Zgłoszenie tematu pracy dyplomowej :: **STUDIA II STOPNIA** ::

na rok akademicki 2024/25

Promotor:	**dr hab. Jozef Kapusta, prof. UKEN**
Temat pracy magisterskiej (j. polski oraz j. angielski):	Tool for Finding Similar Content on a Web Portal Narzędzie do wyszukiwania podobnych treści na portalu internetowym
Zakres i oczekiwane rezultaty pracy:	Document similarity (or distance between documents) is one of the central themes in Information Retrieval. Usually, documents are treated as similar if they are semantically close and describe similar concepts. On the other hand, “similarity” can be used in the context of duplicate detection. One of the main problems of "large" web portals with many publishers is duplicity in the content. Counting the similarities of the documents could effectively help administrators of portals to optimize the web content. In the theoretical part: The theoretical part of the thesis will summarize the methods for word embedding (Word2vec, Tf-Idf, GloVe, etc.), metrics for comparison vectors of documents, techniques for web crawler etc. An important part of the work is to explore models and approaches other researchers have already developed. In the practical part: The aim of the thesis is creating a simple tool for web portals administrators. The tool will calculate and show pages from the web portal with similar content. The student will choose an appropriate document model (TF-IDF, TF, vector, boolean document model, etc.) for indexing web content, and create crawler for web portal indexing. The main task will be to calculate documents similarities and show the result for the user. An administrator could change the structure of web pages or rewrite web content, based on results from this tool.
*Aspekt naukowy, problemowy pracy:	definition and implementation of a documents model, creation method for web crawler, implementation of a method for calculating document similarities.
Literatura	• Bird, S., Klein E., and Loper, E. (2009). Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit. O'Reilly Media. • Bengfort, B., Ojeda, T., Bilbro, R. (2018). Applied Text Analysis with Python: Enabling Language - Aware Data Products with Machine Learning, O'Reilly Media, 332 p. • Natural Language Toolkit, online: https://www.nltk.org/ • Torun H., Inner A.B. (2022). A Method for Similarity Detection in Vector Space by Summarizing News Articles. In. 30th Signal Processing and Communications Applications Conference, SIU 2022, • Mai G., Janowicz K., Yan B. (2018). Combining text embedding and knowledge graph embedding techniques for academic search engines. In. CEUR Workshop Proceedings, Vol. 2241, pp. 77 – 88

	• Giap Y.C., et. al. (2019). Implementation of the levenshtein distance method and similarities in checking the equal content of the document text. In. International Journal of Recent Technology and Engineering, Vol. 7 /6, pp. 38 – 43.
, **Oprogramowanie, język programowania, środowisko systemowe:	Jupyter Notebook Environment (Python)
**Środowisko uruchomieniowe:	
Dodatkowe wymagania i uwagi:	English language

UWAGA:
W polu literatura należy wskazać minimum 1 publikację z listy czasopism punktowanych wg wykazu MEiN z dnia 21 grudnia 2021 r związaną z proponowanym tematem pracy dyplomowej.

* Regulamin studiów § 36 2. Praca dyplomowa na profilu praktycznym, podobnie jak praca inżynierska, powinna mieć charakter aplikacyjny, badawczy, projektowy lub oceniający praktykę w świetle teorii.
** pola opcjonalne