**Text Mining Project - README.txt**

Susanna Maugeri (839365)

Veronica Grazia Morelli (839257)

June 2023

The delivered folder contains the following documents:

* *README.txt*, which is this file;
* *config.json*, it is the file which contains all the paths to the useful files that can be imported. Is it imported at the beginning of each script to import the corpus files.
* *Preprocessing.ipynb*, it is the script which takes the raw file of the corpus and preprocess them. They are then saved in new files.
* *Topic\_modeling\_part1.ipynb*, it is the code which contains the implementation of LSA, LDA with both Sklearn and Gensim and BErtopic.
* *Topic\_modeling\_parte2.ipynb*, it is the code which contains the implementation of Top2Vec.
* *TextSummarization.ipynb*, it is the code which contains the implementation of all the steps necessary to produce the text summarization process, including import, pre-processing, extractive and abstractive summarization methods. Moreover, the first section of this notebook contains the *first\_cleaning\_step()* function used to clean the CNN/Daily mail dataset before applying every method. This function is necessary to create the datasets used for both summarization and topic modeling task.

Topic modeling parts had to be divided in two different files for computational reasons: keeping everything in a single Google Colab file causated it to crush multiple times, while this way everything works fine.

The folder was shared with you, so you can now find it in the “Shared with me” section of Google Drive. The first thing to do is to add a shortcut to this folder in the main section of your Google Drive account (“My Drive”). This way the paths inserted in the code work without the need of changing them.

* For the Topic modeling part: When running *Topic\_modeling\_part1.ipynb*, at the beginning of the script there are some libraries that must be installed before running everything. Anyway, the system requires the runtime to be restarted after installing these libraries, so our advice is to run only this section at the beginning, to restart the runtime as requested by the system, and then to start running all the sections below. Everything that is run before restarting the runtime is lost when doing so.

Google Colab’s resources do not permit setting the number of documents to more than 10 thousands. Some of the methods fail when setting a higher number.

* For Text Summarization part: Unless otherwise specified in the notebook section all codes can be runned in Google Colaboratory platform. All notebooks all already setted to import the necessary packages and also in this way you can easily use a GPU! Don’t forget to change the directories.