternity and)
13. Abadul abad badnom(Abadul abad badnom)
14. Abadul abad turajakdur(It will last forever)
15. Abay singari(Abay suchlike)
16. Abbos degan(Abbos named)
17. Abbos qilichi(The sword of Abbas)
18. Abdulaziz qaytib(Abdulaziz returned)
19. Abdulazizga qaradi(He looked at Abdulaziz)
.....................................
24471. Odamlarni ko'rishadi(They see people)
24472. Odamlarning chehralari(Faces of people)
24473. Odamlarning haqiga(About people)
24474. Odamlarning kamligi(Lack of people)
24475. Odamlarning ko'zidan(From people's eyes)
24476. Odamlarning ko'zini(People's eyes)
24477. Odamlarning nomlarini(The names of the people)
24478. odamlarning og'irini(the weight of people)
24479. odamlarning qaysi(which of the people)
24480. odamlarning va(people and)
24481. odamlarning zilzila(earthquake of people)
24482. odamligi uni(humanity him)
24483. odamligini ham(that he is human)
24484. odamligini ta'minlab(providing humanity)
24485. odamlik qiyofasini(human image)
24486. odamman deb(that I am human)
24487. odamman deganini(I mean man)
24488. odamma saxir(I'm sorry)
24489. odamni ajdodlari(man's ancestors)
24490. odamni ona(mother of man)
```

Figure 2: Examples selected from the list of all stopwords generated by the collocation Algorithm 2.

3.4 Single word (stop word) detection algorithm

In this section we consider the single word stop words detecting algorithm based on TFIDT(Term frequency and inverse document frequency) of the word. To find single word stop words, we use the following Algorithm 3:

The Algorithm 3 applied to the "School corpus" produced 2358 stop words. A few examples are presented in Figure 3.

Algorithm 3: Single word (stop word) detection algorithm

1. $DjTF(a_i) = k_j/h_j$, where h_j is the number of occurrences of the pair words in the document j. k_j is the number of unique pairs in document j.
2. $IDF(a_i) = ln(n/m); n = 25$. m is the number of documents which include unique pairs, in our example among 25 documents.
3. $W_{ij}(a_i) = \frac{1}{25}\sum_{j=1}^{25} IDF(a_i) * D_jTF(a_i)$
4. $W_{ij}(a_i)$ – denotes weight of a word $(a_i$.
5. 5 % of of the 47165 unique words, which $W_{ij}(a_i)$ was close to zero and declared stop words.

3.5 The final stop word detection Algorithm for Uzbek language

In this section we consider the main algorithm of detecting Stop words of text in Uzbek language. We bring this algorithm as in the scheme presented on Figure 4.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2022 doi:10.20944/preprints202204.0234.v1

Algorithm 4: find and remove Uzbek stop words from text (Corpus)

```
Input(Corpus)
Corpus ←Tokenize(Corpus)
Dictionary ←
 Extract_From_Dictionary(pronoun, modal
 verb, particle, part of a rhyme,
 conjunctions, introductory words,
 adverbs, auxiliary words)
;          // Procedure Check(Corpus)
i←1
while i < len(Corpus) do
    if Corpus(i)∈Dictionary then
        Corpus ← Corpus − Corpus(i)
    i←i+1
/* Procedure Collocation_Two_Words
   (Corpus)                          */
Corpus ←Tokenize(Corpus)
i←1
while i < len(Corpus) do
    S(i)← token(i)+token(i+1); i←i+1
/* Procedure IDF()                   */
IDF(S(i))←ln(N/n) ;     // N-number of
 all documents; n- number of
 documents, which include S(i)
/* Procedure TFIDF()                 */
j←1
while j < len(Corpus) do
    TF(j)←0
    i←j; while i < len(Corpus) - 1 do
        if S(j)==S(i) then
            TF(j) ← TF(j) + 1
        i←i+1;
    TFIDF(j)←TF(j)*IDF(S(j) if
    TFIDF(j) close to zero then
        Dictionary(j)←S(j);
        i←1; while i < len(Corpus) do
            if Dictionary(j)==Corpus(i)
            then
                Corpus←Corpus−Dictionary(i);
            i←i+1
    j←j+1
```

```
1. Abdulla(Abdulla)
2. aka(brother)
3. asosida(based on)
4. ayt(say)
5. aytib(telling)
6. aziz(dear)
7. baho(evaluation)
8. bahor(spring)
9. baland(high)
10. beradi(will give)
11. berdi(gave)
12. berib(giving)
13. berilgan(given)
14. bering(read)
15. bichimi(physique)
16. bilan(with)
17. bilib(know)
18. bilim(knowledge)
19. biri(one)
20. birinchi(first)
.....................
