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

na rok akademicki 2020/21

Promotor:	**dr hab. Jozef Kapusta, prof. UP**
Temat pracy magisterskiej (j. polski, j.angielski):	Documents Similarity based on Cluster Analysis *Podobieństwo dokumentów na podstawie analizy skupień*
Zakres pracy i oczekiwane rezultaty praktyczne:	Text clustering is an important application of data mining. It is concerned with grouping similar text documents together. The aim of the practical part is to design several models for clusterization using selected clustering techniques (k-means, k-medoids, etc.). The student will extract dataset from the web site with stories or speeches. The student will create vector representation of documents (TF-IDF, etc.) and apply clustering analysis on the data. The quality of the obtained models will be evaluated and compared.
Aspekt naukowy, problemowy, innowacyjny pracy:	Selection and implementation of natural language processing methods, implementation selected machine learning methods.
Oprogramowanie, język programowania, środowisko systemowe:	Jupyter Notebook environment (Python)
Środowisko uruchomieniowe	Windows or Linux
Dodatkowe wymagania i uwagi:	english language
Literatura:	1. Steven Bird, Ewan Klein, and Edward Loper: Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit. O'Reilly Media, 2009. 2. Benjamin Bengfort, Tony Ojeda, Rebecca Bilbro: Applied Text Analysis with Python: Enabling Language - Aware Data Products with Machine Learning, O'Reilly Media, 2018, 332 p. 3. UDPipe, online: http://lindat.mff.cuni.cz/services/udpipe/info.php 4. Natural Language Toolkit, online: https://www.nltk.org/ 5. scikit-learn: Machine Learning in Python, online: https://scikit-learn.org/stable/