# СТВАРЭННЕ СПІСУ СТОП-СЛОЎ НА АСНОВЕ КОРПУСА ТЭКСТАЎ

*А.А.Русакевіч (Мінск, МГЛУ)*

Дадзеная навуковая работа апісвае тое, як быў створаны першы ў гісторыі камп'ютарнай лінгвістыкі беларускай мовы спіс стоп-слоў. Яна можа быць выкарыстана для стварэння аналагічных спісаў для іншых моў, а яе вынік – спіс стоп-слоў беларускай мовы – у камп'ютарнай лінгвістыцы і SEO-прасоўванні сайтаў.

**Ключавыя словы:** *камп'ютарная лінгвістыка; беларуская мова; стоп-слова; SEO.*

# CREATING A STOPWORD LIST BASED ON A CORPUS OF TEXTS

*A.A.Rusakevich (Minsk, MSLU)*

This scientific paper describes how the first stop-word list in the history of computational linguistics of the Belarusian language was created. The paper can be used to create similar lists for other languages, and its result – the list of stop-words of the Belarusian language – can be used in computational linguistics and SEO website promotion.

**Key words**: *computational linguistics; Belarusian language; stop-word; SEO.*

У наш час няма амаль нікога, хто бы не прызнаў: наш век належыць інфармацыйным тэхналогіям. І сярод гэтых тэхналогій асабліва вылучаюцца пошукавыя рухавікі і ШІ, якія ўжо паспелі заняць значную частку нашага штодзённага жыцця.

Стварэнне і праца вышэй упамянутых рэчаў немагчыма без камп'ютарнай лінгвістыкі, сітуацыя з якой у адносінах да беларускай мовы даволі складаная. Для беларускай мовы існуе Беларускі N-корпус з вялікай і дакладнай граматычнай базай дадзеных. Існуе таксама Эксперыментальны корпус беларускай мовы (далей – ЭКБМ) з вялікім мноствам размечаных тэкстаў на беларускай мове. І ён ідзе разам з лематызатарам і POS-тэгерам, але яны напісаны на мове Perl, якая на дадзены момант не з'яўляецца актуальнай.

Гэта праца апісвае стварэнне спісу стоп-слоў для беларускай мовы як частку маштабнага праекта па стварэнню адсутных праграм і бібліятэк, здольных забяспечыць магчымасць аналізаваць тэксты на Python, каб у перспектыве палепшыць якасць пошукавых рухавікоў і штучнага інтэлекту, што працуюць з беларускай мовай.

Стоп-словы (або шумавыя словы) – гэта словы, без якіх сэнс тэксту ў цэлым захаваецца і якія сустракаюцца вельмі часта сярод розных тэкстаў [2]. Гэта займеннікі, прыназоўнікі, паразіты часу і г.д. Убраць гэтыя словы - адна з частак працы па нармалізацыі тэксту, каб зрабіць яго годным для апрацоўкі камп'ютарам, а таксама частка працы па аптымізацыі тэкстаў для пошукавых рухавікоў SEO-спецыялістамі [3].

Для стварэння спісу шумавых слоў былі скарыстаны бібліятэкі numpy, scipy, з дапамогай якіх быў прааналізаваны ЭКБМ.

Алгарытм выбару стоп-слоў засноўваецца на індэксе TF-IDF, які дае магчымасць вызначыць частату адносна аднаго тэксту і долю адносна ўсіх іншых тэкстаў. TF-IDF можна вылічыць з дапамогай формулы ніжэй [4]:

$$tfidf_{x,y} = tf_{x,y} * \log (N/df_x)$$

Дзе $tf_{x,y}$ – колькасць з'яўленняў x у тэксце y, $df_x$ – гэта колькасць тэкстаў, дзе сустрэлася слова x, а N – агульная колькасць дакументаў.

Гэта формула была рэалізавана ў Scipy у выглядзе TfidfCounter. Дадзеныя туды загружаюцца з ЭКБМ у такой паслядоўнасці:

1. Бярэцца адзін файл корпуса. З яго чытаюцца паслядоўна словы, прапускаюцца знакі прыпынку і ўнутраныя тэгі корпуса. Словы, якія датычацца аднаго і таго ж тэксту, аб'ядноўваюцца праз прабел у адзін радок, які загружаецца ў TfidfCounter.
2. Дадзеныя ў TfidfCounter пераўтвараюцца ў матрыцу, дзе кожны радок замацаваны за тэкстам, стоўб – за словам, а ячэйка ўтрымлівае індэкс TF-IDF.
3. З дапамогай сродкаў матрычнай матэматыкі TF-IDF усіх радкоў складваецца.
4. Атрымліваем суму ўсіх TF-IDF.
5. Кожны індэкс дзеліцца на агульны лік. Усе шумавыя словы, што займаюць менш 1‰, адкідваюцца.
6. Стоп-словы без індэксаў запісваюцца файл праз знак новай лініі.

Усяго атрымліваецца спіс з амаль 250 шумавых слоў, які па саставу і колькасці элементаў падобны на аналагічныя спісы буйных моў. Таксама былі створаны і большыя спісы шляхам змянення парога долі ў 5 пункце [1].

Дадзеная навуковая работа будзе карыснай для тых, хто вывучае камп'ютарную лінгвістыку і жадае стварыць спіс стоп-слоў для іншых моў; вынік працы ж – першы ў гісторыі беларускай мовы спіс шумавых слоў – дапаможа прафесіянальным SEO, якім патрэбна аптымізаваць апісанне і загаловак сайту, а сталым камп'ютарным лінгвістам – лепш апрацоўваць тэксты на беларускай мове.

СПІС ЛІТАРАТУРЫ


1. Русакевіч, А.А. Спісы стоп-слоў для беларускай мовы і іх генератар // Github.com [Электронны рэсурс]. – 2025. Рэжым доступу : https://github.com/alex-rusakevich/stopwords-be/. – Дата доступу : 20.01.2025.
2. Скороходов, И.С. Основные этапы обработки текста и генерации признаков в задачах текстовой классификации // Проблемы Науки. 2016. №15 (57). Режим доступа : https://cyberleninka.ru/article/ n/osnovnye-etapy-obrabotki-teksta-i-generatsii-priznakov-v-zadachah-tekstovoy-klassifikatsii. – Дата доступа : 20.01.2025.
3. Стоп-слова в текстах — роль в SEO и как от них избавиться? // work24.ru [Электронный ресурс]. – 2023. Режим доступа : https://work24.ru/spravochnik/povyshenie-kvalifikacii/teksty/kopirajting/stop-slova-v-tekstah-rol-v-seo-i-kak-ot-nih-izbavitsya. – Дата доступа : 20.01.2025.
4. Introduction to Natural Language Processing with PyTorch // medium.com [Electronic resource]. – 2023. Mode of access : https://medium.com/@thevnotebook/introduction-to-natural-language-processing-with-pytorch-2-5-538d7b167b34. – Date of access : 20.01.2025.