2339. aylamakka(to turn)
2340. aylasak(if we do)
2341. aytishuvlarda(in arguments)
2342. bachalar(children)
2343. bacha (babbie)
2344. badiiyatni(art)
2345. bag'ayri(past)
2346. bag'rimdami(in my heart)
2347. baid(height)
2348. balladalar(ballads)
2349. banddin(occupied)
2350. Bandksoy (Bandkushoy)
2351. barchalarining(all of them)
2352. barglarga(to the leaves)
2353. bastai(composer)
2354. baxilga(stingy)
2355. baxtdan(happily)
2356. baytallarga(beetles)
2357. bazmni(party)
2358. begonani(outsider)
```

Figure 3: Examples selected from the list of all stopwords generated by the single word extraction Algorithm 3.

Table 1: Number of stop words created by each presented algorithm.

Algorithm	Number of stop words
Bigram	4548
Collocation	24490
Single word	2358

4 Results

The first phase of the project consisted of creating a solid base for corpus linguistics as there were o readily available corpora for Uzbek language. A corpus named "School corpus" was created with 731156 running words. The algorithms for stop words detection ere applied to the aforementioned corpus and Table 1

The values are also presented in Figure 5.

5 Data availability

The presented automatically extracted lists (a list for each described method) are freely available at Zenodo repository [21]:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2022 doi:10.20944/preprints202204.0234.v1

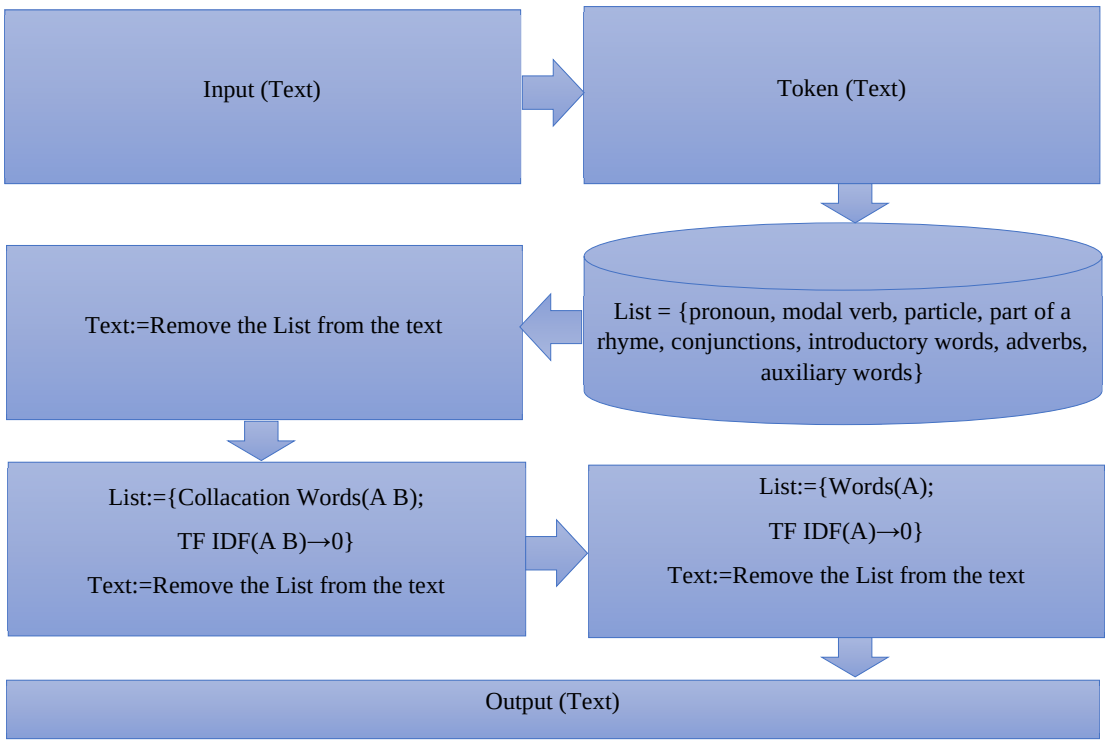

Figure 4: Scheme of the whole process.

