Assignment 3

Marius Aarsnes

October 2018

1 Data loading and preprocessing

In this task i wrote the following functions:

```
read_file # Read a file
split_into_paragraphs # Split the raw data from a text file into paragraphs
remove_containing_word # Remove paragraphs containing the word "Gutenberg"
tokenize # Split each paragraph into single words based on white-space and line breaks
remove_punctuation # Remove string.punctuation and \n \r \t from all paragraphs
remove_punctuation_from_paragraph # Helper-method for removing punctuation from a paragraph
stem_paragraphs # Stem each word in each paragraph

generate_paragraphs # Uses the above functions to produce two lists of paragraphs
```

The result is two paragraphs; one “raw” paragraph used for printing in the later tasks, while the “parsed” paragraph is used for later computation and modeling.

2 Dictionary building

In this task i wrote the following functions:

```
generate_dictionary # Returns a dictionary object for a given list of paragraphs
generate_stopwords # Returns a list of stopwords based on common English words

# Based on a set of stopwords and a dictionary remove stopwords from dictionary
generate_bag_of_words
```

The result is a dictionary containing only non-stopwords and a bag-of-words for each paragraph using the dictionary.

3 Retrieval Models

I only wrote one function for this task, it creates all the corpus, models and similarity matrices:

```
retrieve_models
```

The function creates a a model based on TF-IDF and LSI for the paragraphs. Also, it generates a similarity matrix based on the two models.

The three most popular topic for the LSI model ended up being the following. I have restricted the output to be only three elements from the three most popular topics to keep the printout prettier.

```
[
    (0, '0.146*"labour" + 0.137*"price" + 0.127*"produc"'),
    (1, '0.260*"rent" + 0.225*"labour" + -0.213*"silver"'),
    (2, '0.350*"price" + 0.227*"silver" + -0.213*"trade"')
]
```

4 Querying

In this section the task was to query the models made in section 3. The query used for querying the different models was "*What is the function of money?*". First, the "necessary transformation" was made. Then I used the model to get the TF-IDF weights for the query. Which gave the following result:

```
[
    'money: 0.31356976854565205',
    'function: 0.9495651637745701'
]
```

Only "money" and "function" remained, as the other words were considered stopwords, resulting in them being removed.

According to the TF-iDF model, the three most relevant paragraphs are (**NOTE:** The color formatting has no meaning, it is just a bi-product of using a certain latex package):

```
[Paragraph 676]
Some part of the capital of every master artificer or manufacturer must
be fixed in the instruments of his trade. This part, however, is very
small in some, and very great in others, A master tailor requires no
other instruments of trade but a parcel of needles. Those of the master
shoemaker are a little, though but a very little, more expensive. Those

[Paragraph 987]
Political economy, considered as a branch of the science of a statesman
or legislator, proposes two distinct objects; first, to provide a
plentiful revenue or subsistence for the people, or, more properly, to
enable them to provide such a revenue or subsistence for themselves;
and, secondly, to supply the state or commonwealth with a revenue

[Paragraph 811]
Both productive and unproductive labourers, and those who do not labour
at all, are all equally maintained by the annual produce of the land
and labour of the country. This produce, how great soever, can never
be infinite, but must have certain limits. According, therefore, as
a smaller or greater proportion of it is in any one year employed in
```

After converting the query representation into a LSI-topics representation, the top three topics were:

```
[Topic 4]
(   4, '0.259*"bank" + -0.219*"price" +
    0.213*"circul" + 0.180*"money" +
    0.176*"capit" + -0.170*"corn" +
    0.170*"gold" + -0.160*"import" +
    -0.160*"export" + -0.138*"bounti"'
)

[Topic 63]
(
    63, '-0.177*"money" + -0.177*"class" +
    -0.124*"deriv" + 0.116*"proport" +
    -0.112*"export" + -0.111*"trade" +
    -0.111*"unproduct" + -0.107*"britain" +
    -0.106*"ii" + 0.106*"monopoli"'
)

[Topic 12]
(
    12, '0.371*"bank" + -0.193*"coin" +
    0.179*"money" + 0.161*"tax" + 0.159*"commod" +
    -0.159*"profit" + -0.148*"silver" +
    0.147*"paper" + -0.134*"gold" + 0.125*"demand"'
)
```

And the three most important paragraphs were (**NOTE:** The color formatting has no meaning, it is just a bi product of using a certain latex package):

```
[Paragraph 987]
Political economy, considered as a branch of the science of a statesman
or legislator, proposes two distinct objects; first, to provide a
plentiful revenue or subsistence for the people, or, more properly, to
enable them to provide such a revenue or subsistence for themselves;
and, secondly, to supply the state or commonwealth with a revenue

[Paragraph 1003]
It is partly owing to the easy transportation of gold and silver, from
the places where they abound to those where they are wanted, that the
price of those metals does not fluctuate continually, like that of the
greater part of other commodities, which are hindered by their bulk from
shifting their situation, when the market happens to be either over

[Paragraph 1002]
When the quantity of gold and silver imported into any country exceeds
the effectual demand, no vigilance of government can prevent their
exportation. All the sanguinary laws of Spain and Portugal are not able
to keep their gold and silver at home. The continual importations from
Peru and Brazil exceed the effectual demand of those countries, and
```