README

In this assignment, you are asked to discover and display topics on the given dataset using the Latent Dirichlet Allocation (LDA) model. We recommend using the following libraries:

- NLTK: for text preprocessing
- Gensim: for model training and evaluation

The assignment involves taking the following steps:

Step 1: Text Preprocessing
This is an important step in any natural language processing (NLP) task and mainly consists of tokenization, stop-word removal, lemmatization, and discarding too rare or too frequent words. You can follow the basic steps given in the assignment pdf or use more advanced techniques in the NLTK library. You can read more at the following link:
https://towardsdatascience.com/text-preprocessing-with-nltk-9de5de891658

Step 2: TF/IDF Transformation
We need to represent text data in a vector space before training any machine learning model such as LDA. To this end, in this assignment, we asked you to transform the data using TF/IDF vectorization. If you are not already familiar with TF/IDF, you can read more about it at the following link:
http://www.tfidf.com/

Step 3: Model Training
After vectorizing text data, you can train an LDA model to extract topics. You can use Gensim's LdaModel class with the following recommended hyper-parameters:

```
alpha='auto',
eta='auto',
passes=10,
iterations=500,
eval_every=None,
random_state=12345
```

This class also accepts a "num_topics" parameter indicating the number of topics. You must change this parameter, train, and then evaluate the model described in the next step.

Step 4: Model Evaluation
After training the LDA model, we need to evaluate the quality of the extracted topics. In this assignment, we asked you to use Topic Coherence which scores a single topic by measuring the degree of semantic similarity between high-scoring words in the topic. You can read more about topic coherence at the following link:
http://qpleple.com/topic-coherence-to-evaluate-topic-models/

In this assignment, we use the C_v topic coherence score (introduced in this paper) for the top-20 words of each topic. You can obtain this score for every topic using the "top_topics"

method of Gensim's LdaModel. By averaging the topic scores, you get a single score for the trained model. You need to train the LDA model with different number of topics within {5, 10, 15, …, 50} and draw the average topic coherence against the number of topics curve. The optimal number of topics will be the one achieving the highest topic coherence.

CONTACT

Should you have any question, contact:

sina.sajadmanesh@epfl.ch
lakmal.meegahapola@epfl.ch