MASTER'S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Gažo Alexander** Personal ID number: **495795**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Computer Science**

Study program: **Open Informatics**

Specialisation: **Artificial Intelligence**

II. Master's thesis details

Master's thesis title in English:

Algorithms for Document Retrieval in Czech Language Supporting Long Inputs

Master's thesis title in Czech:

Metody document retrieval nad českými texty vhodné pro zpracování dlouhých vstupů

Guidelines:

The task is to develop methods of document retrieval to be deployed in the fact-checking scenario. Focus on Czech corpora and the ability to deal with long input strings.
1) Familiarize yourself with methods of document retrieval aimed for the fact-checking task. Focus on the Czech language and neural network approaches. Aim for modern model architectures that are able to deal with long inputs (Longformer, Reformer, etc.)
2) Work with the Czech Wiki FEVER and ČTK data, both supplied by the supervisor.
3) Select an appropriate method and modify it for the supplied data. Consider student-teacher training based on models pretrained on different languages and training from scratch.
4) Evaluate and compare the methods.

Bibliography / sources:

[1] Thorne, James, et al. "FEVER: a large-scale dataset for fact extraction and verification." arXiv preprint arXiv:1803.05355 (2018).
[2] Thorne, James, et al. "The fact extraction and verification (fever) shared task." arXiv preprint arXiv:1811.10971 (2018).
[3] Chang, Wei-Cheng, et al. "Pre-training tasks for embedding-based large-scale retrieval." arXiv preprint arXiv:2002.03932 (2020).
[4] Beltagy, Iz, Matthew E. Peters, and Arman Cohan. "Longformer: The long-document transformer." arXiv preprint arXiv:2004.05150 (2020).
[5] Kitaev, Nikita, Łukasz Kaiser, and Anselm Levskaya. "Reformer: The efficient transformer." arXiv preprint arXiv:2001.04451 (2020).

Name and workplace of master's thesis supervisor:

Ing. Jan Drchal, Ph.D., Department of Theoretical Computer Science, FIT

Name and workplace of second master's thesis supervisor or consultant:

Date of master's thesis assignment: **21.02.2021** Deadline for master's thesis submission: _____

Assignment valid until: **19.02.2023**

Ing. Jan Drchal, Ph.D. Supervisor's signature	Head of department's signature	prof. Mgr. Petr Páta, Ph.D. Dean's signature

© ČVUT v Praze, Design: ČVUT v Praze, VIC

III. Assignment receipt

The student acknowledges that the master's thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the master's thesis, the author must state the names of consultants and include a list of references.

.	
Date of assignment receipt	Student's signature

© ČVUT v Praze, Design: ČVUT v Praze, VIC