Математичні моделі й методи в системах автоматизованого управління, проектування та наукових дослідженнях

УДК 004.047

Т.В. Голуб,
М.Ю. Тягунова, канд. техн. наук, доц.
Запорізький національний технічний університет, м. Запоріжжя, Україна
tany-golub@mail.ual

Метод стемінгу україномовних текстів для класифікації документів на базі алгоритму Портера

У статті запропоновано метод стемінгу для україномовних текстів на базі модифікованого алгоритму стемінгу Портера, при використанні якого підвищується якість попередньої обробки текстових документів, поданих українською мовою, для подальшої їх класифікації.

Ключові слова: стемінг, алгоритм Портера, україномовний текст, класифікація.

Вступ

Класифікація текстових документів проводиться для полегшення процесу пошуку необхідної інформації у великому об'ємі накопичених матеріалів, представлених різними мовами світу. У літературі наведено велику кількість реалізацій класифікаторів текстових документів, але, не зважаючи на свою універсальність, вони не вирішують усі існуючі проблеми. Так, для нашого регіону найбільш розповсюдженими мовами є англійська, російська та українська. Наведені в літературі автоматичні класифікатори охоплюють здебільшого матеріали, надані англійською та російською мовами. При цьому україномовні тексти залишаються осторонь.

Процес класифікації текстових документів передбачає проходження декількох етапів. Першим етапом є попередня обробка текстів, тобто перетворення текстових документів у форму, придатну для подальшої комп'ютерної обробки – вектору даних, подальша обробка отриманого вектору з метою його зменшення шляхом відсікання суфіксів і закінчень слів (стемінг), виключення неінформативних слів. Другим етапом є безпосередньо класифікація документу у відповідну категорію [1].

У даній роботі з метою подальшої класифікації пропонується метод реалізації одного із ключових етапів попередньої обробки текстів, а саме процесу стемінгу, спрямованого на аналіз україномовних текстів.

Стемінг слів

Під стемінгом розуміють виділення основи слова, яка буде незмінною для ряду форм незалежно від відмінку та наявних суфіксів. При цьому отримана форма слова не обов'язково відповідатиме його кореню.

Через існування значних відмінностей у процесі словотворення навіть споріднених та близьких мов, у залежності від мови тексту, що аналізується, необхідно застосовувати різні правила стемінгу. З цієї причини неможливо розробити універсальний перелік правил, придатних для будь-якої мови. Тому в даній статті пропонується реалізація стемінгу україномовного тексту.

У літературі наведено велику кількість алгоритмів стемінгу. Але вони найчастіше передбачають обробку текстів, наведених на розповсюджених у світі мовах, до яких українська не належить.

На даний момент в літературі запропоновано стемери, що спираються на алгоритми з використанням підходу опису правил відсікання слів (стемер Портера [2], Paice/Husk

стемер [3]), а також ті, що спираються на лінгвістичні дані та використовують семантичну інформацію слова (Y -стеммер [4]), або використовують статистичну інформацію (анализ К-грамм [5], Stemka [6], MyStem [7]).

Для аналізу текстів наведених на російській та українській мовах розроблено такі алгоритми, як Stemka, MyStem та стемер Портера.

При реалізації алгоритму Stemka формується набір потенційних закінчень з урахуванням умови їх використання та представляється у вигляді таблиці переходів. Алгоритм MyStem виконує порівняння усіченого слова з існуючими у словнику. Тобто для обох реалізацій необхідне створення деякої опорної бази, що потребує додаткових вимог до об'ємів пам'яті.

Стемер Портера (Snowball) побудовано за принципом попереднього описання правил відсікання суфіксів та закінчень, тому процес стемінгу

не вимагає наявності додаткової бази для формування критеріїв усічення слів.

Таким чином, алгоритм Портера, на відміну від двох інших алгоритмів, не потребує генерування баз основ, що зменшує вимоги до необхідного обсягу пам'яті для роботи програми, що надає можливість створити автономну програму невеликого обсягу. Проблеми, які виникають при роботі даного стемера (сильне або слабке усічення слів), не є критичними для вирішення задачі класифікації текстових документів.

Тому далі пропонується для розгляду метод стемінгу для україномовних текстів на основі алгоритму Портера.

У літературі існує велика кількість реалізацій стемера Портера [8], призначеного для обробки текстів, наданих на різних мовах світу. Та для української мови жодна реалізація не передбачалась через фонетичну та орфографічну близькість її до російської. При цьому існує суттєва сукупність відмінностей між мовами, що розглядаються, яка обмежує можливість їх взаємного використання. Так, спостерігається значна різниця у формах суфіксів та закінчень для цих мов, подвоєнні приголосних, наявності специфічних літер, характерних для української та російської мов. Приклади таких відмінностей наведені в таблиці 1. Їх сукупність не дозволяє якісно використовувати стемер, розроблений для російськомовних документів [2], для попередньої обробки україномовних текстів.

