НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Факультет прикладної математики

Кафедра програмного забезпечення комп'ютерних систем

"На правах рукопису"
УДК 004.021

«До захисту допущено»
Науковий керівник
кафедри
_____ І. А. Дичка
(підпис)
"___" _____ 2017 р.

Магістерська дисертація

на здобуття ступеня магістра

зі спеціальності 8.05010301 "Програмне забезпечення систем"
на тему: КОМБІНОВАНИЙ СТАТИСТИЧНИЙ МЕТОД АВТОМАТИЗОВАНОЇ ОРФОКОРЕЦАЦІЇ ПРИРОДНОМОВНИХ ТЕКСТОВИХ ДАНИХ

Виконав: студент 6 курсу, групи КП-51м
Лецик Андрій Олександрович _____
(підпис)

Науковий керівник к.т.н., доц., доц. Заболотня Т.М. _____
(підпис)

Рецензент к.т.н., доц., доц. Петрашенко А.В. _____
(підпис)

Рецензент к.т.н., доц., доц. Дідковська М.В. _____
(підпис)

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.
Студент _____
(підпис)

Київ – 2017

РЕФЕРАТ

Актуальність теми. На сьогоднішній день кількість інформації продовжує невпинно зростати. Це породжує наступні проблеми:

- швидкість обробки даних стає не достатньою;
- апаратні ресурси для обробки необхідно збільшувати;
- розробка нових алгоритмів і методів орфокорекції для мов, які відмінні від англійської, є невелика кількість;
- розробка нових методів потребує все більшої і більшої кількості людино-годин, оскільки їх складність збільшується.

На сьогоднішній день вже недостатньо просто виділити помилку в тексті. Комп’ютер має її визначити і виправити, так як кількість нової інформації невпинно збільшується і для ручного виправлення потребується значна кількість людей і затраченого часу.

Для роботи з україномовними текстами дуже мало методів. Деякі з них є загальними, але їх робота занадто повільна для обробки великої кількості інформації й їх результати не задовольняють вимоги. Існує велика кількість методів, які знаходяться у закритому доступі. Вони розроблені провідними компаніями у даній сфері: Google, Microsoft, Facebook та багато іншими. Доступ до них закритий, тому для вільного програмного забезпечення є потреба використання безкоштовних методів.

Об’єктом дослідження є процес орфокорекції природномовних текстових даних.

Предметом дослідження є статистичні методи орфокорекції природномовних текстових даних.

Мета роботи: підвищення ефективності виправлення орфографічних

помилок комп’ютерними системами автоматизованої обробки природномовних текстів за рахунок розробки статистичного методу орфокорекції природномовних текстових даних.

Методи дослідження: метод морфологічного аналізу, метод N-грам, метод Левенштейна, методи Стеммінга, методи повних обернених перетворень.

Наукова новизна:

1) Розроблено комбінований метод орфокорекції природномовних текстових даних, що відрізняється від існуючих використанням їх переваг:
    - визначення префікса і закінчення морфологічним методом;
    - пошук місця помилки методом N-грам;
    - сортування ймовірних рішень методом сортуванням помилок за імовірністю їх виникнення;
    - визначення точного результату методом Левенштейна.
2) Вперше використано метод сортування помилок за імовірністю їх появи у слові як частину розробленого комбінованого методу.

Практична цінність отриманих результатів полягають в тому, що вони дозволять розроблювати програмне забезпечення з автокорегуванням текстових даних. Метод може використовуватись для наступних типів програмного забезпечення:

- текстові редактори;
- *web*-портали;
- засоби для програної розробки;
- автоматизовані засоби для перевірки написання текстів, які

вбудовані як віджети у браузерах і автоматично вбудовуються на сайти.

Розробка, що використовуватиме даний метод, дозволить автоматично виправляти помилки у словах. Середні показники виправлення мають високий показник (більше 80%) і метод знаходить однозначне рішення. Якщо рішення не знайдено, то програма позначає слово як те, у якому знайдено помилку.

Апробація роботи. Основні положення, які представлені у методі, йог використання та практична значущість були обговорені на IX науковій конференції магістрантів та аспірантів «Прикладна математика та комп'ютеринг» ПМК-2017 (Київ 19-21 квітня 2017 р.).

Структура та обсяг роботи. Магістерська дисертація складається з вступу, п'яти розділів, висновків та додатків.

У вступі надано загальну характеристику проблеми, яка була поставлена для вирішення, охарактеризовано загальний напрямок роботи, обґрунтовано актуальність, сформульовано мету дослідження і практичне застосування результатів дослідження, вказано про впровадження результатів дослідження.

