UNIVERSITY OF ILLINOIS AT CHICAGO, INFORMATION AND DECISION SCIENCES

Topic Model in Mahout Under Hadoop

In this assignment, please read the description and readme file of the dataset first.
http://archive.ics.uci.edu/ml/datasets/Bag+of+Words

You don't need to run topic models on all dataset, only need to focus on NIPS full papers. This is the clean dataset with stop words removal, word stemming and other preprocessing done. The format of the docword.nips.txt file extracted from docword.nips.txt.gz is 3 header lines, followed by NNZ triples:

```
- - -
D
W
NNZ
docID wordID count
docID wordID count
docID wordID count
docID wordID count
...
docID wordID count
docID wordID count
docID wordID count
- - -
```

The format of the vocab.nips.txt file is line contains wordID=n.

You need to extract all documents first. Each document can be named by docID.txt.

You should hand in a report containing:

1. The source code: extracting all documents (20 points)
2. For each topic, print out top 10 words with their corresponding probabilities. (Topic term distribution) (20 points)
3. For each document, print out 10 topic probabilities. (Document-topic distribution) (20 points)
4. All Mahout and Hadoop commands you used and all corresponding temporary file summaries. (30 points)
5. Include a README file to describe commands you run LDA under Mahout and functionality of every code your wrote. (10 points)
6. Any additional codes with comments (if needed) you use to process files.

You can specify the number of topics K (e.g. $K = 10$)