Information Retrieval 2020/21

Assignment 3

Task 1 – Group work

Stay with the corpus consisting of 2249 short news articles (AssociatedPress.txt) from Assignment 2 and install the full text indexing library Whoosh https://pypi.org/project/Whoosh/.

See the documentation at https://whoosh.readthedocs.io/en/latest/intro.html as a reference.

Designing a schema:

For the given corpus you only have to index a single field, which is the body text ans already a predefined field type in Whoosh: whoosh.fields.TEXT

Since the TEXT is your only field use TEXT(stored=True) to specify that text should be stored in the index and TEXT(phrase=True) to allow searching for phrases.

```
from whoosh.fields import Schema, TEXT

schema = Schema(content=TEXT)
TEXT(phrase=True)
TEXT(stored=True)
```

Create an index and Import the corpus

```
import os.path
from whoosh.index import create_in

if not os.path.exists("index"):
    os.mkdir("index")
ix = create_in("index", schema)
```

```
writer = ix.writer()

# read textual documents from file
documents_path = './AssociatedPress.txt'
with open(documents_path, 'r', encoding='utf-8') as doc_f:
    corpus_list = doc_f.readlines()

# index documents
for x in corpus_list:
    writer.add_document(content= x)

writer.commit()
```

Now explore the capabilities of the library by exploring the core concepts learned up to now like different analyzers for indexing such as adding the lowercase filter, running different queries like standard queries vs. phrase queries and comparing results for different scoring and ranking functions.

Concretely try:

- Index the corpus using a lowercase filter
- Search for “Michael Dukakis”, “Dukakis OR Bush” (Boolean) and “graduate of Syracuse University” (phrase) and output results
- **Bonus: provide an example** for the use of an additional feature such as indexing n-grams, apply different analyzers (lemmatization, stemming) or pull suggestions for spell corrections **of your choice**.

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