Lab Sheet 1

NB. In our

- In our practical work, we use free and open-source tools and software;
- We use Linux shells to run our tasks;
- We use also Java and Python programming languages;
- Under Python, we need some packages and modules which will be asked gradually;
- The textual corpora and data will be in English and Arabic;
- The encoding of texts will be in utf8;
- When developing in Python, we prefer use simple Python IDE such as IDLE.

Task 1. Using the following Linux shell commands open and display the downloaded corpus.
The commands

- cat: Concatenate and display files: cat file1.txt file2.txt
- sort: Sort lines of text file: sort file.txt
- uniq: Remove duplicate lines from a sorted file: sort file.txt | uniq
- grep: Search for patterns in files: grep "pattern" file.txt
- cut: Extract columns of text from files: cut -f1,3 file.txt
- sed: Stream editor for filtering and transforming text: sed 's/old/new/' file.txt
- wc: Count lines, words, and characters in a file: wc file.txt
- nl: Number lines in a file: nl file.txt
- iconv: Convert character encoding of a file: iconv -f utf-8 -t iso-8859-1 file.txt
- dos2unix: Convert DOS line endings to UNIX line endings: dos2unix file.txt
- rev: Reverse lines of a file: rev file.txt

Task 2. Apply the regular expressions on the associated corpus.
Write the regular expression (RegEx) that matches the following patterns :

- the pattern Samsung,
- the pattern Samsung Electronics,
- the pattern Samsung or samsung,
- the pattern بيانات,
- all variants of the previous pattern,
- isolated and concatenative conjunction particles
- all imperfective verbs
- all 1st person imperfective verbs
- all plural imperfective verbs conjuncted with و