**Duty: study of LDA method**

**Professor Jean-Charles Lamirel**

**Deadline for feedback**: 30 June

Email the duty document filled with answers to questions at: lamirel@loria.fr

The duty include one Python3 program name **CoherenceTestIwor.py** and 2 different datasets (i.e. **TokenVieuxM.txt** and **TokenVieuxN.txt**) that will be used for the proposed study.

To make the Python3 program workable on your computer, you must at least install Python3 and nltk and gensim libraries as well. Some additionnal complementary libraries from which these two libraries depends should perhaps also be installed.

You must send me the list of students participating to the group and provide me a duty document in which you present, for each question, the results of your experiments and the answer to my specific questions.

The datasets contains the tokens (i.e. the words) extracted from the abstracts of research papers related to **research on healthcare**. The two datasets have different sizes. Files contain one document description in the form of list of appearing tokens per line.

For better understanding of the Python program, you can refer to gensim related page: https://radimrehurek.com/gensim/models/ldamodel.html

**1) Launch the CoherenceTestIwor.py program with default dataset used in the Python code that is TokenVieuxM.txt**

a) Launch the program several times. What did you observe regarding the obtained coherence and perplexity values? Try to provide a reason for your observation (looking to papers or Web infos referring to LDA topic identification process). Consequently, explain how it can be possible to obtain the best results regarding coherence and perplexity.

b) Look at the topic descriptions obtained with one specific launch. Present the topic description output you obtain along with coherence and perplexity values. What can you conclude regarding these descriptions and theses values? Explain if you consider that the results are good are bad. Justify your answer in different ways.

**2) Modify the CoherenceTestIwor.py program to remove stopwords from document descriptions.**

a) Explain what are representing stopwords exactly (looking to papers or Web infos on that topic).

b) Go back to the results of question 1) b) and considering your response to question 2) a) explain what is the main content of topics obtained in question 1) b). Try to explain why.

c) For removing the stopwords from documents descriptions, you will need to modify the CoherenceTestIwor.py, as it is mentioned on 2). For that purpose, you can make use of some commented instructions in the program alongside with some slight modifications of some other intructions of the program. Describe which instructions you have modified and why.

Relaunch your modified program and present your obtained results including topic decriptions, coherence and perplexity values. Compare the results with the ones of question 1) b). Consequently, estimate the relationship between coherence, perplexity and topic quality.

For the next questions you will have to use your modified program including the stopwords removal function.

d) Launch the modified program several times making varying the number of expected topics from 5 to 20 and use the obtained values of coherence for each different number of topics to draw a chart. Looking to the obtained chart, try to conclude what should be the optimal number of topics for the document collection TokenVieuxM.txt. Justify your answer.

e) Change the dataset used by your modified program including the stop words removal function in order to use TokenVieuxN.txt document file instead of TokenVieuxM.txt document file. Use an expected number of topics equal to 10 to compare the results of this experiment with the experiment achieved with TokenVieuxM.txt in 2) c). What can you conclude?

f) Make the same experience as the one achieved in 2) d) but using TokenVieuxN.txt instead of TokenVieuxM.txt and number of expected topics varying from 5 to 30. Looking to the new obtained chart (that should be presented), draw a new conclusion.

Can you conclude that the optimal number of topic depend on the exploited dataset? Observe the content of the two datasets TokenVieuxM.txt and TokenVieuxN.txt more carefully and try to find the differences between the two. What can you conclude regarding the dependance between the optimal number of topic and the charateristics of the datasets after your observations.

**3) Most typical documents for topics**

a) For the experience conducted in 2) c) add a new step by modifying again the CoherenceTestIwor.py including stopwords removal to highlight which document is the most typical of each topic. For that purpose you can rely on topic proportion (or probability) in documents (looking to gensim page https://radimrehurek.com/gensim/models/ldamodel.html) or on measuring cosine similarity (looking to papers or Web infos on that topic) between topic profile (list of topic words) and document profiles (list of document words).

Present the new program and the obtained results.