Assignment - February 18th, 2021 - Text Mining & Sentiment Analysis
Module "Text Mining"
Master in Data Science & Economics[1]

Issued: Thursday, February 18th, 2021 **Due:** Sunday, February 28th, 2021

Part (a): (***Text data pre-processing***[2]).

****

Consider the corpus you choose. If it is the case, consider a subset of the corpus.

Task 1:

1 Clean the corpus by eliminating punctuation and stop words.

2 Tokenize it.

3 Try to obtain bi-grams.

Output type: Python code (preferably in a Jupyter notebook format).

Task 2:

1 Split the original corpus in sentences.

2 Vectorise it with bag-of words and TF-IDF methods.

3 Try to form a document-term matrix.

Output type: Python code (preferably in a Jupyter notebook format).

Task 3:

- Try to create a pipeline for implementing Task 1, parts 1 and 2 .

[1] This assignment is not mandatory. Must be done alone. If you pass this assignment, mark will range from 18 to 30. For the final exam mark you have the option to consider or not this assignment's mark. If you decide to opt for using this mark, an average with the final essay mark will be your final exam's mark. You can keep and use this assignment's mark in the first or in the second exam scheduled after the end of the course. All the developed Python code should be sent to **giancarlo.manzi@unimi.it** in **a single Jupyter notebook file**. Please provide your student ID number in your message. If you have any questions, contact the instructor for assistance.

[2] You can use any corpus for this assignment. Some suggestions are: 20 Newsgroup (`http://qwone.com/~jason/20Newsgroups/`, `https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html`), the Gutenberg data set (`https://web.eecs.umich.edu/~lahiri/gutenberg_dataset.html`), European language corpora (`https://www.linguistik.hu-berlin.de/en/institut-en/professuren-en/korpuslinguistik/links-en/korpora_links?set_language=en`). A comprehensive list of corpora can be found here: `https://guides.library.uq.edu.au/research-techniques/text-mining-analysis/language-corpora`.

Output type: Python code (preferably in Jupyter notebook format).

****

Part (b): (*Classification and clustering, topic model and summarisation*).

****

Consider the corpus you choose or another corpus suitable for the tasks included in this part[3] . If it is the case, consider a subset of the corpus. Exploit what you have done in Part (a).

Task 1:

- Perform classification and clustering and provide comments (within your Python code) on your results (commenting your code).

Output type: Python code (preferably in Jupyter notebook format).

Task 2:

- Perform topic model and provide comments on your results.

Output type: Python code (preferably in Jupyter notebook format).

Task 3:

- Perform summarisation and provide comments (within your Python code) on your results.

Output type: Python code (preferably in Jupyter notebook format).

****

[3] For example, for text classification you may use a corpus of messages / e-mails, etc., for topic models and summarisation a book paragraph, etc.