# Natural Language

Practical Classes

Luísa Coheur

2025

# P1

## Dealing with Natural Language

or *How did I become a detective*

Image generated by ChatGPT

- **Summary**:
  - Dealing with unstructured data (natural language)
  - Basic Unix commands
  - Extracting basic information from data
- **Operational objectives**:
  - Make students aware that we are dealing with unstructured data (challenges ahead!)
  - Understand that (simple) Unix commands can be very useful when processing text
  - Bring to memory the following commands: less/cat, wc (ahahaha!), grep, head/tail, sort/uniq, cut, tr,... (use “man command” to have more information about them)
  - Understand the importance of properly analysing data and results
  - Have great fun with a “Friends” corpus (this sounds weird)
- **This class needs**: a computer and a terminal where you can run Unix commands
- **Class material**: “Friends” dataset (adapted from Kaggle[1])

[1] There are many interesting datasets on Kaggle`https://www.kaggle.com/`

# An unexpected request

You receive a strange email:

> Hi,
>
> I'm a dedicated fan of the series F·R·I·E·N·D·S, and I'm currently attempting to collect funds to raise a statue in honor of Monica, my favorite character, in my small hometown. I was close to achieving this goal when the despicable Mayor, Miss Early, made insulting remarks about the show. Specifically, she stated (and I quote):
>
> *It makes no sense to have a statue of Monica or any of the other friend. For several reasons: a) The cast is a group of egocentric and selfish people where the most frequently used word is 'me'; b) Monica, in particular, dominates conversations excessively; c) Monica is also the most sarcastic character, making her presence intolerable.*
>
> Regrettably, I'm unable to respond to her criticisms, supported by scientific evidence. I know that you are an expert in NLP. Could you assist me with this matter? I am prepared to compensate you for your help.
>
> Sincerely yours,
>
> Mary Moniac

Hmm... The quarter has just started and you don't feel like an expert in NLP yet, but you think you can manage this... How difficult can it be?

# Hands on

You found a dataset on Kaggle containing the dialogues from Friends. You name this dataset as P1_dataset_friends.txt. You decide to use Unix to get the answers you search for.

## 1. Basic characterisation of the dataset

In the theoretical classes, your professor talked a lot about datasets and how important it was to gave them a look before processing. Perhaps you should begin with a basic analysis.

- How many lines and words can be found in the whole dataset?
  *Tip*: use the wc command (again: this sounds weird). What is/should be considered a word? (see what happens with echo "it's ok" | wc -w. Why?)

- How many main characters (personas) are there in the show?
  *Tip*: capture them with cut -f1 (first column) and then use sort and uniq in a pipeline (attention: uniq should be applied after sort; -f applied to sort ignore case).
  Look at the results! Was this enough to get the *EXACT* number of characters?

## 2. Let us examine the assertions

**Is "me" the most frequent word used by the entire cast?**

- Find the frequency of each word in the whole document.
  *Tips*: a) put every word in a single line (use: tr -sc 'A-Za-z' '\n' < friends.txt | less). tr is to translate characters; c is for all the characters that are not (A-Za-z); s merges newlines); b) play with uniq -ci (c for counts and i for ignore-case) and sort.

- Find the 10 most frequent words in the series dialogues. *Tip*: use head (or tail).

Let's now talk about Monica ...

**Does Monica dominate the conversations excessively?**

- Extract all lines with the word Monica. *Tip*: use grep. Are those Monica's dialog lines?
- How many lines contain the word Monica? *Tip*: use a pipeline with grep and wc.
- Extract all Monica's dialogue lines.
  *Tip*: use grep with a regular expression (we will dedicate more time to regular expressions in a near lab). For now just use ^ to signal the sentence start. Use > to redirect those lines to a file. grep -e allows you to extract patterns (probably don't need it now).
- Consider the 6 main characters of the series (Monica, Rachel, Ross, Phoebe, Joey or Chandler), who has more dialogue lines? *Tip*: use uniq -c (counts)

**Is Monica the most sarcastic character?**

- You look at the dataset again and you find/confirm that there are some actions in parentheses "()". *Tip*: grep -o prints only the matching part of the lines.
- Find the most sarcastic friend[2].

## Your answer

Based on your acquired results, you reply to Mary Moniac. Now you have the answers to:

- Is "me" the most frequent word used by the entire cast?
- Does Monica dominate the conversations excessively?
- Is Monica the most sarcastic character?

Next day, you receive an email from Mary Moniac:

> "You are amazing! The despicable Mayor, Miss Early, had to publicly apologize after I published your study in a local newspaper. As I promised, I am prepared to compensate you: I will send you the entire F·R·I·E·N·D·S series. I only have it in VHS. Hope this is not a problem for you. Thank you so much for contributing to Monica's statute!"

Ohhh...

## A major decision

That night, in your bed, this episode makes you think:

> "Maybe I can use my knowledge in NLP and really win some money. And if I create a company that solves NLP cases? Not a company that develops NLP applications (there are so many), but one that uses expertise in NLP to solve NLP challenges? After attending the NL course and paying lots of attention to all classes I will be an expert! And even along the way... my knowledge about the field is improving everyday. So, why not?"

Before you fall into sleep, a name for your future business comes to your mind:

`NLP Detectivezzzzzzzzzz`.

[2] If you watched the series, who do you think they will be?