Hmw II : Spark, NLP, Tables and visualization

2025-05-11

- Due : May 29, 2025
- Work in pairs
- Deliver your work as a `qmd` file through a github repository
- Use the `quarto` package for reproducible research
- Use `pyspark` or `sparlyr`
- Use `spark-nlp` for text annotation
- The report should be rendered at least in HTML format, and possibly also in PDF format

Objectives

This homework is an opportunity to use `pyspark`/`sparlyr`/`spark-nlp`.

- Extract/Load/Transform the Balzac corpus
- Annotate the corpus with `spark-nlp`
- Perform Stylometric Analysis and Visualize the results using either `plotly` or `altair`
- Design a way to store results using `parquet` files. Motivate your solution.

Compare annotations from Spark NLP and annotations from Spacy (usability, agreement).

In Stylometric analysis, you should at least

- Compute Flesch-Kincaid and Kandel-Moles readability indices and design a visualization
- Compute *sliding* readability indices over sliding windows defined by different window sizes. How stable are readability indices?
- Display Zipf plots for the different documents
- Segment the different texts into *dialog* and *narration* parts.

Tune your spark session so as to minimize shuffles, use multi-core architectures as much as possible.

👉 Your deliverable shall consist in a `qmd` file that can be rendered in HTML format.

You shall describe the downloaded data.

Plots shall be endowed with titles, legends and captions,

Data, NLP pipelines and graphical pipelines shall be given in an appendix.

Data

Data can be downloaded/scrapped from different sources

- https://github.com/dh-trier/balzac/tree/master (not complete)
- https://www.gutenberg.org/ (17 volumes of *Comédie humaines*)
- …

👉 Table wrangling should be performed using `Spark Dataframes` Dataframes.

👉 Your extraction (ELT) pipeline shall be *reproducible* and shall be given and motivated in an appendix.

You are not supposed to deliver the text files as a zipped archive.

👉 Annotate with `soark-nlp`

👉 Annotation shall be done on a per novel basis. It should be performed in a parallel (and distributed) way.

👉 Graphical pipelines should be reproducible and shall be given in an appendix.

👉 Keep the downloaded data in a separate subdirectory. Your working directory (working tree) should look like something like that:

```
.
   .git/
   DATA/

|   :
   _extensions/
   _outdir/
   _metadata.yml
   _quarto.yml
   our_report.qmd
   :
   README.md
```

Report organization

The first part (introduction) of the report shall be dedicated to the description of the data to be extracted and to the extraction pipeline (not different from Homework I).

The second part of the report shall be dedicated to the description of load/transform pipeline.

The third part of the report shall be dedicated to the description of the annotation pipeline

The fourth part of the report shall be dedicated to the stylometric analysis: which questions did you pick up (and why?), plots, summary tables and comments. Refrain from overplaying your hand: yours plots are not likely to provide a new literary interpretation of Balzac opera. Comment the data, all the data, and nothing but the data.

The fifth part is the appendix. The first four parts should be mostly text and plots. The fifth part should be code only.

The appendix shall be dedicated to the details of the pipelines. You shall give the code.

You shall also give the code of the graphical pipelines in the appendix.

You shall avoid copy-paste coding. Don't Repeat Yourself. `knitr` provide the tools to organize the Quarto file so that you can write your code once and use it many times, once for data wrangling and plotting (without echoing), then for listing and explanation.

> **Organizing a report using the `jupyter` engine**

References

- Data Humanities with R
- Spacy
- Spark NLP
- `scrapy`

- Computational stylistics

Data sources

- `Project Gutenberg`: you can find the `La Comédie Humaine` using a simple search. All volumes can be downloaded as text files from there.

Grading criteria

Criterion	Points	Details
Narrative, spelling and syntax	20%	English/French
Plots correction	15%	choice of `aesthetics`, `geom`, `scale` …
Plot style	10%	Titles, legends, labels, breaks …
ETL	20%	ETL, SQL like manipulations
Annotation	10%	Annotations
Computing Statistics	5%	…
DRY compliance	20%	DRY principle at W Wikipedia