Information Retrieval 2020/21

Assignment 2

Task 1 – Individual work

Install the *langdetect* library in your Anaconda environment

https://pypi.org/project/langdetect/

and test it for German, Italian and English sentences. You do not need to submit your Jupyter notebook.

Task 2 – Group work

Create a Jupyter notebook that reads a corpus consisting of 2249 short news articles (AssociatedPress.txt) and computes basic statistics of the corpus like on slide 13 (03_slides_dictionary). See https://www.nltk.org/book/ as a reference.

```
#read textual documents from file

documents_path = './data/AssociatedPress.txt'

with open(documents_path, 'r', encoding='utf-8') as doc_f:

   corpus_list = doc_f.readlines()
```

Concretely, tokenize documents (use nltk.tokenize) and compute:

- Number of unfiltered (distinct) terms
- Distinct terms without numbers
- Distinct terms after case folding
- Distinct terms after removing English stopwords (from nltk.corpus)
- Distinct terms after stemming (from nltk.stem.porter)
- Compute the frequencies of distinct terms after case folding

Submit your notebook via OLE and explain the difficulties/issues you encountered.