

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

na rok akademicki 2023/24

|  |  |
|--|--|
| <b>Promotor:</b>   | <b>dr hab. Jozef Kapusta, prof. UP</b>   |
| Temat pracy magisterskiej (j. polski oraz j. angielski):       | Automatic Text Anonymization<br><i>Automatyczna anonimizacja tekstu</i>  |
| Zakres i oczekiwane rezultaty pracy:                           | <p>Text anonymization plays a significant role in today's world. When sharing documents that contain sensitive data, we can easily mask or anonymize this data and share the document without compromising privacy.</p> <p>In the theoretical part:<br/>The thesis focus on natural language processing (NLP). The most essential part of the theoretical part of the work will be the analysis of existing solutions focused on text anonymization.</p> <p>In the practical part:<br/>The main goal of this thesis is to create a tool that can process a document and automatically remove sensitive information from it (surname, company name, invoice amount, address, etc.). The created tool will use available libraries to recognize named entities in the text. Named entities will be replaced by placeholders, and the tool will store information about the document and placeholders. The prepared document can be safely shared. Subsequently, the tool can convert the data back into a document.</p>                          |
| *Aspekt naukowy, problemowy pracy:                             | implementation natural language processing methods, implementation machine learning methods.   |
| Literatura   | <ul style="list-style-type: none"> <li>• Bird, S., Klein E., and Loper, E. (2009). Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit. O'Reilly Media.</li> <li>• 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.</li> <li>• Natural Language Toolkit, online: <a href="https://www.nltk.org/">https://www.nltk.org/</a></li> <li>• Kapusta, J., Drlik, M., Munk, M. (2021). Using of N-grams From Morphological Tags for Fake News Classification. PEERJ COMPUTER SCIENCE, PeerJ, 7, e624.</li> <li>• Nagy, K., Kapusta, J. (2021). Improving fake news classification using dependency grammar. PLOS ONE, Public Library of Science, 16(9), e0256940</li> <li>• Zhong, XS., Cambria, E., Hussain, A. (2021). Does semantics aid syntax? An empirical study on named entity recognition and classification. NEURAL COMPUTING &amp; APPLICATIONS, 34/11, pp. 8373-8384</li> </ul> |
| , **Oprogramowanie, język programowania, środowisko systemowe: | Jupyter Notebook Environment (Python)  |

Zgłoszenie tematu pracy dyplomowej :: **STUDIA II STOPNIA** ::  
na rok akademicki 2023/24

---

|                              |                  |
|------------------------------|------------------|
| **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