TM&S PROJECT INSTRUCTIONS

Proff. Gabriella Pasi
and Marco Viviani

Lab: Dr. Pranav Kasela

gabriella.pasi@unimib.it
marco.viviani@unimib.it
pranav.kasela@unimib.it

Tasks to be accomplished (1)

- **Text pre-processing**
  (text-representation-dependent, task-dependent):
  - Tokenization;
  - Normalization;
  - Stop-words removal;
  - Stemming/Lemmatization;

- **Text representation**
  - Choose suitable representation(s) and explain the rationale behind this choice.
    - BoW (binary, tf, tf-idf)
    - Word Embeddings (word2vec, Glove, etc.)
    - Contextualized Word Embeddings (BERT, …)

Tasks to be accomplished (2)

- **“Core” tasks** (please select TWO at your choice):
  - Text classification (e.g., with respect to different topics);
  - Text clustering;
  - Topic modeling;
  - Text summarization.

- The above-mentioned tasks must be performed on **suitable datasets**.
  - The same dataset can be used by AT MOST two groups.

Possible datasets for Text Classification

- **Different possibilities**:
  - Text Classification Dataset Repositories
  - Review Datasets
  - Online Content Evaluation Datasets
  - Sentiment Analysis Datasets

- You can have **access** to SOME of the above-mentioned datasets at the **following links**:
  - https://lionbridge.ai/datasets/14-best-text-classification-datasets-for-machine-learning/
  - https://analyticsindiamag.com/10-open-source-datasets-for-text-classification/

Possible datasets for Text Clustering

- Datasets employed for Text Classification can be also employed for **Text Clustering**.

- Other useful Datasets for Text Clustering:
  - https://archive.ics.uci.edu/ml/datasets.php?format=&task=clu&att=&area=&numAtt=&numIns=&type=text&sort=attUp&view=table
  - https://www.kaggle.com/snap/amazon-fine-food-reviews

Possible datasets for Topic Modeling

- Datasets employed for Text Classification and Text Clustering can also be used for **Topic Modeling**.

- Other useful Datasets for Topic Modeling:
  - https://github.com/nytimes/covid-19-data
  - https://catalog.ldc.upenn.edu/LDC2008T19
  - https://www.yelp.com/dataset/

Possible datasets for Text Summarization

- **CNN/Daily Mail**
  - The dataset contains online news articles paired with multi-sentence summaries
  - https://github.com/abisee/cnn-dailymail

- **Gigaworld**
  - The dataset represents a sentence summarization/headline generation task with very short input documents and summaries
  - https://drive.google.com/file/d/0B6N7tANPyVeBNmlSX19Ld2xDU1E/view

- **X-Sum**
  - Data is collected by harvesting online articles from the BBC. The idea of this dataset is to create a short, one sentence news summary. More suitable for abstractive summarization.
  - https://github.com/EdinburghNLP/XSum

Other datasets at your choice

- **Dataset described in scientific papers** used or generated specifically to solve text mining tasks.

- **Any other dataset** that may be of interest to you but has particular characteristics:
  - Constituted by **textual documents**.
  - Characterized by an **adequate number** of documents.
  - Possibility of **preprocessing** text.
    - Datasets that already provide the representation of the text after the preprocessing phases are not adequate.
  - **Adequacy** with respect to the **text mining task** to be addressed.
    - Independently from the considered task, it is necessary to have available or be able to easily generate a "ground truth" with respect to the task addressed to provide suitable evaluations.

Tasks to be accomplished (3)

- **Evaluation**
  - Provide suitable evaluation metrics, depending on the considered task.

- **Important**: the proposed datasets contain textual content that refers to **different contexts**. This has to be taken into account in the development of the project.
  - Sub-sets of the data within each dataset can be considered (e.g., text referring to a specific topic), by motivating this choice.

Other instructions (1)

- The work must be done in groups of 2 or 3 people **at most**.

- **Requirements**:
  - All must be written in ENGLISH.
  - Delivery of all the material (packages, libraries, etc.) necessary to run the developed project.
  - A README.txt document of the how-to install and run the project.
  - Source code.
  - A report describing the project, the implemented solutions, the evaluations.
  - A PowerPoint presentation of the project. There will be an oral presentation and a discussion.

- The programming languages to be used for the development of the project are **R** or **Python**.

Other instructions (2)

- All the material must be shared with both Prof. Gabriella Pasi, Prof. Marco Viviani, and Dr. Pranav Kasela at least **7 days before** the date of the written exam → **Google Drive folder**.

- The written examination and the project must be conducted in the **same examination session**.
  - If you do not pass the written examination, or if you intend to decline the grade, the mark taken in the project will be kept valid for the entire academic year.

Evaluation Dimensions

- The project will be **evaluated** against:
  - Clarity in:
    - the presentation of the problem;
    - the adequate choice and treatment of the dataset(s).
  - Correctness and completeness in:
    - the pre-processing and representation of the text (use of several techniques);
    - dealing with the considered text mining task(s);
    - the carried-out evaluations.
  - Adequacy of:
    - the report;
    - all material sent.

Evaluation Score

- The project will make it possible to obtain **from 0 to 4 points**. 4 points will be assigned only to particularly original projects.

- **Projects that will be better evaluated** in terms of scoring will be those that:
  - Propose non-discounted datasets and models;
  - Compare their models with any available models trained on the same dataset;
  - Will implement models described in scientific articles, but which do not have an implementation available on GitHub.

- These points will be **added** to the evaluation obtained in the written (theoretical) exam.
  - E.g., written exam: 25, project: 3 → Final score: 28/30
  - Praise (*lode*) is acquired with a total grade equal to or greater than 32/30 → 30 e lode

Filling in the Google Sheet

- Groups are requested to fill in a **Google Sheet**, indicating:
  - Surnames of group members, separated by commas
  - Project abstract
  - Dataset the group intends to use
    - Please note that the same dataset can be used by a maximum of two groups.

- **Link** to the Google Sheet:
  - https://docs.google.com/spreadsheets/d/1cNglMaEjdmzQI7jVy-aeFx4sKv9_Aq0A2FGp79y_Ba0/edit?usp=sharing