

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «ЛЬВІВСЬКА
ПОЛІТЕХНІКА» ІНСТИТУТ КОМП'ЮТЕРНИХ НАУК ТА
ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра систем штучного інтелекту



ЗВІТ

про виконання практичної роботи №3
з курсу «Науковий процес та робота з науковими джерелами»

Пошук наукової літератури за допомогою спеціалізованих пошукових систем
та бібліографічних баз

Виконав:

ст. гр. КН-409

Тимошенко Павло Олександрович

Перевірив:

доцент каф. СШІ, к.т.н., Ізонін І. В.

Львів – 2021

1. Постановка завдання

Далі описано пункти, які потрібно виконати у межах цієї практичної роботи.

1. Написати тему та мету дипломної роботи, при можливості об'єкт та предмет дослідження. Визначити та оформити у таблицю 3-6 ключових слів, які найбільш повно можуть описати тематику.
2. Знайти не менше трьох найбільш цитованих статей за темою роботи у бібліометричній базі даних Google Scholar (українською мовою). Включити у пошук лише статті, які опубліковано у 2020 році. Виключити із пошуку патенти та посилання. Додати знайдені роботи у Zotero.
3. На основі обраних ключових слів сформулювати пошуковий запит в наукометричній базі Scopus. Отриманий набір робіт зменшити із використанням наступних критеріїв:
 - публікації лише з відкритим доступом;
 - тип документу – лише статті (а не тези конференції);
 - документи лише англійською;
 - публікації за останні три роки.
4. Здійснити опис проведеної роботи.

2. Хід роботи

В цьому розділі описано процес виконання практичної роботи та пророблені кроки.

Тема дослідження: Використання та оптимізація алгоритму Вітербі при визначення частин мови в тексті українською.

Мета дослідження:

- розкрити застосовність алгоритму Вітербі до текстів українською;
- з'ясувати, чи цей алгоритм буде так само надійним, як і в аналізі тексту англійською, де його використали вперше;
- розробити надійний та швидкий спосіб використання алгоритму на тих словах, на яких не проводилось навчання;
- виявити надоліки алгоритму та продемонструвати можливі способи оптимізації.

Звідси можна сформулювати такі ключові слова: *алгоритм Вітербі* (англ. *Viterbi algorithm*), *NLP*, *розмітка частин мови* (англ. *POS tagging*), *українська мова* (англ. *Ukrainian language*).

2.1. Пошук в Google Scholar

Сформовані запити пошуку в Google Scholar показано на Рис. 1 та Рис. 2. Посилання на знайдені статті подано далі:

1. https://www.researchgate.net/profile/Vasyl-Starko/publication/344841042_Semantic_Annotation_for_Ukrainian_Categorization_Scheme_Principles_and_Tools/links/5f92d758458515b7cf96d83c/Semantic-Annotation-for-Ukrainian-Categorization-Scheme-Principles-pdf (цитовано в 9 джерелах);
2. <http://ceur-ws.org/Vol-2604/paper21.pdf> (цитовано в 9 джерелах);
3. https://www.emerald.com/insight/content/doi/10.1016/j.aci.2018.12.003/full/html?utm_source=rss&utm_medium=feed&utm_campaign=rss_journalLatest (цитовано в 8 джерелах).

Google Академія

viterbi algorithm nlp

Статті

Приблизна кількість результатів: 556 (0,04 сек.)

Будь-коли
3 2021
3 2020
3 2017
Спеціальний діапазон...
2020 — 2020
Пошук

Сортувати за відповідн.
Сортувати за датою

Усі види
☐ включаючи патенти
☐ включаючи цитування

Оглядові статті
☒ Створити сповіщення

Parts of Speech Tagging in NLP: Runtime Optimization with Quantum Formulation and ZX Calculus
AK Bishwas, A Mani, V Palade - arXiv preprint arXiv:2007.10328, 2020 - arxiv.org
... In the process of building the solution, we develop a quantum version of the **Viterbi algorithm**. Additionally, we also demonstrate the optimization in implementing the solution in a quantum computer with the help of the ZX-calculus [7] keeping the target in the context of ...
☆ Пов'язані статті Кількість версій: 3