Стемер для україномрвних слів

Стемер Портера для української мови, що пропонується, має кілька особливостей: виключення із слова символу апострофа та рівнозначність букв «г» та «ґ». В результаті аналізу правил побудови слів української мови було отримано перелік закінчень, характерних для іменників, прикметників, дієслів, дієприкметників, дієприслівників, виділено форми суфіксів для найвищого ступеня порівняння прикметника та зазначено подвоєння приголосних. Форми словотворення, притаманні дієприкметникам, аналізуються після перевірки слів на наявність закінчень прикметника. Відсікання відбувається за найбільшим співпадаючим суфіксом і закінченням та виконується над частиною слова після першого та другого сполучення «голосна-приголосна» для виключення перетворення слів, що складаються з одного складу.

З урахуванням отриманих даних пропонується наступна послідовність дій при реалізації стемера Портера для української мови, представлена блок-схемою алгоритму на рисунку 1.

На першому етапі виконується перевірка наявності апострофа та букви «ґ» в слові, що аналізується. У разі їх знаходження апостроф видаляється, а буква «ґ» замінюється на «г». Далі визначається область слів після першого та другого сполучення «голосна-приголосна». Якщо слово містить лише одну голосну літеру, воно не підлягає усіканню суфіксів. На наступному етапі виконується пошук та видалення закінчень дієприслівника. Якщо закінчення видаляється, етап закінчується. В іншому випадку перевіряється наявність закінчень дієслів на –ся, –ться або іменників на –ість, –іст.

Таблиця 1. Деякі відмінності словотворення російської та української мов

Критерії	**Російська мова**	**Українська мова**
Голосні літери	а, е, и, о, у, ы, э, ю, я	а, е, є, и, і, ї, о, у, ю, я
Специфічні символи	ё	ґ, «'»
Подвоєння приголосних	н	д, т, з, с, л, н, ж, ш, ц, ч
Закінчення дієприслівників	Одна з груп закінчень: 1. *–ив, -ивши, -ившись, -ыв, -ывши, -ывшись*; 2. *–в, -вши, -вшись* после *-а-, -я-;*	Одна з груп закінчень: 1. *–учи, -ючи, -ачи, -ячи;* 2. *–ши, -вши;*
Окончания существительных	*-а, -ев, -ов, -ие, -ье, -е, -иями, -ями, -ами, -еи, -ии, -и,-ией, -ей, -ой, -ий, -й, -иям, -ям, -ием, -ем, -ам, -ом, -о, -у, -ах, -иях, -ях, -ы, -ь, -ию, -ью, -ю, -ия, -ья, -я;*	Одна з груп закінчень: 1. *–ею, -єю, -їв, -ів, -ей, -єй, -ою, -ові, -ом;* 2. *–м, -ми, -х, -ти, -ті, -та, -тами* після *–а-, -я-;* 3. *–ві, -м, -ні, -нем, -нами* після *–е-, -є-.*

Далі виконується один із варіантів: пошук і видалення закінчень прикметника, за ним дієприкметника; або пошук і видалення закінчень дієслова; або пошук і видалення закінчень іменника. Наступний етап передбачає видалення голосної літери наприкінці отриманого слова. І на останньому етапі проводиться перевірка на наявність суфіксів найвищого ступеня порівняння прикметника, подвоєних приголосних, наявності букви «ь» з подальшим видаленням суфіксу, «ь» та однієї з подвоєних приголосних.

В результаті виконання даного алгоритму отримується форма заданого слова з відсіченими суфіксом та закінченням, спрощенням слова до форми без подвоєнь приголосних і наближеної до кореня.

Експериментальна реалізація методу

З метою реалізації запропонованого методу та проведення експериментів було розроблено

програму, що реалізує стемінг Портера для україномовних текстів, на основі її аналогу для російської мови. Було враховано всі зазначені відмінності словотворення. Для створення програмної реалізації використовувалася мова програмування Python, а також окремі функції спеціалізованої надбудови – модуля NLTK.

