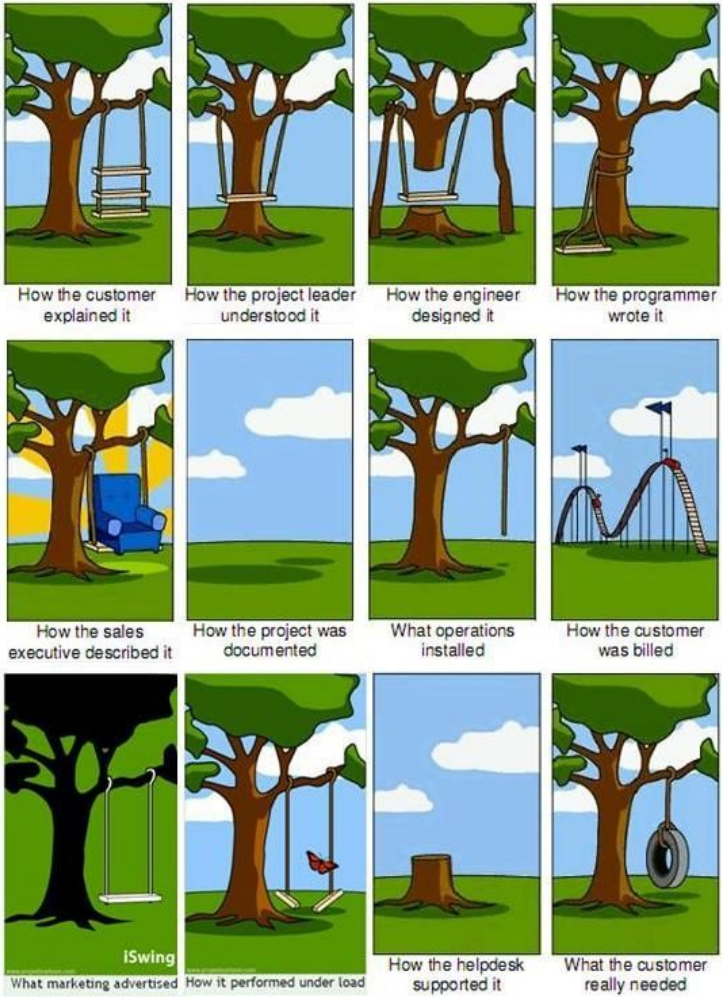
Text Mining and Natural Language Processing

Alessandro Raganato alessandro.raganato@unimib.it

The 12 Phases of Project Development



General Requirements

- The project can be done in groups of **up to 2 people (at most)**
- **General Requirements (more details in the next slides)**:
  - Delivery of all the material necessary to install and run the developed system:
    - A ReadMe Document (txt file) explaining the code and/or the installation procedure (if required);
    - Source Code (for Colab send the downloaded notebook);
    - External used resources (if used).
  - A detailed and concise report (A4 pdf file) describing the system and the choices undertaken.
  - A PowerPoint presentation illustrating the system and the choices undertaken. There will be an oral presentation and a discussion!
  - The presentation and report must be written in English.

General Requirements

- Don’t be afraid to make use of advanced techniques at any (or every) step of the project: they will evaluated positively!
- Any reasonable additional step done by your initiative will be evaluated positively!
- The projects must be submitted:

4 Days prior to the exam!

All the material must be uploaded through a Google form (check the e-learning website)
Moreover, please, send it also via email to this address (add in CC all members of the group): prof.alessandro.raganato@universitadipavia.it using the following subject:

[TM&NLP] Project submission 2023-2024 - ACRONYM OF YOUR PROJECT

General Requirements

- The written examination and the project must be conducted in the same examination session (for all group members).
- Please note that the same dataset can be used by a maximum of two groups
- The project, once submitted, **CANNOT** be re-done. Only in the case of a fail (grade 0), you can re-do it.
- Once graded, you can keep the score of the project until the last exam session of the current academic year, i.e. February 2025.
- The time slot for each project presentation will be of 10 minutes. All members of a group need to present in the same session and so they need to split the talk among themselves.
- Project slides can be uploaded later until the same day of the presentation session
- All Project materials (code, models, readme, report, etc.) must be in a compressed folder as one .zip archive. Its name must be ACRONYM OF YOUR PROJECT.zip

General Requirements

- **AI policies**:
  - Using assistance from AIs such as ChatGPT to complete your project is allowed.
  - **If you take advantage of any sort of AI assistance, you will be required to submit the specific prompts you used as well as a description of how the AI helped (or did not help) you complete the project.**
  - This should go without saying, but if you are using AI assistance, you are also responsible for making sure it is correct before submitting it.

Filling in the Google Sheet

- Groups are requested to fill in a **mandatory Google Sheet**, indicating:
  - **Ids** (matricola) of each group member, separated by commas
  - Project **acronym**
  - **Dataset(s)** 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 in the e-learning website

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 **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 3 points**.

- **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 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 31/30 → 30 e lode

Steps to be accomplished (part 1)

- Text pre-processing (task-dependent):
  - Tokenization;
  - Lemmatization;
  - Additional pre-processing operations can be implemented.

- Use of linguistic features (for example for analyzing the data)

- Text representation:
  - Choose suitable representation(s) and explain the rationale behind this choice.
    - Sparse representation (ppmi, tf-idf)
    - Dense word Embeddings (word2vec, Glove, fasttext, etc.)
    - Contextual Word Embeddings (Elmo, BERT, ...)
    - Large Language Models (LLMs)

Steps to be accomplished (part 2)

- "Core" task: **Text classification**

- Some available online resources:
  - https://pytorch.org/text/0.17.0/datasets.html#text-classification
  - https://huggingface.co/docs/datasets/en/index
  - https://www.kaggle.com/datasets?tags=13204-NLP
  - https://semeval.github.io/
  - https://www.ics.uci.edu/~smyth/courses/cs175/text_data_sets.html
  - https://paperswithcode.com/datasets?task=text-classification&mod=texts&lang=english&page=1
  - https://imerit.net/blog/17-best-text-classification-datasets-for-machine-learning-all-pbm/

Steps to be accomplished (part 2)

- Hint: look for datasets published with a paper

  For example in https://huggingface.co/datasets/ look for the citation information section:

Citation Information

```
@InProceedings{maas-EtAl:2011:ACL-HLT2011,
  author    = {Maas, Andrew L.  and  Daly, Raymond E.  and  Pham, Peter T.  and  Huang, Dan  and  Ng, Andr
  title     = {Learning Word Vectors for Sentiment Analysis},
  booktitle = {Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Hu
  month     = {June},
  year      = {2011},
  address   = {Portland, Oregon, USA},
  publisher = {Association for Computational Linguistics},
  pages     = {142--150},
  url       = {http://www.aclweb.org/anthology/P11-1015}
}
```