У першому розділі наведено головні базові методи для орфокорекції природномовних текстових даних, наведені їх приклади, вказуються основні переваги та недоліки існуючих базових методів. Вказано сучасні методи, які розроблені за останні 20-30 років і є базовими для комерційних методів і також наведені їх ключові переваги і недоліки. Зроблено детальний опис переваг, які можна використати для комбінованого методу, що можуть дати перевагу у точності, часі і розміру словника. Описано особливості реалізації

існуючих методів. Зроблено висновки по даним методам і описано підґрунтя для подальшого дослідження.

У другому розділі описано переваги, які було використано для розробки комбінованого методу і обґрунтовано їх вибір. Було детально проаналізовано вибір необхідних методів, які дали необхідну перевагу над існуючими методами. Описано етапи, які були розроблені для комбінованого методу, обґрунтовано вибір послідовності даних етапів і зроблено попередній аналіз можливих результатів програмної реалізації. Були зроблені висновки про розроблені етапи комбінованого методу орфокорекції природномовних текстових даних.

У третьому розділі було проведено аналіз вибору технологій для реалізації розробленого комбінованого методу, описано технології, які використовувались для реалізації методу. Описано інтерфейс програмної реалізації методу. Наведені деталі програмної розробки: описано базу даних, серверну та клієнтську частину. Описано деталі реалізації та зроблено висновки по реалізації.

У четвертому розділі наведено опис даних, які використовувались для тестування роботи програмної розробки і остаточного підтвердження результатів, які були досягнуті роботою методу. Було описано прикладне використання даної розробки, наведено можливі варіанти покращення і збільшення функціональних можливостей розробки. Останній підрозділ даного розділу вказує варіанти покращення існуючого методу і наступні варіанти продовження дослідження. Зроблено висновки по проведеному аналізу результатів отриманих під час тестування методу і програмної реалізації.

У п’ятому розділі наведено бізнес-ідею та бізнес-план створення і розвитку стартапу на основі розробленого методу.

У висновку проаналізовані результати дослідження.

У додатках наведені схема алгоритму розробленого методу, схема бази даних, таблиці з результатами роботи програми і діаграми прецедентів.

Робота виконана на 87 аркушах, містить 5 додатків та посилання на список використаних літературних джерел з 14 найменувань. У роботі наведено 17 рисунків та 1 таблиці.

Ключові слова: комбінований метод, орфокорекція, природномовні текстові дані, автоматизоване виправлення орфографічних помилок.

ABSTRACT

Relevance of the topic. Information amount increases nowadays. These caused next problems:

- speed of data processing is low;
- need to improve hardware;
- count of no English methods for spell correction is low;
- new methods developing require more and more human labor.

Only note mistake is not enough in the text nowadays. The computer have to find and fix it, because count of new information increases and manually correction need more humans and more time.

A small amount of methods for Ukrainian. Some of them are general but they work slowly for large amount of data processing or they have bad results. There are many methods that are unavailable for free using. Google, Microsoft, Facebook and many other companies created these methods. These methods have not free access. They are not suitable for open-source software.

Object of research is process of spell correction of text data.

Subject of research are statistic methods of spell correction of natural language text data.

Research objective: improving the efficiency of correction spelling mistakes by automated processing of spelling text data computer systems by developing statistical automatic correction method of spelling correction of text data.

Research methods. morphological analysis method, N-gram method, Levenshtein method, Stemming methods, full inverse transformations method.

Scientific novelty:

1) The combined method spell language correction of text data, that it is differ from existing by using their advantages:
    - searching prefix and end by morphological method;
    - searching mistake place by N-gram method;
    - sorting of probable variants of solution by the method of sorting errors on the probability by their occurrence;
    - determine the exact result by Levenshteyn method.
2) First used the method for sorting mistakes probability of their occurrence in the word as part of the combined method.

Practical value of results allow to develop software for automatic mistakes correction. Method can used for same types of software:

- text editors;
- *web*-sites;
- integrated development environment;
- software for automated text checking that it can be used as widget for site into the browser.

Software allows automatic correct mistakes. Average statistic of developed software has high percent of correct corrections (more 80%). Method finds one answer. Program sets word as incorrect when it is not find correct word.

Approbation. The results of the testing have been or are in the process of publication at the IX scientific conference of graduate and post-graduate students "Applied Mathematics and Computing" PMK-2017 (Kyiv, 19-21 April 2017).

Structure and content of the thesis. Master's thesis consists of an introduction, four chapters, conclusions and applications.

In the introduction, given the general description of the work performed

assessment of the current state of the problem, relevance of the research, formulated goals and objectives of the study.