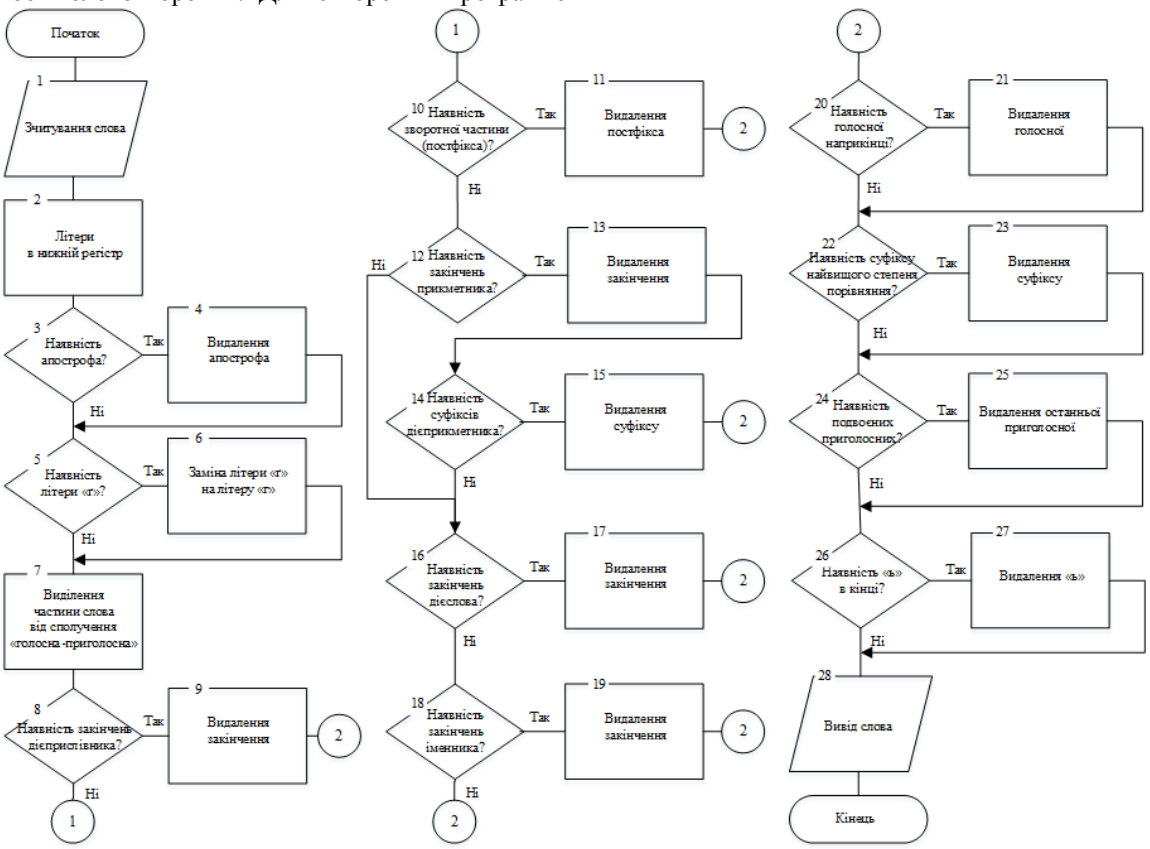

Рисунок 1 – Модифікація алгоритму стемінга Портера для української мови

У якості експериментального матеріалу використовувалися прості та складні слова різного відмінку, а також уривки тексту.

Робота реалізованого програмно запропонованого методу, що базується на запропонованому модифікованому алгоритмі Портера, була перевірена на прикладі аналізу різних форм окремого слова та частини тексту.

В таблиці 2 наведено отриманий результат аналізу окремого слова (на прикладі різних відмінків і форм слів «привласнюваний» та «суттєвий»), в таблиці 3 – аналіз блоку вихідного тексту з виділеними відсіченими частинами слів.

Таблиця 2. Приклад роботи запропонованого алгоритму для окремого слова

Слово	**Стемінг**	**Слово**	**Стемінг**
привласн*ю*ваний	привласн	суттєв*ий*	суттєв
привласн*ю*ваного		суттєв*ість*	
привласн*ю*ваною		суттєв*ого*	
привласн*ю*ванім		суттєв*а*	
привласн*ю*ваній		суттєв*ою*	
привласн*ю*ваний		суттєв*ими*	
привласн*ю*ваних		суттєв*ого*	

Таблиця 3. Приклад роботи запропонованого алгоритму для блоку тексту

Вихідний текст	Результуючий вектор
Стемінг - це процес скорочен*ня* слов*а* до основ*и* шлях*ом* відкидан*ня* допоміжн*их* частин, так*их* як закінчен*ня* чи суфікс. Результат*и* стемінг*у* іноді дуже схож*і* на визначен*ня* кореня слов*а*, ал*е* його алгоритм*и* баз*уються* на інш*их* принцип*ах*. Том*у* слов*о* після обробк*и* алгоритм*ом* стемінг*у* мож*е* відрізн*ятися* від морфологічн*ого* кореня слов*а*.	стемінг це процес скорочен слов до основ шлях відкидан допоміжн частин так як закінчен чи суфікс результ стемінг інод дуж схож на визначен корен слов ал йог алгоритм баз на інш принцип том слов післ обробк алгоритм стемінг мож відрізн від морфологічн корен слов

Часові витрати на роботу алгоритму наведені на рисунку 2.

