ОДЕСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені І.І.МЕЧНИКОВА

(повне найменування вищого навчального закладу)

Факультет математики, фізики та інформаційних технологій

(повне найменування інституту, назва факультету (відділення))

Кафедра математичного забезпечення комп'ютерних систем

(повна назва кафедри (предметної, циклової комісії))

Дипломна робота

на здобуття ступеня вищої освіти «магістр»

(освітньо-кваліфікаційний рівень)

на тему: Розробка програмної підтримки для створення корпусу української мови та його застосування

Development of software for the creation of the corpus of the Ukrainian language and its use

Виконав: студент денної форми навчання
спеціальності 123 – Комп'ютерна інженерія.

(шифр і назва напряму підготовки, спеціальності)

Керпель Олексій Ігорович

(прізвище, ім'я, по-батькові)

Керівник доц., к.т.н. Пенко Валерій Георгійович

(науковий ступінь, вчене звання, прізвище та ініціали, **підпис**)

Рецензент доц., к.ф.-м.н Антоненко О. С.

(науковий ступінь, вчене звання, прізвище та ініціали)

Рецензент д.т.н., проф. Кобозєва А.А.

(науковий ступінь, вчене звання, прізвище та ініціали)

Рекомендовано до захисту:

Протокол засідання кафедри

№_____ від «___»_____ 2020 р.

Завідувач кафедри

_____ _Є.В. Малахов_

(підпис) (прізвище, ініціали)

Захищено на засіданні ЕК № ___

протокол № __ від «___» ____ 2020 р.

Оцінка_____/_____/_____

(за національною шкалою, шкалою ECTS, бали)

Голова ЕК

_____ _Н.Ф. Казакова_

(підпис) (прізвище, ініціали)

Одеса – 2020

АНОТАЦІЯ

Тема дипломної роботи: «Розробка програмної підтримки для створення корпусу української мови і його використання».

Актуальність дипломної роботи полягає в необхідності аналізу українських текстів з метою вивчення української мови.

Об'єкт дослідження – засоби складання POS-тегованих корпусів українських текстів.

Предмет дослідження – процес створення програмних засобів для створення корпусів українських текстів, зокрема, алгоритмів POS тегування.

Мета роботи – дослідження і розробка програмних засобів для створення анотованого корпусу української мови.

Для досягнення поставленої мети були вирішені наступні задачі: аналіз предметної області; вибір відповідних програмних засобів; створення навчальної множини; реалізація системи; навчання системи; тестування системи.

В результаті проведеної роботи отримано навчений тегер української мови.

АННОТАЦИЯ

Тема дипломной работы «Разработка программной поддержки для создания корпуса украинского языка и его использования».

Актуальность дипломной работы заключается в необходимости анализа украинских текстов с целью изучения украинского языка.

Объект исследования – средства составления POS-тегированных корпусов украинских текстов.

Предмет исследования – процесс создания программных средств для создания корпусов украинских текстов, в частности, алгоритмов POS тегирования.

Цель работы – исследование и разработка программных средств для создания аннотированного корпуса украинского языка.

Для достижения поставленной цели были решены следующие задачи: анализ предметной области; выбор подходящих программных средств; создание обучающего множества; реализация системы; обучение системы; тестирование системы.

В результате проведеной работы был получен обученный теггер украинского языка.

ABSTRACT

Thesis topic: "Development of software for the creation of the corpus of the Ukrainian language and its use"

The relevance of the thesis lies in the need to analyze Ukrainian texts in order to study the Ukrainian language.

The object of the research is the variety of tools for the task of POS tagging of Ukrainian texts.

The subject of the research is the process of creating software for creating corpora of Ukrainian texts, in particular, POS-tagging algorithms.

The purpose of the work is to research and develop software for creating an annotated corpus of the Ukrainian language.

To achieve this goal, the following tasks were solved: analysis of the subject area; selection of suitable software tools; creating a training set; system implementation; system training; system testing.

As a result of this work, a trained Ukrainian language tagger was obtained.

ЗМІСТ

ВСТУП.....7
1 ОГЛЯД МЕТОДІВ І СИСТЕМ СТВОРЕННЯ МОВНИХ КОРПУСІВ.....8
1.1 Основні властивості корпусів.....8
1.2 Класифікація корпусів.....9
1.3 Розмітка корпусів.....12
1.4 Огляд іноземних корпусів.....15
1.4.1 Корпус Брауна.....15
1.4.2 Британський національний корпус.....16
1.4.3 Corpus of Contemporary American English.....17
1.5 Огляд існуючих корпусів української мови.....17
1.5.1 КНУ ім. Т. Шевченка.....18
1.5.2 НУ Острозька академія.....19
1.6 Існуючі системи, які використовуються для створення корпусів.....20
1.6.1 Apache OpenNLP.....20
1.6.2 CoreNLP.....20
1.6.3 NLTK.....21
1.7 Постановка задачі.....21
2 КЛАСИФІКУЮЧІ МОДЕЛІ ЯК ІНСТРУМЕНТ ДЛЯ СТВОРЕННЯ ТЕГОВАНИХ КОРПУСІВ.....22
2.1 Задачі класифікації.....22
2.2 Огляд актуальних классифікуючих моделей.....24
2.2.1 SGD.....25
2.2.2 Дерево рішень.....26

