Practical No. 7

AIM: Implement topic modelling using Latent Dirichlet Allocation.

LDA's approach to topic modeling is it considers each document as a collection of topics in a certain proportion. And each topic as a collection of keywords, again, in a certain proportion.

Once you provide the algorithm with the number of topics, all it does it to rearrange the topics distribution within the documents and keywords distribution within the topics to obtain a good composition of topic-keywords distribution.

When I say topic, what is it actually and how it is represented?

A topic is nothing but a collection of dominant keywords that are typical representatives. Just by looking at the keywords, you can identify what the topic is all about.

The following are key factors to obtaining good segregation topics:

1. The quality of text processing.
2. The variety of topics the text talks about.
3. The choice of topic modeling algorithm.
4. The number of topics fed to the algorithm.
5. The algorithms tuning parameters.

We will need the `stopwords` from NLTK and spacy's `en` model for text pre-processing. Later, we will be using the spacy model for lemmatization.

Lemmatization is nothing but converting a word to its root word. For example: the lemma of the word 'machines' is 'machine'. Likewise, 'walking' –> 'walk', 'mice' –> 'mouse' and so on.

Steps: I needed to install openNLP version 0.9 instead of 0.15 presently available because some methods didn't execute on it.

1. Import NewsGroups Dataset
2. Tokenize Sentences and Clean
3. Build the Bigram, Trigram Models and Lemmatize
4. Build the Topic Model

Steps: For Mahout LDA we did the following:

1. Extract the data to HDFS in news-all directory.
2. Go to the MAHOUT_HOME
>bin/mahout seqdirectory -i news-all -o news-seq
3.
> bin/mahout seq2sparse \
–i news-seq
-o news-tf \
–wt tf \
-a org.apache.lucene.analysis.WhitespaceAnalyzer

4. Convert the TF vectors from SequenceFile<Text, VectorWritable> to SequenceFile<IntWritable,Text> :
>bin/mahout rowid -i news-tf/tf-vectors -o news-tf-int
5. Run the following command to perform the LDA computation:
> bin/mahout cvb \
-i news-tf-int/matrix -o lda-out \
-k 10
-x 20 \
-dict news-tf/dictionary.file-0 \
–dt lda-topics \
–mt lda-topic-model
6. Dump and inspect the results of the LDA computation:
>bin/mahout seqdumper -i lda-topics/part-m-00000
7. Join the output vectors with the dictionary mapping of term to term indexes:
>bin/mahoutvectordump -i lda-topics/part-m-00000 --dictionary
20news-tf/dictionary.file-0 --vectorSize 10 -dt sequencefile

Code:

We have information in the form of a Jupyter Notebook.

Conclusion:

The main conclusion is understanding the LDA algorithm and how it works. Latent basically means that topics are generated later after analyzing the dataset. We also learned about gensim. Gensim is basically a library for unsupervised topic modeling and natural language processing, using modern statistical machine learning. Gensim is implemented in Python and Cython for top performance and scalability.

We learned how we can further optimize our code using multi core especially if we are executing it on the Google Collab.

About LDA done in Mahout: We had this cookbook on Mahout which helped us in performing topic modelling. To see how it works we need to see the CVB command.

Mahout CVB version of LDA implements the Collapse Variable Bayesian inference algorithm using an iterative MapReduce approach:

>bin/mahout cvb -i 20news-tf-int/matrix -o lda-out -k 10 -x 20
20news-tf/dictionary.file-0 -dt lda-topics -mt lda-topic-model
-dict points to the dictionary containing the mapping to term-indexes. The -i parameter provides the input path, while the -o parameter provides the path to store
the output. -k specifies the number of topics to learn
-x specifies the maximum number of iterations for the computation. -dict points to the dictionary containing the mapping of terms to term-indexes. Path given in the -dt parameter stores the training topic distribution. Path given in -mt is used as a temporary location to store the intermediate models.