Instructions

Assignment 1: Sentiment classification (graded part)

General information

- The graded part of Assignment 1 will earn you **30%** of your overall course grade. You will get a grade between 0 and 10.
- **Groups:**
  - Apples will do the assignment on their own (plus maybe a kiwi; see below).
  - Oranges will do the assignment in pairs (plus maybe a kiwi; see below).
  - Kiwis will work with either an apple or a pair of oranges. They will learn as much as they can and contribute as much as they can (see below for more information).
  - Apples and oranges that work with kiwis will need to invest some time in helping them learn such that they can contribute to the assignment. This will be rewarded with +0.5 points in the grade for the assignment.
- Please communicate the groups to us (via email to lucas.weber@upf.edu) by **Wednesday October 13 midnight**. Please include the names and fruits of all the members of the group, and make sure to email us also if you do the assignment on your own.

Introduction: SemEval shared tasks

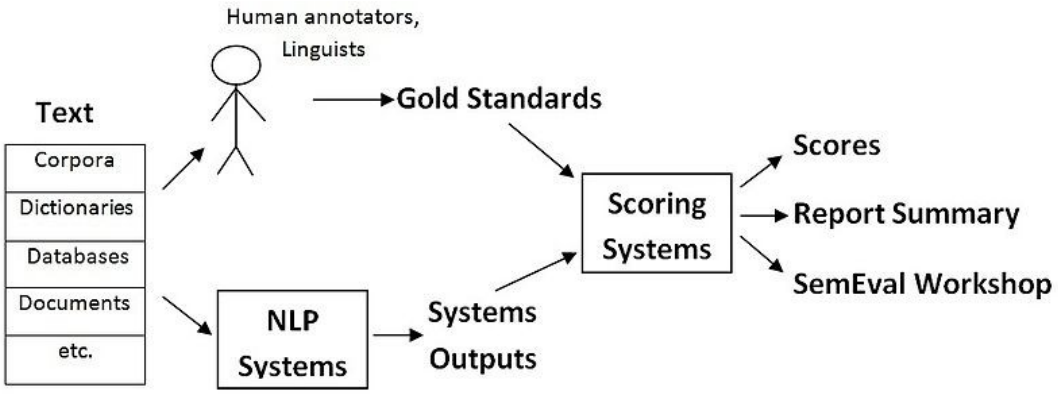
In the graded part of Assignment 1, we will emulate a **shared task** of the kind that are provided in the **SemEval competitions** hosted by the Association for Computational Linguistics (ACL; this is the main organization for CL and NLP).

- Background on SemEval: https://en.wikipedia.org/wiki/SemEval
- Tasks for SemEval 2022: https://semeval.github.io/SemEval2022/tasks
- Website of the ACL: https://www.aclweb.org/portal/

In what follows, we will describe the typical SemEval task; but bear in mind that the exact procedure and specifications can vary depending on the task and the organizing team.

Each task poses a challenge in computational semantics, such as identifying mysogynous memes, establishing the similarity of news articles in a multilingual setting, or generating dictionary definitions from word representations and viceversa (all these are example tasks from SemEval 2022, see link above). Tasks are proposed by members of the ACL community and evaluated by the SemEval organizers.

The selected tasks follow a process that can be schematized as follows:



(Figure credits: alvations via Wikipedia, CC-BY-SA 3.0)

First, the task organizers (there is a different organizing team for each task) create a resource containing semantic information, typically through a mix of semi-automatic selection of materials and manual annotation. An example could be the sentiment data that we have been working with. The organizers then publicly release this Gold Standard resource for the competition; but they initially only provide part of the data (typically, training and development sets). Teams interested in participating in the competition are given a few weeks to develop their system using the data released by the organizers. At a determined deadline, participants submit their best systems to the organizers, who then evaluate them on the test data. Often, a leaderboard is published with the scores of the best systems.

Participants write a short (4-page) article describing their system. Task organizers typically write an overview article explaining trends observed in the competition: overall performance, what kinds of systems tend to work best, what the remaining challenges are for solving the task, etc. Tasks, systems, and results are presented at the annual SemEval workshop, and articles are published in the proceedings of the workshop (see https://aclanthology.org/venues/semeval/).

Winning a SemEval competition is prestigious in the field; and the resources developed remain public, to be used as benchmarks for computational semantics beyond SemEval.

[Question (not for the assignment, but as food for thought): Why is SemEval a good idea?]

Our shared task: description and procedure

In the computational semantics course, our shared task for Assignment 1 will be **sentiment analysis for parts of movie reviews in English**, using the data that you have been working on during the first 3 weeks of the course.

- During **Week 4**, you will develop the best classifier you can, using the training and development (validation) data available on AulaGlobal. Then, by **Tuesday October 19 at 23:59** (Barcelona time), you will submit the system and we will evaluate it on the test data. We will provide you with the accuracy on the test data, your system's predictions, and the gold data (including the texts). Depending on the results, we will consider publishing a leaderboard with the scores of the best systems.
    - We provide the file *sentiment_classifier.py* on AulaGlobal with a bit of code and further instructions.
    - For submission instructions, see AulaGlobal.
- During **Week 5** of the course, you will do error analysis or your system, using the test data, and write a report encompassing the system construction and the

evaluation and error analysis of your system. The deadline for the report will be **Tuesday October 26 at 23:59** (Barcelona time).

- All you need to submit is the report, in a PDF file.
- We are providing a **template** for the report (see the folder for Assignment 1 on Aula Global), please follow it, either using it directly, or reproducing the content and format in the editor of your choice.

Notes

- We expect clean code (see the instructions for the non-graded part for tips).
- For the classifier,
  - you can intervene in all the steps of the process: text pre-processing, feature definition and extraction, and creation of a rule-based or Machine Learning classifier;
  - apart from the material that we have seen in class, you can consult whichever additional source of information you want, such as other parts of the textbook or the internet;
  - you can use Machine Learning algorithms other than Logistic Regression if you want, **provided that you understand how they work**;
  - we expect you to go through several rounds of improvement, with the cycle that we have practiced during the assignment: develop classifier based on training data => test on validation data => do errror analysis => improve classifier.
- For error analysis,
  - we expect that you will acquire insights into the strengths and weaknesses of your classifier, and in particular into the kinds of semantic phenomena that are troublesome for it;
  - and we expect you to use the tools that we have practiced in class to do so: evaluation scores, confusion matrix, targeted inspection of true/false positive and true/false negative cases, analysis of logistic regression coefficients (if applicable). You can use additional methods if you want.