LING 410X: Language as Data

Semester: Spring '18

Instructor: Sowmya Vajjala

Iowa State University, USA

29 March 2018

Topic modeling -resources page

https://github.com/trinker/topicmodels_learning

"Learning" a topic model with mallet: steps in the code

1. Pre-processing of the corpus (in author's version: splitting it into chunks, pre-processing of chunks)
2. