Project Instructions
Machine Learning for Natural Language Processing

March 10, 2021

Here are all the instructions needed for the final evaluated project for the Machine Learning for Natural Language Processing course (ENSAE 2021).

Important Dates

March 17th 8pm	March 22nd 8pm	April 24th 8pm
Registration deadline	Project proposal deadline	Final Report deadline
Register yourself and your group here	Mail to your lab instructor a detailed project proposal .	Send the 2 links to your lab instructor by email (git link and colab link).

1. Before **March 17th 8pm:** Register yourself in https://docs.google.com/spreadsheets/d/1arvMCFMjI7Bu-Ya8T3L8tq8x6v70ei1IHrTY22az-yU/edit?usp=sharing
2. Before **March 22nd 8pm ,** mail to your lab instructor a **detailed project proposal** providing details on your project (dataset, modelling approach…).
3. Before **April 24th 8pm** send to **benjamin.muller@inria.fr** by email with the object **[ML FOR NLP final project 2021]** the TWO following links
    a. Github/Gitlab link which includes **your final report (2 pages max) as a PDF file** (based on the latek template provided) **AND** your uploaded google colab notebook as a **ipynb file.**
    b. Google Colab link which points to your google colab notebook

    NB: one email per group cc-ing your partner is required

NB: the notebook is “submitted” two times, first in the git project, second in the google colab

Instructions

- Project Content Instructions

1. You must frame your project yourself

➔ Define the context in which you want your project to be

➔ Frame and Write down one or several key questions that you'll try to answer based on the data you will choose and the experiments you will run.

➔ Define and explain a clear experiment protocol to answer those questions (what techniques you will use, based on what tasks, what models, what preprocessing, what training, what evaluation you intend to do)

2. You can choose among those datasets or pick a dataset of your choice (after validation from your instructors)

➔ Universal Dependencies: POS tagging,

➔ Topic: New York Times Comments (1Gb) : https://www.kaggle.com/aashita/nyt-comments

➔ Yelp: https://www.yelp.com/dataset/documentation/main

➔ Name Entity Recognition

➔ Sentiment analysis :

➔ https://www.kaggle.com/sid321axn/amazon-alexa-reviews Sentiment analysis (multidomain): http://jmcauley.ucsd.edu/data/amazon/

➔ Sequence Classification (french official jobs' taxonomy): https://www.data.gouv.fr/fr/datasets/repertoire-operationnel-des-metiers-et-des-emplois-rome/ (Mostly for the "Arborescence principale" file)

➔ Language Modelling/Sequence Generation:

➔ https://blog.einstein.ai/the-wikitext-long-term-dependency-language-modeling-dataset/

➔ A small list of well-known standard datasets for common NLP tasks (Speech Recognition and Image Captioning excluded!): https://machinelearningmastery.com/datasets-natural-language-processing/

➔ An alphabetical list of free or public domain text datasets: https://github.com/niderhoff/nlp-datasets

NB: If you want to use your own datasets, you should send it to benjamin.muller@inria.fr AND gael.guibon@gmail.com for validation.

3. You must use one or several of these word/sentence embedding techniques :

➔ Skip-gram word2vec, Fastext (bag-of-n-grams), Glove, ELMo, BERT (Roberta, CamemBERT, …)

4. You must build a task-specific model for performing one of the three types of tasks among *Sequence Labelling, Sequence Classification, Sequence Generation* (e.g. POS tagging, NER, sentiment analysis, intent detection, translation, QA…).

NB: One of these approaches should be based on embedding vectors.

5. You must evaluate your model with both **qualitative analysis** and **a quantitative analysis**

6. Present your experiments and results in a synthetic way. Draw future interesting work to do following your project.

● Project Example

Here is a project example to make the instructions more concrete. You should not use this project but you can take inspiration from it.