https://doi.org/10.5281/zenodo.6319953

6 Conclusion and further work

The article presents the first attempt at the automatic detection of stop words for Uzbek language. A corpus named "School corpus" was created for this purpose, it contains 25 documents and 731155 running words, of which 47165 are unique words. Three methods were applied to the corpus in order to extract (or detect) stop words: a method that extracts single word stop words and two methods that aim at pairs of words, a bigram and collocation method. Each method is described and presented in a form of an algorithm. The methods can be used in a series and the results can be added together to form the final list of stop words.

Taking account the conception of stop words depending on the text every word can be stop words. According to this approach (based on TFIDF). A quick comparison of the methods shows an increase in stop words detection using the collocation method

Only a two-word collocation is considered in this article and the motivation behind this is that three or more word collocations that act as stop words are not that common, but we still believe that a further work needs to be done in this direction. The proposed methodology does not change for longer collocations.

References

[1] S. Matlatipov, X. Madatov, G. Matlatipov, A. O'razbayev, M. Raximboyev, I. Avezmatov, U. Babajanov, L. Kurbanova, D. Xujamov, and D. Matjumayeva, ""o'zbek tilining statistik electron lug'at" exm dasturi uchun guvohnoma," *Intellektual mulk agentligi*, 2020.

[2] A. W. Pradana and M. Hayaty, "The effect of stemming and removal of stop words on the accuracy of sentiment analysis on indonesian-language texts," *Game Technology, Information System, Computer Network, Computing, Electronics, and Control Journal*, vol. 4, no. 3, pp. 277–288, 2019.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2022 doi:10.20944/preprints202204.0234.v1

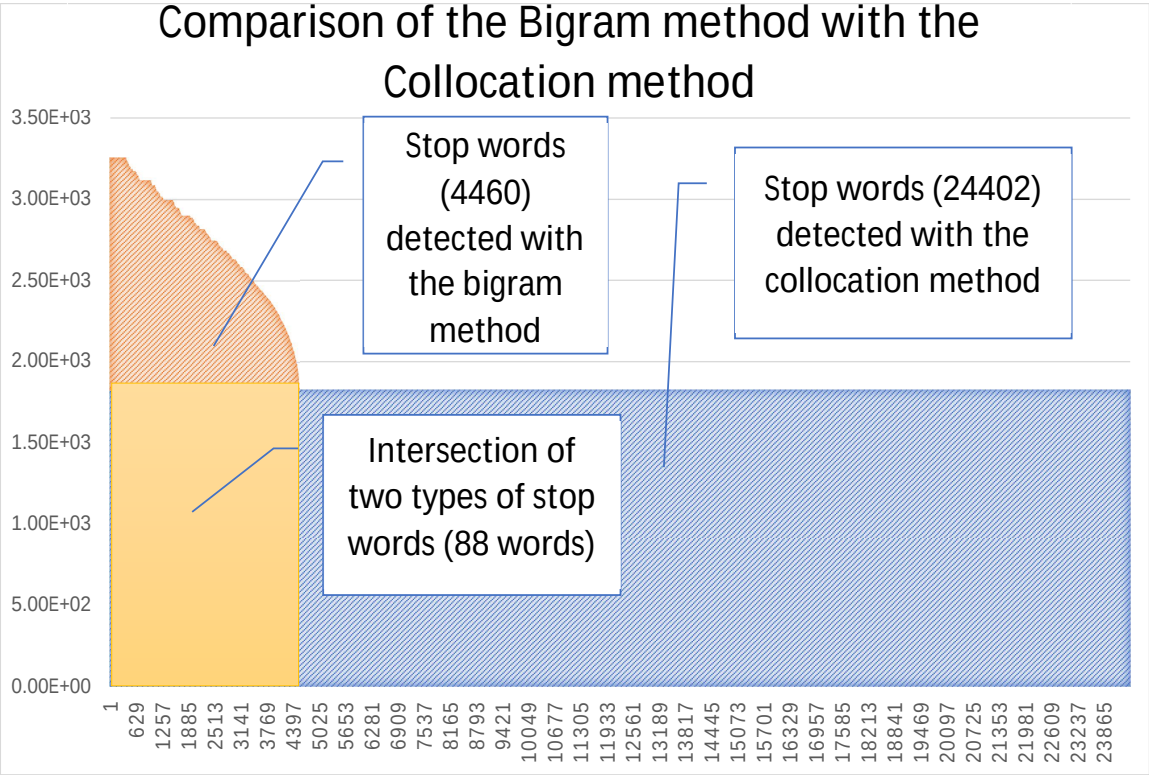

Figure 5: Number of stop words for each algorithm applied to the "School corpus".

[3] R. U. Haque, P. Mehera, M. F. Mridha, and M. A. Hamid, "A complete bengali stop word detection mechanism," in *Conference Paper · May 2019*. Conference, 2019.

