

Одеса - 2019

АННОТАЦІЯ

У данній дипломній роботі розроблена тема «Розробка граматик вилучення синонімів і гіперонімів з тексту заданного словника».

Метою роботи є проектування і реалізація системи, яка замінює вхідний текст на синоніми та гіпероніми. Розроблена система надає можливість користувачам подавати на вхід будь-який текст російської мови та отримати на виході текст з семантично близьким значенням, заміненний синонімами та гіперонімами.

Мовою розробки є Python, для написання коду і компіляції використане середовище розробки PyCharm.

АННОТАЦИЯ

В данной дипломной работе разработана тема «Разработка грамматик извлечения синонимов и гиперонимов из текста заданного словаря».

Целью работы является проектирование и реализация системы, которая заменяет входной текст на синонимы и гиперонимы. Разработанная система дает возможность пользователям подавать на вход любой текст русского языка и получить на выходе текст с семантически близким значением, замененный синонимами и гиперонимами.

Языком разработки является Python, для написания кода и компиляции использована среда разработки PyCharm.

ANNOTATION

In this thesis, the theme «Develop the grammars for extracting synonyms and hyperonyms from the text of given dictionary» has been developed.

The aim of the work is to design and implement a system that replaces the input text with synonyms and hyperonyms. The developed system allows users to submit any text of the Russian language to the input and to receive a text with a close semantical meaning, replaced by synonyms and hyperonyms.

The development language is Python, the PyCharm development environment is used for writing code and compiling.

ЗМІСТ

ВСТУП.....	4
1 ОГЛЯД ІСНУЮЧИХ АНАЛОГІВ.....	5
1.1 Онлайн синонімайзер USyn	5
1.2 SeoGenerator.....	6
1.3 Online-Sinonim.ru	8
2 ПОСТАНОВКА ЗАВДАННЯ	9
3 ТЕОРИТИЧНА ЧАСТИНА ПРОЕКТУ.....	10
3.1 NLP (Natural Language Processing)	10
3.2 Предобробка тексту.....	12
3.3 Нейронні мережі.....	14
3.4 Векторне представлення слів.....	18
3.5 Word2Vec	19
3.6 CBOW (Continuous Bag of words).....	21
3.7 Skip-Gram.....	23
3.8 Ієрархічний Softmax.....	26
3.9 Негативне Семлірування.....	28
4 ПРОГРАМНА РЕАЛІЗАЦІЯ ПРОЕКТУ.....	30
4.1 Вибір бібліотек	31
4.2 Опис функцій.....	32
4.3 Вибір засобів розробки.....	33
5 РЕЗУЛЬТАТИ РОБОТИ ПРОГРАМИ.....	36
ВИСНОВКИ	40
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ	41

ВСТУП

Робота з текстом є однією з найважливіших завдань в житті кожної людини. В кожному столітті, в суспільстві велику повагу та визнання отримували люди, які могли з набору слів скласти цікаві та цінні тексти.

В наш час робота з текстом є не тільки забавою чи способом увійти в історію, як письменник, дуже часто це бізнес та робота для багатьох людей. Що стосується бізнесу, то це SEO менеджмент та SEO оптимізація: наповнювання сайтів унікальним контентом для просування сайту у високі позиції Google пошуку для залучення нових клієнтів чи описання якогось товару різними словами, щоб контент не був одноманітним.

Що стосується роботи для багатьох людей, то з поширенням Інтернету з'явилась необхідність наповнювати сайти, журнали та різноманітні інформаційні портали статтями на різну тематику. Подібні статті люди купують чи продають самі. Такі люди звуться копірайтерами, в обов'язки яких входять писати статті на теми, які цікавлять заказчика. Роками копірайтери робили усю роботу самі, витрачаючи дуже багато часу. З огляду на це, з'явилась необхідність в розробці системи, яка б могла сама перекладати абсолютно будь-який текст в близький по сенсу.

Також зараз у світі дуже стрімко розвивається робототехніка та Інтернет речей. У наше життя входять такі елементи як: розумні дома, роботи-помічники і т.д. Однією із задач є створення умов для розуміння роботизованих систем та людини.

Головною метою роботи є розробка системи, яка не просто замінює слова в тексті на синоніми та гіпероніми, а яка видає текст семантично близький по сенсу до вхідного.

