Information Retrieval 2020/21

Assignment 5
Topic modeling

If you want to implement your analysis in R download the IDE RStudio.

You find a corpus of tweets messages in ole "Lab Corpus.csv"

Use the Gensim library for Python and deliver your code as Jupyter notebook
You can find the API here https://radimrehurek.com/gensim/apiref.html

Pre-process the corpus of tweets and describe your choices
Prepare Corpus

Input : clean document
Purpose: create term dictionary of our corpus and Converting list of documents (corpus) into Document Term Matrix
Output : term dictionary and Document Term Matrix

Task 1 – _Group work
Latent Semantic Indexing (or Analysis)

Select N documents.
Create a Vocabulary from the documents and create a vector space of terms.

Compute the LSI for a term-document incidence matrix of size MxN with M={5,10,15} and respectively N={10,15,20}.

How did you choose the terms and the documents?

Discuss the rank of the matrix and the Frobenius errors for different choices of k.

How long did the computation take in each case? Discuss the computation effort by extending the size of the matrix.

Bonus: Choose one incidence matrix and compute U and V^T and discuss what are the eigenvectors more relevant.

Task 2 – _Group work

- Create an LSI model using Gensim
  `model = gensim.models.LsiModel(corpus, num_topics=k)`

Bonus: Task 1 and Task 2 for the corpus of news in ole "AssociatedPress.txt" that you can find in Assignment 2

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