[4] R. Rania and D.K.Lobiyal, "Automatic construction of generic stop words list for hindi text," in *International Conference on Computational Intelligence and Data Science*, vol. 132, International Conference on Computational Intelligence and Data Science. ICCIDS 2018, 2018, pp. 362–370.

[5] P. J. Burns, "Constructing stoplists for historical languages," *Digital Classics Online*, vol. 4, no. 2, 2018.

[6] R. M. Rakholia and J. R. Saini, "A rule-based approach to identify stop words for gujarati language," in *In Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications*, 2017, pp. 797–806.

[7] J. K. Raulji and J. R. Saini, "Generating stopword list for sanskrit language," in *In: 2017 IEEE 7th International Advance Computing Conference*. IEEE 7th, 2017, pp. 799–802.

[8] O. D. Tijani, A. T. Akinwale, S. A. Onashoga, and E. O. Adeleke, "An auto-generated approach of stop words using aggregated analysis," in *In: Proceedings of the 13th International Conference of the Nigeria Computer Society*, 2017, pp. 99–115.

[9] M. Mhatre, D. Phondekar, P. Kadam, A. Chawathe, and K. Ghag, "Dimensionality reduction for sentiment analysis using pre-processing techniques," in *In Proceedings of the IEEE 2017 International Conference on Computing Methodologies and Communication*. ICCMC, 2017, pp. 16–21. [Online]. Available: https://doi.org/10.1109/ICCMC.2017.8282676

[10] C. Sammut and G. I. Webb, Eds., *TF–IDF*. Boston, MA: Springer US, 2010, pp. 986–987. [Online]. Available: https://doi.org/10.1007/978-0-387-30164-8_832

[11] Y. Wang, K. Kim, B. Lee, and H. Y. Youn, "Word clustering based on pos feature

for efficient twitter sentiment analysis," *Human-centric Comput*, vol. 8, no. 17, pp. 1–25, 2019. [Online]. Available: https://doi.org/10.1186/s13673-018-0140-y

[12] N. Ousirimaneechai and S. Sinthupinyo, "Extraction of trend keywords and stop words from thai facebook pages using character n-grams," *International Journal of Machine Learning and Computing*, vol. 8, no. 6, 2018.

[13] C. Slamet, A. R. Atmadja, D. S. Maylawati, R. S. Lestari, W. Dharmalaksana, and M. A. Ramdhani, "Automated text summarization for indonesian article using vector space model model," in *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 288, no. 1, Conference. IOP, 2018. [Online]. Available: https://doi.org/10.1088/1757-899X/288/1/012037

[14] G. Li and J. Li, "Research on sentiment classification for tang poetry based on tf-idf and fp-growth," in *Proceedings of 2018 IEEE 3rd Advanced Information Technology*, Electronic and Automation Control Conference. IAEAC, 2018, pp. 630–634. [Online]. Available: https://doi.org/10.1109/IAEAC.2018.8577715

[15] H. M. Zin, N. Mustapha, M. A. A. Murad, and N. M. Sharef, "The effects of pre-processing strategies in sentiment analysis of online movie reviews," in *AIP Conf. Proc.*, vol. 1891, no. 1. AIP Conf., 2017, pp. 1–7. [Online]. Available: https://doi.org/10.1063/1.5005422

[16] S. K. Metin and B. Karaog'lan, "Stop word detection as a binary classification problem," *Anadolu University Journal of Science and Technology A- Applied Sciences and Engineering*, vol. 18, no. 2, pp. 346–359, 2017.

[17] J. K. Raulji and J. R. Saini, "Generating stop word list for sanskrit language," in *In Advance Computing Conference IEEE 7th International*. IEEE, 2017, pp. 799–802.

[18] S. J. R. Rakholia R. M., "A rule-based approach to identify stop words for gujarati language," in *Suresh Chandra Satapathy Vikrant Bhateja Siba K.*, 2017.

[19] R. M. Rakholia and J. R. Saini, "Information retrieval for gujarati language using cosine similarity based vector space model," in *Theory and Applications*. Springer_Singapore, 2017, pp. 1–9.

[20] X.Madatov and S. Matlatipov, "Kosinus o'xshahshlik va uning o'zbek tili matnlariga tatbiqi haqida," *O'zMU xabarlari*, vol. 2, no. 1, 2016.

[21] K. Madatov, S. Bekchanov, and J. Vičič, "Lists of uzbek stopwords (1.1) [data set]," Zenodo.