The first section presents main basics methods of spelling language text data correction, describes basics methods advantages and disadvantages. There are specified modern methods that have been developed over the last 20-30 years and they are the base for business methods and also given their description. There are describes of software that are based on this methods. There are conclusions in this method and described the basis for further research.

The second section describes advantages that was used for combined method and justified their choice. There are describing method steps and analyze future results of the method, describing conclusions for software developing.

The third section describes about programming languages choosing, technologies for software developing, describing interface of web site, describing software developing: data base, server and client parts.

The fourth section describes about data for software testing, analyzes results and confirms results. There are describing applied using of software, gives possible improvements. Last subsection describes about variants of improvements and next researching.

The fifth section proposed business idea and business plan of creating startup using created method.

The conclusions contains analysis of results.

The applications contain results of testing software, DB diagram, algorithm scheme and use case diagram.

The thesis is presented in 87 pages, it contains 18 references to the used information sources, 18 figures and 1 table are given in the thesis.

Key words: combined method, spell correction, natural language text data, automatic spell correction of text data.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Заболотня, Т.М. Інверсний контекстно-асоціативний метод та програмні засоби автоматизованої орфокорекції природомовних текстових об'єктів [Текст] : дис. канд. техн. наук : 07.10.2008 / Т.М. Заболотня / НТУУ «КПІ» - К., 2008 – 17с.
2. Даревич, Р.Р. Автоматизована метапошукова система на основі адаптивної онтології [Текст] : дис. канд. техн. наук : 10.09.2007/ Р.Р. Заболотня / Нац. універ. «Львівська політехніка» - Л., 2007 — 18 с.
3. Wagner, R. A. The String-to-String» Correction Problem [Електронний ресурс] / R.A. Wagner — Режим доступу : http://www.inrg.csie.ntu.edu.tw/ algorithm2014/homework/Wagner-74.pdf. – (07.04.2016)
4. Розробка підсистеми морфологічного аналізу інформаційної системи [Електронний ресурс] — Режим доступу : http://bukvar.su/informatika_ programmirovanie/171742-Razrabotka-podsistemy-morfologicheskogo-analiza-informacionnoiy-sistemy.html. – (07.04.2016)
5. Розпізнавання текстового зображення з урахуванням морфології слова [Електронний ресурс] — Режим доступу : http://technomag.bmstu.ru/doc/ 350020.html. – (08.04.2016)
6. Знаходження найдовшої зростаючої підпослідовності [Електронний ресурс] — Режим доступу : http://roinet.livejournal.com/2706.html. – (08.04.2016)
7. Гниловська Л.П. Гниловська Н.Ф. Автоматична корекція орфографічних помилок // Культура народів Причорномор'я. — 2004. - № 48. – с. 171 – 180.

8. Нечіткий пошук в тексті і словнику [Електронний ресурс] // Алгоритми – 2011. — Режим доступу : https://habrahabr.ru/post/114997/. – (08.04.2016)
9. Метод побудови N-грамної моделі адаптованої для слов'янських мов [Електронний ресурс] // Інновації в науці – 2014. — Режим доступу : http://sibac.info/conf/innovation/xxxiii/38409. – (08.04.2016)
10. Н. Є. Бузикашвілі, Г.А. Крилова, Д.В. Самойлов. N-грами в лінгвістиці [Електронний ресурс] — Режим доступу: www.cognitive.ru/assets/docs/ scienwork/ sbornic2/samoilov.doc. – (08.04.2016)
11. Задача про відстань Дамерау-Левенштейна [Електронний ресурс] — Режим доступу: http://neerc.ifmo.ru/wiki/index.php?title=Задача_о_ расстоянии-Дамерау-Левенштейна – (08.04.2016).
12. About Python [Електронний ресурс] — Режим доступу: http://www. plug.org.ua/documentation/about-python – (05.05.2017)
13. Django[Електронний ресурс]. — Режим доступу : http://uk.wikipedia.org /wiki/Django. Дата доступу : квітень 2015. Назва з екрана. — (05.05.2017)
14. Writing your first Django app, part 1[Електронний ресурс]. — Режим доступу: https://docs.djangoproject.com/en/1.8/intro/tutorial01/ – (06.05.2017)
15. PostgreSQL [Електронний ресурс]. — Режим доступу: https://uk. wikipedia.org/wiki/PostgreSQL – (07.05.2017)
16. Взаємовдія програми з PostgreSQL [Електронний ресурс]. — Режим доступу: http://www.kytok.org.ua/?p=447 – (07.05.2017)
17. CSS3 [Електронний ресурс]. — Режим доступу: https://uk.wikipedia.

org/wiki/CSS – (07.05.2017)

18. HTML5 [Електронний ресурс]. — Режим доступу: https://ru.wikibooks .org/wiki/HTML5 – (07.05.2017)