An Empirical Study on Efficiency of a Dictionary Based Viterbi Algorithm for Word Segmentation
S Aggarwal, S Houshmand... - ... Conference on Big ..., 2020 - ieeeexplore.ieee.org
... INTRODUCTION Strings consisting of words without spaces appear in relation to many contexts in natural language processing (**NLP**). Twitter and Instagram hash-tags (for example #photooftheday or #internationalwomensday) and URLs contain strings of words without clear ...
☆ Пов'язані статті Кількість версій: 2

The Optimization of the Dynamic Mechanism of Jilin's Economic Revitalization from the Perspective of Viterbi Algorithm's Supply-Side Structural Reform
Y Jin - International Journal of Uncertainty, Fuzziness and ..., 2020 - World Scientific
... The techniques used for the extraction are basic **NLP** and syntactic rule techniques. The syntactic patterns and rules used token based rules, regular expression at word level to extract the reasonable data from text. This text extraction would be beneficial for several text ...
☆ Пов'язані статті

A Stochastic Part of Speech Tagger for the Sinhala Language based on Social Media Data mining
SG Withanage, T Silva - ... Conference on Advances in ICT for ..., 2020 - ieeeexplore.ieee.org
... Abstract—Part of Speech (POS) taggers plays a critical role in **NLP** applications for analyzing the behavior and the construction of a ... This research uses the **Viterbi algorithm** for the decoding task. **Viterbi algorithm** is used to find the **Viterbi** path which consists of the most likely ...
☆ Пов'язані статті

Prediction of POS Tagging for Unknown Words for Specific Hindi and Marathi
K Chiplunkar, M Kharche, T Chaudhari... - ... 2020), Volume 2, 2020 - books.google.com
... The knowledge of part of speech plays a vital role in **NLP** because it tells us how a word is used in every sentence. POS tagging is the pre-requisite for natural language ... **Viterbi algorithm** uses above probability values and finds the most likely sequence of tags for the given input. ...
☆ Пов'язані статті

Рис. 1: Пошуковий запит у Google Scholar з текстом «pos tagging ukrainian»

2.2. Пошук в Scopus

Для пошуку в Scopus я сформував такий запит: TITLE-ABS-KEY (("pos"OR nlp) AND "viterbi algorithm") AND PUBYEAR > 2019 AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (OA , "all")).

Обрані статті є 4, 5 та 6-ою позицією в бібліографії наступного розділу.

Google Академія

pos tagging ukrainian

Статті

Приблизна кількість результатів: 248 (0,02 сек.)

Будь-коли
3 2021
3 2020
3 2017

Спеціальний діапазон...

2020 — 2020

Пошук

Сортувати за відповідн.
Сортувати за датою

Усі види

☐ включаючи патенти

☐ включаючи цитування

Оглядові статті

☒ Створити сповіщення

[PDF] Semantic Annotation for Ukrainian: Categorization Scheme, Principles, and Tools. [\[PDF\] researchgate.net](#)
V Starko - COLINS, 2020 - researchgate.net
... by assigning semantic tags to individual lemmas in VESUM, a POS dictionary for Ukrainian. TagText, the POS tagger for Ukrainian by the r2u group, can then be modified to carry out semantic tagging of texts. Complex queries enabled by a combination of morphological (...)
☆ 99 Цитовано в 9 джерелах Пов'язані статті Кількість версій: 3

DEVELOPMENT OF SOFTWARE FOR THE CREATION OF THE CORPUS OF THE UKRAINIAN LANGUAGE AND ITS USE.
Ol Kerpel, VG Penko - Informatics & Mathematical Methods ..., 2020 - search.ebscohost.com
... 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 ... work, a trained tagger of the Ukrainian language was obtained, which, for a sample of literary Ukrainian texts, ...
☆ 99 Пов'язані статті

[PDF] Using a Distributional Semantic Model for Collocation Identification [\[PDF\] lp.edu.ua](#)
A Mosinyan, S Petrasova - ... and intelligent systems: proceedings of the ..., 2020 - ena.lp.edu.ua
... This paper proposes the approach to automatic collocation identification using both the distributional semantic model and POS-tagging. The authors suggest calculating PMI to obtain a sequence of collocations from the designed corpus of research abstracts. Then ...
☆ 99 Пов'язані статті

