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

rok akademicki 2025/26

Promotor:	**dr hab. Jozef Kapusta, prof. UKEN**
Temat pracy magisterskiej (j. polski, j.angielski):	Comparison of Methods for Finding Similar Content on a Web Portal *Porównanie metod wyszukiwania podobnych treści na portalu internetowym*
Zakres pracy i oczekiwane rezultaty praktyczne:	Document similarity (i.e., the distance between documents) is a core problem in Information Retrieval. Documents are considered similar when they are semantically close and describe related concepts; at the same time, similarity measures are widely used for detecting near-duplicates and redundant content. In large information systems and web portals, duplicated or highly overlapping texts can reduce search quality and complicate content management. Reliable similarity estimation is therefore important both for improving retrieval and for supporting content curation and deduplication workflows. The theoretical part of the thesis will provide a structured overview of approaches to representing text for similarity estimation, including classical and modern embedding methods (e.g., TF-IDF, Word2Vec, GloVe, FastText, contextual transformer embeddings such as BERT/SBERT). It will also summarize common similarity and distance metrics for document comparison (cosine similarity, Euclidean distance, Jaccard similarity, etc.), and discuss their assumptions, advantages, and limitations. The aim of the thesis is to experimentally compare selected document representations and similarity metrics on real pages from selected domain. The practical part will culminate in a case study demonstrating a simple tool for web portals administrators. The tool will calculate and show pages from the web portal with similar content. An administrator could change the structure of web pages or rewrite web content, based on results from this tool.
Aspekt naukowy, problemowy, innowacyjny pracy:	definition and implementation of a documents model, creation method for web crawler, implementation of a method for calculating document similarities.
*Oprogramowanie, język programowania, środowisko systemowe:	
*Środowisko uruchomieniowe	
Dodatkowe wymagania i uwagi:	English language
*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.

*pola opcjonalne