Zgłoszenie tematu **INŻYNIERSKIEJ** pracy dyplomowej

STUDIA I STOPNIA rok akademicki 2019/20

Promotor:	**dr hab. Jozef Kapusta, prof. UP**
Temat pracy dyplomowej (j. polski, j. angielski):	*Design of Tool for Searching Similarities in Web Content* *Projektowanie narzędzia do wyszukiwania podobieństw w treści stron WWW*
Zakres pracy i oczekiwane rezultaty praktyczne:	Document similarity (or distance between documents) is one of the central themes in Information Retrieval. How humans usually define how similar document look like? 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. 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 inżynierski*:	Creation a tool for web portals administrators, definition and implementation of a documents model, creation web crawler, implementation of a method for calculating document similarities.
Wymagane oprogramowanie/języki programowania**:	Jupyter Notebook Environment (Python)
Środowisko uruchomieniowe**:	Windows or Linux
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 Analy-

	sis with Python: Enabling Language - Aware Data Products with Machine Learning, O'Reilly Media, 332 p. • Natural Language Toolkit, online: https://www.nltk.org/

***należy uzasadnić/wskazać, czy praca spełnia wymagania inżynierskie**

**pola opcjonalne