З рисунку 2 видно, що для аналізу тексту, що має розмір більше 17 тис. символів, алгоритм

витратив менше 0,1 секунди, тобто для тексту зазначеної величини є несуттєвим.

Результати експериментів, які наведено в таблицях, показують, що запропонований метод стемінгу, що базується на описаній модифікації алгоритму Потрета, є працездатним.

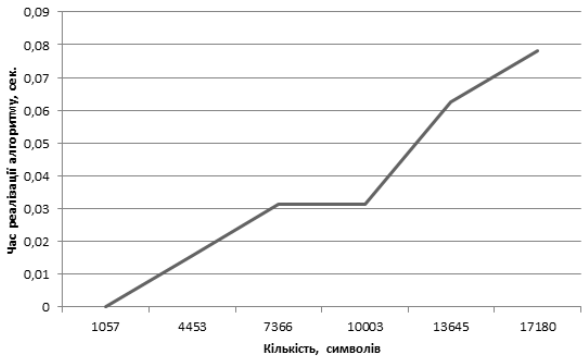

Рисунок 2 – Часові витрати на роботу запропонованого алгоритму

Аналіз тексту власноруч та за допомогою програмних засобів показав, що отримані результати в більшості випадків співпадають на 100%, тобто наведена реалізація є цілком працездатною.

Це дозволяє говорити про подальшу можливість використання запропонованого методу для класифікації україномовного тексту на етапі попередньої підготовки.

Висновки

Даний варіант методу стемінгу виконує відсікання суфіксів та закінчень слів україномовних текстів. Це дає можливість спростити вектор даних зазначених текстів шляхом виключення повторів однакових за змістом слів, зумовлених різними формами та відмінками їх використання. У результаті зменшуються вимогти до об'єму пам'яті, необхідної для реалізації класифікатора.

Наукова новизна роботи полягає у виділенні характеристичних особливостей словотворення в україномовних текстах та формуванні методу стемінгувказаних текстів.

Практична значимість запропонованого методу полягає у можливості стемінгу україномовних текстів. Його програмна реалізація не потребує великого об'єму пам'яті на диску.

Список літератури

1. Golub T., The Analysis of Text Documents Classifiers Constructing Methods / T. Golub // XIII International conference: Modern Problems Of Radio Engineering, Telecommunications, And Computer Science 2016, 23-26 February 2016. : тези доп., – Lviv-Slavsko, Ukraine, 2016. – P.742-745.

2. Porter M.F. An algorithm for suffix stripping / M.F. Porter // Program, 2006, - Vol 40 Iss: 3- P .211 - 218

3. Moral Cristian, A survey of stemming algorithms in information retrieval / Cristian Moral, Angélica de Antonio, Ricardo Imbert and Jaime Ramírez // Information research - vol. 19 no. 1, – 2014, p. 605-625.

4. Yatsko Viatcheslav, Y-stemmer / Yatsko Viatcheslav – Yatsko's Computational Linguistics Laboratory [Электронный ресурс], - Режим доступа: http://yatsko.zohosites.com/y-stemmer.html

5. Lama Prabin, Clustering system based on text mining using the k-means algorithm / Prabin Lama – Bachelor's thesis (UAS) of Information Technology "Text Mining and Clustering 2013", [Электронный ресурс], - Режим доступа: https://www.theseus.fi/bitstream/handle/10024/69505/Lama_Prabin.pdf?sequence=1

6. Коваленко А. «Стеммер». Морфологический анализ для небольших поисковых систем / А. Коваленко // Системный администратор, 2002, - Выпуск №1 (1) - [Электронный ресурс] Режим доступа: http://samag.ru/archive/article/47

7. Segalovich Ilya. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. / Ilya Segalovich // Conference: Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications. MLMTA'03, 23 – 26 June 2003, Las Vegas, Nevada, USA, 2003. – P. 721-729.

