TÉCNICO LISBOA

Natural Language Lab
Tagus and Alameda
(2021)

Goals of this lab:

- Make students aware of some challenges related with the fact that we are dealing with unstructured data.
- Bring to memory that (simple) Unix commands – less/cat, wc, grep, head/tail, sort/uniq, cut, tr,... (use "man command" to have more information about them) – can be very useful when processing text.
- Understand the importance of properly analysing results.
- Have great fun with a Friends corpus[1] (this sounds weird): find out who is the top friend, the most sarcastic one, which are Rachel's most used words, etc. (just by using Unix).

Using Unix, please find:

1. How many lines and words can be found in the whole document? Tip: use the wc (again: this sounds weird). Question: what is/should be considered a word?

2. Extract all lines with the word "Monica". Tip: use grep. Question: are those Monica's dialog lines?

3. How many lines with the word "Monica" can be found? Tip: use a pipeline with grep and wc.

4. Extract all Monica's dialogue lines. Tip: use grep with a regular ex- pression; use ">" if you want to put those lines in a file. Tip: grep -e allows you to extract patterns (probably don't need it now).

5. Find all the actions within this file (assume that actions appear between "()"). Tip: grep -o prints only the matching part of the lines..

6. Find the most sarcastic friend and the one that sings more often.

[1] friends.txt, adapted from Kaggle at https://www.kaggle.com/ryanstonebraker/ friends-transcript

7. How many characters are there in this file? Tip: capture the characters with cut (first column) and then use sort and uniq (attention: uniq should be applied after sort; -f applied to sort uniq ignore case). Look at results. Was this enough?

8. Who is the top star? That is, who, between the 6 main characters of the series (Monica, Rachel, Ross, Phoebe, Joey or Chandler) has more dialogue lines? Tip: use uniq -c (counts)

9. Find the frequency of each word in the whole document. Tip:
   • put every word in a line (use: tr -sc 'A-Za-z' '\n' < friendsLN.txt | less). tr: translate characters; c is for all the characters that are not (A-Za-z); s merge newlines);
   • play with uniq -c (counts) and sort.

10. Find the 20 most frequent words in the series dialogues. Tip: use head (or tail).

11. Find the top ten words for each one of the main characters (you can try to understand who has the biggest ego, that is the one that uses "I" more often). You can do it one by one.