assignment3

October 19, 2023

[ ]:
```
import random; random.seed(123)
import codecs
import string
import gensim
from nltk.stem.porter import PorterStemmer
```

Task 1.1 & 1.2 Henter inn dokumentet og splitter til en liste

[ ]:
```
f = codecs.open("pg3300.txt", "r", "utf-8")
paragraphs = f.read().split("\r\n\r\n")
```

Task 1.3 Fjerner "gutenberg"

[ ]:
```
paragraphs_sans_gutenberg = []
for i in range(len(paragraphs)-1, -1, -1):
    if not "gutenberg" in paragraphs[i].lower():
        paragraphs_sans_gutenberg.append(paragraphs[i])

paragraphs_sans_gutenberg.reverse()
```

Task 1.4 Splitter paragrafene til ord

[ ]:
```
words = []
for paragraph in paragraphs_sans_gutenberg:
    words.append(paragraph.split())
```

Task 1.5 & 1.6 Stemmer ordene og fjerner tegnsetting

[ ]:
```
stemmer = PorterStemmer()
for i in range(len(words)):
    for j in range(len(words[i])):
        words[i][j] = words[i][j].strip(string.punctuation+"\n\r\t\ufeff").
 ↪lower()
        words[i][j] = stemmer.stem(words[i][j])
```

Task 2.1 Lager en ordliste med ord og og stoppord

```
[ ]: dictionary_words = gensim.corpora.Dictionary(words)
stop_words = codecs.open("common-english-words.csv", "r", "utf-8").read().
 ↪split(",")
stop_words_ids = []
for stop_word in stop_words:
    try:
        stop_words_ids.append(dictionary_words.token2id[stop_word])
    except:
        pass
```

Task 2.2 Fjerner stoppord og lager corpus

```
[ ]: dictionary_words.filter_tokens(stop_words_ids)
corpus = [dictionary_words.doc2bow(word) for word in words]
```

Task 3

```
[ ]: tfidf_model = gensim.models.TfidfModel(corpus)
tfidf_corpus = tfidf_model[corpus]

matrix_similarities = gensim.similarities.MatrixSimilarity(tfidf_corpus)

lsi_model = gensim.models.LsiModel(tfidf_corpus, id2word=dictionary_words,␣
 ↪num_topics=100)
lsi_corpus = lsi_model[tfidf_corpus]
lsi_index = gensim.similarities.MatrixSimilarity(lsi_corpus)

print(lsi_model.show_topics(num_topics=3))
```

```
[(0, '0.180*"thi" + 0.167*"those" + 0.166*"countri" + 0.162*"upon" +
0.156*"price" + 0.155*"hi" + 0.151*"more" + 0.150*"wa" + 0.147*"part" +
0.145*"great"'), (1, '-0.777*"0" + -0.291*"2" + -0.286*"1" + -0.237*"8" +
-0.197*"4" + -0.152*"6" + -0.114*"10" + -0.100*"barrel" + -0.093*"£" +
-0.092*"3"'), (2, '0.687*"chapter" + 0.223*"divis" + 0.223*"iv" + 0.204*"v" +
0.191*"iii" + 0.183*"ii" + 0.159*"stock" + 0.150*"labour" + 0.116*"book" +
0.110*"system"')]
```

Task 4.1 & 4.2 Lager spørring og prosesserer den

Lager BoW og TF-IDF representasjoner av spørringen

```
[ ]: def preprocessing(text):
    stop_words = codecs.open("common-english-words.csv", "r", "utf-8").read().
 ↪split(",")
    stemmer = PorterStemmer()
    tokens = gensim.utils.tokenize(text)
    tokens = [stemmer.stem(token.strip(string.punctuation)) for token in tokens]
    tokens = [token for token in tokens if token not in stop_words]
    return tokens
```

```
query = preprocessing("What is the function of money?")
# query = preprocessing("How taxes influence Economics?")
print("Query:", query)
bow_query = dictionary_words.doc2bow(query)
tfidf_query = tfidf_model[bow_query]

for i in range(len(tfidf_query)):
    print(f"{dictionary_words[bow_query[i][0]]}: {tfidf_query[i][1]:.3f}",␣
 ↪end=", ")
```

```
Query: ['function', 'money']
money: 0.352, function: 0.936,
```

Task 4.3 Finner de 3 mest relevante dokumentene

[ ]:
```
doc2similarity = sorted(enumerate(matrix_similarities[tfidf_query]), key=lambda␣
 ↪kv: -kv[1])[:3]

for doc_id, sim in doc2similarity:
    print(f"[paragraph: {doc_id}] [similarity: {sim:.3f}]")
    print("\n".join(paragraphs_sans_gutenberg[doc_id].split("\n")[:5]))
```

```
[paragraph: 29] [similarity: 0.124]

CHAPTER IV.
OF THE ORIGIN AND USE OF MONEY.
[paragraph: 79] [similarity: 0.092]


     When the stock which a man possesses is no more than sufficient to
     maintain him for a few days or a few weeks, he seldom thinks of deriving
     any revenue from it. He consumes it as sparingly as he can, and
[paragraph: 80] [similarity: 0.088]

CHAPTER II.
OF MONEY, CONSIDERED AS A PARTICULAR
BRANCH OF THE GENERAL STOCK OF THE SOCIETY, OR OF THE EXPENSE OF MAINTAINING
THE NATIONAL CAPITAL.
```