[PDF] Collection and Processing of a Medical Corpus in Ukrainian. [\[PDF\] ceur-ws.org](#)
O Cherednichenko, O Kanishcheva, O Yakovleva... - COLINS, 2020 - ceur-ws.org
... We can underline some features of Ukrainian medical texts. Medical texts contain many borrowed words that came from Latin, for instance. ... We used POS tagger Pymorph2 for the analysis of speech parts of our corpus (https://pymorph2.readthedocs.io/en/latest/index.html). The ...
☆ 99 Цитовано в 9 джерелах Пов'язані статті

Experimental Investigation of Significant Keywords Search in Ukrainian Content
O Bisikalo, V Vysotska, V Lytvyn, O Brodyak... - ... on Computer Science ..., 2020 - Springer
... Hence, our first objective is to create a POS-tagger and lemmatizer dedicated to the processing of Ukrainian texts. Building of the annotated corpus has already started and several dozens of manually tagged documents with syntactic and morphological information in ...
☆ 99 Пов'язані статті

Computer Based Stylometric Analysis of Texts in Ukrainian Language
A Mazurko, T Walkowiak - ... Conference on Artificial Intelligence and Soft ..., 2020 - Springer
... The most common founded NLP tool called NLTK doesn't support Ukrainian, therefore for POS tagging was selected UDPipe [17]. It is a generic powerful linguistic tool that is associated with UD (Universal Dependencies) framework that supports more than 70 languages, ...
☆ 99 Пов'язані статті

Рис. 2: Пошуковий запит у Google Scholar з текстом «viterbi algorithm nlp»

2.3. Сформована бібліографія

Обрані статті я додав у Zotero. На Рис. 3 показано вигляд вікна зі статтями. На Рис. 4 подана сформована у стилі IEEE бібліографія.