2.2.3 MLP 27
2.2.4 Метод k-найближчих сусідів 28
2.2.5 Випадкові ліси 29
2.2.6 AdaBoost 30
2.3 Крос-валідація 31
2.4 Оцінка якості роботи системи 33
3 ПРОЕКТУВАННЯ І РЕАЛІЗАЦІЯ ПІДХОДУ ДО РОЗМІТКИ УКРАЇНСЬКИХ ТЕКСТІВ 35
3.1 Частини мови в українській мові 35
3.2 Прості автоматизовані тегери 36
3.3 Складання навчальної вибірки 37
3.4 Опис стратегії виконання експериментів машинного навчання 38
3.5 Пошук оптимального набору ознак 38
3.6 Порівняння ефективності роботи класифікаторів 40
3.7 Реалізація системи 41
3.8 Модуль навчання тегеру 42
3.9 Модуль класифікації 42
3.10 Графічний інтерфейс 43
3.11 Залежність показників тегування від розмірів навчальної вибірки 44
3.12 Опис роботи системи 45
ВИСНОВОК 47
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ 48
ДОДАТОК А КОД КЛАСИФІКАТОРА 50
ДОДАТОК Б ТОЧНІСТЬ КЛАСИФІКУЮЧИХ МОДЕЛЕЙ 52

ВСТУП

На даний момент обробка текстових даних стає все більш і більш затребувана як в системах розпізнавання мови (Cortana, Siri), так і в пошукових системах (Google, Yandex, Yahoo). Також комп'ютерна обробка текстів необхідна для аналізу творів різних письменників, знаходження певних закономірностей в текстах різних епох.

Сама по собі обробка текстів може бути корисна в таких задачах як організація пошуку в пошукових системах, розпізнавання мови, автоматичне визначення тематики веб-сторінок, перекладі текстів, і в багатьох подібних задачах, що вимагають просунутої роботи з текстами.

Для аналізу текстів використовуються «корпуси» текстів – збірники різних текстів, які дозволяють розглянути мову з різних сторін і містять розмітку (мета-дані, такі як морфологічна (POS або part of speach) розмы, семантична, синтаксична, і т.д.). До недавнього часу єдиним можливим способом складання подібного корпусу було тегування текстів «вручну», але з розвитком інформаційних технологій, перед нами відкрилися нові можливості, що дозволяють автоматизувати процес розмітки і знизити (або повністю прибрати) необхідність втручання людини (експерта) у процес обробки текстів.

На жаль, для української мови досі не існує достатніх для практичного використання розмічених корпусів і програмних ресурсів, що дозволяють здійснити розмітку текстів.

У зв'язку з цим метою даної роботи є створення автоматизованої системи, що дозволяє розмічати тексти і, в перспективі, створювати повноцінні корпуси українських текстів.

ВИСНОВОК

В ході виконання даної магістерської роботи спроектована, реалізована і протестована система, яка виконує автоматичну розмітку українських текстів за частинами мови.

Основний механізм класифікації базується на методах машинного навчання з учителем, реалізованих в де-факто стандартному пакеті Scikitlearn. Ефективне використання можливостей цього пакету передбачає реалізацію кількох дослідницьких процедур.

Зокрема, для реалізації системи проведений пошук оптимального набору ознак класів і їх конструювання з навчальної множини. Також проаналізовано ефективність роботи різних класифікуючих моделей і обрано оптимальну класифікуючу модель, якою опинилася SGD класифікатор. Для отримання високої узагальнюючої здатності класифікатора використовувалася техніка крос валідації методом K блоків.

Реалізований POS-тегер дозволяє користувачам в короткі терміни і з високою ефективністю протегувати великі обсяги інформації. Таким чином представлений тегер може бути використаний для створення корпусу української мови, що дасть лінгвістам додаткові можливості дослідження української мови.

Для підвищення показників якості класифікації перспективним представляється включення в просторі ознак контекстної інформації, в тому числі інформації про теги сусідніх слів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. The Collins Corpus [Електронний ресурс] – Режим доступу: https://collins.co.uk/pages/elt-cobuild-reference-the-collins-corpus
2. Georgian Language Corpus [Електронний ресурс] – Режим доступу: http://corpora.iliauni.edu.ge
3. Russian Language Corpus [Електронний ресурс] – Режим доступу: https://ruscorpora.ru/new
4. The iWeb Corpus [Електронний ресурс] – Режим доступу: https://www.english-corpora.org/iweb
5. Bucknell University [Електронний ресурс] – Режим доступу: https://www.departments.bucknell.edu/linguistics/lectures/05lect09.html
6. UCREL Corpus Annotation [Електронний ресурс] – Режим доступу: http://ucrel.lancs.ac.uk/annotation.html
7. ICAME CORPUS MANUALS [Електронний ресурс] – Режим доступу: http://korpus.uib.no/icame/manuals
8. Корпус української мови MOVA.info [Электронный ресурс] – Режим доступу: http://www.mova.info/syntaxis_search.aspx – 25.06.2018
9. Проект створення корпусів текстів | Національний університет «Острозька академія» [Електронний ресурс] – Режим доступу: https://www.oa.edu.ua/ua/departments/filologist/filol_literature/lexilab/project3
10. Вступ до мовознавства: Підручник для студентів філологічних спеціальностей вищих навчальних закладів. — К.: Видавничий центр «Академія», 2001))
11. А.М. Земсикй, С.Е. Крючков, М.В. Светлаев – Русский язык в двух частях. М.: Просвещение, 1986.
12. Новий довідник: Українська мова та література. – К.: Тов "Казка", 2005.

13. NLTK 3.5 documentation [Електронний ресурс] – Режим доступу: https://www.nltk.org/api/nltk.tag.html
14. PART OF SPEECH TAGGING WITH NLTK [Електронний ресурс] – Режим доступу: https://streamhacker.com/2008/11/10/part-of-speech-tagging-with-nltk-part-2