ВИСНОВКИ

В ході даної роботи були розібрані теоретичні питання, які властиві для обробки природніх мов, методів представлення слів векторами та її архітектура. Результатом роботи є робоча система, яка реалізує усі поставленні цілі.

У першому розділі були розглянені існуючі аналоги та проаналізовані загальні переваги та недоліки доступних систем.

У другому розділі була поставлена основна мета роботи та сформульован функціонал програми.

У третьому розділі були розглянуті загальні теоретичні питання, які допомогли розкрити суть методу Word2Vec, зрозуміти, що являє собою модель. Були розглянуті проблеми, а також переваги даної моделі. Була розглянута теоретична частина теми нейронних мереж. Були розглянуті нейромережеві архітектури, на яких побудована модель Word2Vec.

У четвертому розділі був розглянут та аргументован вибір бібліотек для створення системи, були описані основні функції та обгрунтован вибір засобів розробки.

У п'ятому розділі були приведені результати роботи програми зі скриншотами та їх описанням.

На поточний момент розробка систем з аналізом та маніпуляцією природних мов швидко розвиваються і є перспективним напрямком.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Стаття Natural Language Processing (NLP) for Machine Learning [Електронний ресурс] – Режим доступа: <https://towardsdatascience.com/natural-language-processing-nlp-for-machine-learning-d44498845d5b>
2. Стаття A Simple Introduction to Natural Language Processing [Електронний ресурс] – Режим доступа: <https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32>
3. Стаття Natural Language Processing is Fun! [Електронний ресурс] – Режим доступа: <https://medium.com/@ageitgey/natural-language-processing-is-fun-9a0bff37854e>
4. Основные термины в Natural Language Processing [Електронний ресурс] – Режим доступа: <https://singularika.com/ru/nlp/natural-language-processing-terms/>
5. Как работает нейронная сеть: алгоритмы, обучение, функции активации и потери [Електронний ресурс] – Режим доступа: <https://neurohive.io/ru/osnovy-data-science/osnovy-nejronnyh-setej-algoritmy-obuchenie-funkcii-aktivacii-i-poteri/>
6. Нейронные сети для начинающих Часть 1 [Електронний ресурс] – Режим доступа: <https://habr.com/ru/post/312450/>
7. Искусственные нейронные сети [Електронний ресурс] – Режим доступа: <https://moluch.ru/archive/138/38781/>
8. Get Busy with Word Embeddings – An Introduction [Електронний ресурс] – Режим доступа: <https://www.shanelynn.ie/get-busy-with-word-embeddings-introduction/>
9. What is One Hot Encoding? Why And When do you have to use it? [Електронний ресурс] – Режим доступа: <https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f>

10. Немного про word2vec: полезная [Электронный ресурс] – Режим доступа: <http://nlpx.net/archives/179>
11. Чудесный мир Word Embeddings: какие они бывают и зачем нужны? [Электронный ресурс] – Режим доступа: <https://habr.com/ru/company/ods/blog/329410/>
12. Understanding Word Vectors and Word2Vec [Электронный ресурс] – Режим доступа: <http://www.stokastik.in/understanding-word-vectors-and-word2vec/>
13. Introduction to Word Embedding and Word2Vec [Электронный ресурс] – Режим доступа: <https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa>
14. Hierarchical softmax and negative sampling: short notes worth telling [Электронный ресурс] – Режим доступа: <https://towardsdatascience.com/hierarchical-softmax-and-negative-sampling-short-notes-worth-telling-2672010dbe08>
15. Gensim – Руководство для начинающих [Электронный ресурс] – Режим доступа: <https://webdevblog.ru/gensim-rukovodstvo-dlya-nachinajushhih/>
16. Морфологический анализатор pymorphy2 [Электронный ресурс] – Режим доступа: <https://pymorphy2.readthedocs.io/en/latest/user/index.html>
17. Марк Лутц, Изучаем Python, 4-е издание. – Пер. с англ. – СПб.: Символ-Плюс, 2011. – 1280 с
18. Mastering Python for Data Science , Samir Madhavan, 2016 – 1060 с.
19. PyCharm — интеллектуальная Python IDE [Электронный ресурс] – Режим доступа: <https://jetbrains.ru/products/pycharm/>