8. Пересторонин П. Стеммер Портера для русского языка/ Павел Пересторонин // 2014 [Электронный ресурс] Режим доступа: https://eigenein.xyz/snowball/

Надійшла до редакції 15.03.2014

Т.В. ГОЛУБ, М.Ю. ТЯГУНОВА
Запорожский национальный технический университет (Украина)

МЕТОД СТЕММИНГА УКРАИНОЯЗЫЧНЫХ ТЕКСТОВ ДЛЯ КЛАССИФИКАЦИИ ДОКУМЕНТОВ НА ОСНОВЕ АЛГОРИТМА ПОРТЕРА

В статье предложен метод стеммига для украиноязычных текстов на основе модифицированного алгоритма стемминга Портера, при использовании которого повышается качество предварительной обработки текстовых документов, предоставленных на украинском языке, с целью дальнейшей их классификации.

Ключевые слова: стемминг, алгоритм Портера, украиноязычный текст, классификация.

T.V. GOLUB, M.YU. TYAGUNOVA
Zaporizhzhya National Technical University (Ukraine)

STEMMING METHOD OF UKRAINIAN LANGUAGE TEXT FOR DOCUMENT CLASSIFICATION BASED ON PORTER ALGORITHM

The Ukrainian texts stemming method for documents classification based on Porter algorithm has been proposed. Using the proposed method the quality of text documents preliminary processing for the purpose of further classification can be significantly increased.

The stemming process is one of the key stages of preliminary text processing. It is intended to convert the document into a form, applicable for further computer processing to conduct the classification.

The Porter stemmer is built on a principle of suffixes and endings cutting off description rules. The additional basis for words truncation criterions generation is not required here. This fact diminishes the required memory volume needed for program to function. It provides the ability to create an autonomous moderate-sized software.

The proposed Porter stemmer for the Ukrainian language has the following peculiarities: the apostrophe symbol is being excluded from the words; the "г" and "ґ" letters are being interpreted as equal. As a result of Ukrainian words building rules analysis conducted the list of specific endings, typical for different speech parts and word forms, including the consonants duplication indication, has been obtained. The truncation is taking place for the biggest coincident suffix and ending and is carried out on a word part after the first and second "vowel-consonant" conjunction.

Taking into consideration the obtained results, the following sequence of actions for Porter stemmer for the Ukrainian language implementation has been proposed: 1. if the apostrophe is taking place, it has to be removed; 2. if the "ґ" letter is taking place, it has to be substituted with "г" letter; 3. the number of words between the first and the second "vowel-consonant" conjunctions is being calculated; 4. depending on a part of speech, both parts of certain word – suffix and ending – are being deleted; 5. if the resulting word ends with a vowel, then the latest should be deleted; 6. then the removal of duplicating consonants and "ь" should be carried out.

The proposed variant of algorithm allows conducting the truncation of Ukrainian texts words suffixes and endings. It allows simplifying the data vector, formed from such texts, by excluding the duplication of words, stipulated by different forms of their usage. As a result, the memory requirements for classifier implementation can be reduced.

Key words: stemming, Porter algorithm, Ukrainian-text, classification.

References

1. Golub T. (2016), *The Analysis of Text Documents Classifiers Constructing Methods*, XIII International conference: Modern Problems Of Radio Engineering, Telecommunications, And Computer Science 2016, Lviv-Slavsko, Ukraine, pp.742-745.
2. Porter M.F. (2006), *An algorithm for suffix stripping*, Program, Vol 40 Iss: 3, pp .211 - 218
3. Moral Cristian, Angélica de Antonio, Ricardo Imbert and Jaime Ramírez (2014), *A survey of stemming algorithms in information retrieval*, Information research,Vol. 19, No. 1, pp. 605-625.
4. Yatsko Viatcheslav (2016), *Y-stemmer*, Yatsko's Computational Linguistics Laboratory, available at: Access mode: http://yatsko.zohosites.com/y-stemmer.html
5. Lama Prabin (2013), *Clustering system based on text mining using the k-means algorithm*, Bachelor's thesis (UAS) of Information Technology "Text Mining and Clustering 2013", available at: https://www.theseus.fi/bitstream/handle/10024/69505/Lama_Prabin.pdf?sequence=1
6. Kovalenko A., (2002), *"The Stemmer." Morphological analysis for small search engines* [«Stemmer». Morfologicheskiy analiz dlya nebolschih poiskovih system], System Administrator, [Sistemniy administrator], Vol №1 (1), available at: http://samag.ru/archive/article/47
7. Segalovich Ilya. (2003), *A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine, Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications. MLMTA'03,* Las Vegas, Nevada, USA, 2003, pp. 721-729.
8. Perestoronin P. (2014), *Porter Stemmer for Russian language* [Stemmer Portera dlya russkogo yazika], available at: https://eigenein.xyz/snowball/