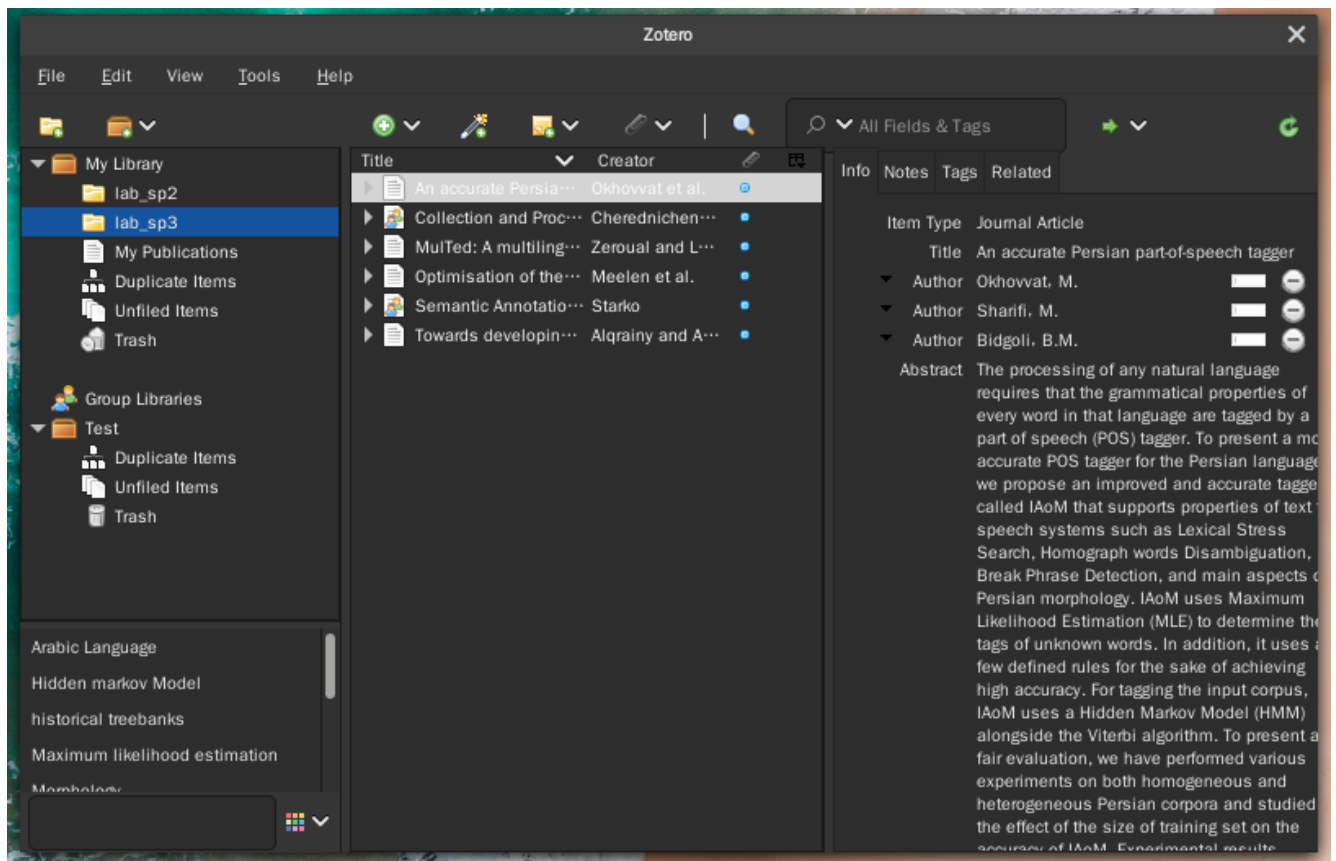


Рис. 3: Вікно Zotero з доданими статтями.

- [1] M. Okhovvat, M. Shanifi, i B. M. Bidgoli, «An accurate Persian part-of-speech tagger», *Computer Systems Science and Engineering*, вип. 35, вип. 6, с. 423–430, 2020, doi: 10.32604/csse.2020.35.423.
- [2] O. Cherednichenko, O. Kanishcheva, O. Yakovleva, i D. Arkatov, «Collection and Processing of a Medical Corpus in Ukrainian», в *COLINS*, 2020, с. 272–282.
- [3] I. Zeroual i A. Lakhouaja, «MulTed: A multilingual aligned and tagged parallel corpus», *Applied Computing and Informatics*, 2020.
- [4] M. Meelen, E. Roux, i N. Hill, «Optimisation of the Largest Annotated Tibetan Corpus Combining Rule-based, Memory-based, and Deep-learning Methods», *ACM Transactions on Asian and Low-Resource Language Information Processing*, вип. 20, вип. 1, 2021, doi: 10.1145/3409488.
- [5] V. Starko, «Semantic Annotation for Ukrainian: Categorization Scheme, Principles, and Tools.», в *COLINS*, 2020, с. 239–248.
- [6] S. Alqrainy i M. Alawairdhi, «Towards developing a comprehensive tag set for the Arabic language», *Journal of Intelligent Systems*, вип. 30, вип. 1, с. 287–296, 2021, doi: 10.1515/jisys-2019-0256.

Рис. 4: Сформована бібліографія.

3. Результати та висновки

У цій практичній роботі я сформував список ключових слів на основі поданої теми та мети дослідження. За ключовими словами я знайшов 6 статей у каталогах Google Scholar та Scopus, використовуючи не тільки текстуальний пошук, але й змінюючи категорії. Окрім того я навчився будувати пошуковий запит у системі Scopus.

Отримані вміння є важливими при написанні наукової роботи, оскільки дозволяють знайти інші праці на схожу тематику та посилатись на них. Це допоможе мені у написанні дипломної роботи.

4. Список використаної літератури

1. Пошукова та наукометрична система Гугл Академія (Google Scholar) [Електронний ресурс] – Режим доступу: <http://dnsgb.com.ua/files/%D0%A1%D0%98%D0%A1%D0%A2%D0%95%D0%9C%D0%90%20GOOGLE%20%D0%90%D0%9A%D0%90%D0%94%D0%95%D0%9C%D0%86%D0%AF.pdf> (відвідано 08.11.2020)
2. Веб-адреса бази даних Google Scholar. [Електронний ресурс] – Режим доступу: <https://scholar.google.com/> (відвідано 08.11.2020)
3. Scopus – докладна інструкція для вченого [Електронний ресурс] – Режим доступу: <https://ua.publ.science/uk/blog/scopus---podrobnaya-instruktsiya-dlya-uchenogo> (відвідано 08.11.2020)