1. *Twitter has now an ubiquitous impact on politics. More specifically, the former President of the US, Donald Trump, used it as his main communication channel with American people as with other foreign politicians. Our hypothesis is that the sentiments expressed in those tweets were likely to be predictive of many phenomenons in the world (geopolitical interactions, financial markets, …) . Therefore, this project aims to extract quantitative metrics from President Trump's tweets, specifically based on sentiments.*
2. *We will use President Trump's tweets from https://www.kaggle.com/liston/sentiment-analysis-of-donald-trumps-tweets as well as the sentiment analysis treebank SST*
3. *We will use Fastext vector in English https://fasttext.cc/docs/en/english-vectors.html as well as English version of BERT https://github.com/google-research/bert with Hugging Face library (https://huggingface.co/transformers/)*
4. *We will first explore the database.*
   Then we will build a sentiment analysis model based on BERT.
   We will test it qualitatively with some predicted examples. We will then evaluate it quantitatively using the F1 score computed on the test set of the source dataset;
5. *We will then draw conclusions and possible future work*

● Report Instructions

➔ Write and present clearly and synthetically your project
➔ **2 pages** maximum (*any group breaking this rule will be penalized*) in **English** only
➔ You must point to your notebook (colab or github/gitlab (e.g. with footnote \footnote{https://...}) as well as your github project)
➔ You must use this latex template from overleaf (to copy) https://www.overleaf.com/read/hdcgdbhmcxxr
➔ Add all the references in Appendix (use .bib \cite{10test})
➔ You can add any non-textual content (plots, table, images, schemes...)
➔ You can use any concepts and methodology studied during the lectures and the lab sessions.
➔ You can reuse any pieces of code from the lab session (though we will judge your project based on its originality also)

- Code Instructions

➔ Your code should be a Google Colab jupyter notebook
➔ Your notebook should **be self-contained**
  ◆ This means we should be able to run all the experiments presented in the report from your notebook (see for instance the "! wget" code cells from the labs to programmatically download file from a git repository)
  ◆ This does not mean all the code must be in the notebook. But all the code useful to demonstrate your work should **be runnable from the notebook**
➔ Your code should be based in python
➔ You can re-use any code used during the labs

- Evaluation

You will be evaluated on :

- Respecting all the **instructions** detailed in this document
- Your **scientific approach:**
  - Able to frame a problem, ask relevant questions which includes textual data
  - Able to answer fully or partly this problem using NLP techniques seen during the course
  - Present everything in a reproducible (code) and convincing (report) way
- The **originality** of your project

Resources

Embeddings:

- Fastext (Skip-Gram or CBOW-ngram) https://fasttext.cc/docs/en/crawl-vectors.html
- Glove https://nlp.stanford.edu/projects/glove/
- BERT https://huggingface.co/transformers/pretrained_models.html
- ELMo https://allenai.github.io/allennlp-docs/tutorials/how_to/elmo/

Pretrained models — transformers 2.8.0 documentation

Features:

- Wordnet (en) https://pythonprogramming.net/wordnet-nltk-tutorial/
- Lexicon (POS and morphological features) : http://atoll.inria.fr/~sagot/UDLexicons.0.2.zip

Logistics

- ➔ **2 students per group**
- ➔ **1 *google colab notebook* + 1 report per group uploaded to github or gitlab.**
- ➔ **Due date is April 24th 2021**

Submission Process

4. Before **March 17th 8pm:** Register yourself in https://docs.google.com/spreadsheets/d/1arvMCFMjI7Bu-Ya8T3L8tq8x6v70ei1IHrTY22az-yU/edit?usp=sharing
5. Before **March 22nd 8pm ,** mail **to your lab instructor** a **detailed project proposal** providing details on your project (dataset, modelling approach…).
6. Before **April 24th 8pm** send to **benjamin.muller@inria.fr** by email the TWO following links with the object: **[ML FOR NLP final project 2021]**
    a. Github/Gitlab link which includes **your final report (2 pages max) as a PDF file** (based on the latek template provided) **AND** your uploaded google colab notebook as a **ipynb file.**
    b. Google Colab link which points to your google colab notebook

NB: the notebook is “submitted” two times, first in the git project, second in the google colab

Github Help
https://reproducible-science-curriculum.github.io/sharing-RR-Jupyter/01-sharing-github/