

СТВАРЭННЕ СЛОЎНІКАВАГА ЛЕМАТЫЗАТАРА ДЛЯ БЕЛАРУСКАЙ МОВЫ

А.А.Русакевіч (Мінск, МГЛУ)

Дадзеная работа апісвае працэс стварэння першага Python-лематызатара для беларускай мовы. Ён можа быць выкарыстаны для стварэння корпусаў беларускай мовы, палепшання якасці пошуку ў слоўніках і пошуку па беларускім тэкстам у цэлым.

Ключавыя словы: *беларуская мова; камп'ютарная лінгвістыка; лематызатар.*

CREATING A VOCABULARY LEMATIZER FOR BELARUSIAN LANGUAGE

A.A.Rusakevich (Minsk, MSLU)

This work describes the process of building the first Python lemmatizer for the Belarusian language. It can be used for building corpora for the Belarusian language, improving search performance in dictionaries and searching Belarusian texts in general.

Key words: *Belarusian language; linguistics; lemmatizer.*

Камп'ютарная лінгвістыка няўстанна рухаецца з часам. Яна дапамагае будаваць сайты, ШІ, пашуковыя рухавікі. Развіваюцца інструменты для камп'ютарнай апрацоўкі моў, такія як, напрыклад, NLTK, але колькасць падобных інструментаў для беларускай мовы значна меншая ў параўнанні са сусветнымі мовамі, што абмяжоўвае магчымасці аналізу камп'ютарам беларускамоўных тэкстаў.

Адным з кампанентаў вышэйадзначанай нястачы з'яўляецца адсутнасць лематызатара для беларускай мовы на Python. Ён неабходны для стварэння корпусаў тэкстаў, для палепшання пошука ў слоўніках, выдавання прапаноў у анлайн-слоўніках і пашукавіках. Дадзеная работа апісвае тое, як гэта праблема адсутнасці Python-лематызатара была вырашана [1].

Варта заўважыць, што лематызатар ужо існуе як частка праекта ЭКБМ, але ён напісаны на мове Perl, якая на момант 2025 году страціла папулярнасць і сваіх карыстальнікаў [3]. Да таго ж, яна працуе на старой аднабайтавай кадыроўцы `sr1251`, калі галоўная кадыроўка на момант 2025 года – гэта мнагабайтавая UTF-8. Гэты канфлікт кадыровак стварае рад праблем з так званымі “іерогліфамі” замест літар.

Існуе таксама API лематызатара ад `corpus.by`, але яго ўвод і вывад мае занадта многа параметраў у параўнанні з другімі лематызатарамі. Да таго ж, яго слоўніковая база вельмі абмежаваная: напрыклад, `corpus.by` не ведае слоў “прывітанне” і “амлет” і г.д.

Дадзеная работа не апісвае працэс “перакладу” з праграмы з Perl на Python або тое, як рабіць запыты да API. Яна апісвае тое, як лематызатар пісаўся на Python, апіраючыся выключна на яго функцыі і магчымасці, на ідэю лематызатара ЭКБМ і на арыенцір у выглядзе лематызатара NLTK.

Для аналізу была ўзята граматычная база дадзеных праекта “Беларускі N-корпус”, дзе змешчаны формы слоў і іх лемы. Гэтая база існуе ў фармаце XML, якая з’яўляецца даволі складанай для аналізу, таму гэтыя файлы былі перапрацаваны праграмным шляхам у базу дадзеных SQLite3, якая больш лёгка для аналізу за кошт адсутнасці парсінгу XML пры запуску, таму будзе хутчэй апрацоўвацца.

SQLite3 таксама дае магчымасць загрузіць базу дадзеных у апэратыўную памяць для павышэння эфектыўнасці пры дастатковай колькасці свабоднай RAM і адсутнасці неабходнасці ў імгненным запуску. Іменна гэты рэжым быў абраны ў якасці безумоўнага з магчымасцю вярнуцца да рэжыма чытання з дыску.

Створаны намі лематызатар дае магчымасць карыстацца тымі ж магчымасцямі і функцыямі, што і лематызатар WordNet ад NLTK. Метадам класу лематызатара былі абраны аналагічныя яму імёны і тыпы аргументаў з вяртаемым значэннем. Гэта было зроблена з мэтай паскорыць навучанне тых, хто ўжо знаёмы з NLTK, і каб у перспектыве прасунуць лематызатар у сам праект NLTK.

Як частка праекту быў напісаны яшчэ і міні-сервер на FastAPI. Гэта дазваляе карыстацца лематызатарам як мікрасэрвісам у сайтах, што дазволіць праграме працаваць больш эфектыўна, карыстацца плюсамі асінхроннага праграмавання і мнагядзернасці.

Лематызатар ужо апублікаваны ў архіве PyPI [2] і даступны для ўсіх жадаючых. Ён будзе зручны для камп’ютарных лінгвістаў, што працуюць з нашай мовай, студэнтаў, што вывучаюць камп’ютарную лінгвістыку і для праграмістаў, якія планіруюць адаптаваць свае праграмныя прадукты пад беларускую мову.

СПІС ЛІТАРАТУРЫ

1. Русакевіч, А.А. Лематызатар на аснове Беларускага N-корпуса // Github.com [Электронны рэсурс]. – 2025. Рэжым доступу : <https://github.com/alex-rusakevich/lemmatizer-be/>. – Дата доступу : 20.01.2025.
2. Rusakevich, A. Belarusian lemmatizer // A. Rusakevich // PyPI.org [Electronic resource]. – 2025. Mode of access : https://pypi.org/project/lemmatizer_be/. – Date of access : 24.01.2025.
3. TIOBE Index [Electronic resource]. – 2025. Mode of access : <https://www.tiobe.com/>. – Date of access : 24.01.2025.