Correction test mi-semestre

1. (large) Collection of machine-readable texts compiled with a purpose. These texts tend to be naturally occurring.

2. Allow for verifiable results when doing NLP experiments and provide a source for NLP applications to be trained on, given their size and the fact that the language used is representative of natural language (vs. synthetic linguistic data).

3. Text classification, named entity recognition, natural language generation.

4. a) mkdir MyProject
   b) cd MyProject
   c) touch NLP_task.py

5. A virtual environment allows the user to install packages and dependencies for their project in a secluded environment — meaning it won't be affected by any library updates. So launching the program later on, once newer python versions exist for example and your system version of python has been updated, will not be an issue.

6. A token is a basic unit used to analyze sentences: a sentence is split with regard to any spaces or separating characters such as \n. It usually includes both words and punctuation.

7.
```
def tokenize(sentence):
        return sentence.split()
```

8.
```
with open(‘MyCorpus.txt’, ‘r’) as file:
        my_corpus = file.read()        or        my_corpus = file.readlines()
```

9. a.1/6
   b.2/6
   c.3/6

10. a.4/52
    b.13/52
    c.1/52
    d.16/52 => P(jack) + P(hearts) – P(jack and hearts)