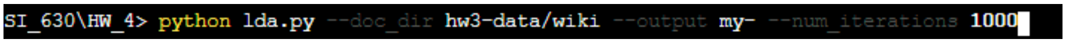
Topic Modeling

Firstly, run the lda.py and get the result in “my-topic.txt”.



```
SI_630\HW_4> python lda.py --doc_dir hw3-data/wiki --output my- --num_iterations 1000
```

Then, run Gensim and Mallet and get results in “genism.txt” and “mallet.txt” respectively.

For Gensim and Mallet, the top words for a topic make it seem like a coherent theme. For the LDA, “one” and “also” are top words for all the topics and they make no contribution and other top words make each topic seem like a coherent theme.

To be specifically, here are all the topics these methods contained.

The Gensim Model:

The first topic is related to sports.

The second topic is about environment.

Topic 3 is about tourism.

Topic 4 is related to nuclear.

Topic 4 is related to history.

The Mallet Model:

Topic 1: art

Topic 2: daily life

Topic 3: sports

Topic 4: environment

Topic 5: politics

The LDA model I implemented:

Topic 1: daily life

Topic 2: history

Topic 3: family

Topic 4: sports

Topic 5: politics

To wrap up, there are a lot of topics in common in these three methods, but they still have difference. For example, the Gensim model has a topic with “nuclear”, “radiation” and “war” as its top words, and none of the other methods selected these words. But they still have a lot in common. All of the methods have topics “sports”, “daily life” and “art”.

Time difference (1000 iterations for each methods):

Gensim: 9.78037 seconds

Mallet: 36.26380 seconds

LDA: 645.48927 seconds

Gensim is the fastest among all the methods, but it extracted a topic “nuclear” which is not a main topic in this wiki data. Mallet works best but it takes about half a minute which is way slower than Gensim model. So, the advantage of Gensim is the computational cost, while the advantage of Mallet is the topic accuracy. Compared to Gensim and Mallet, LDA implementation is slower as well as not did well in the topic selection.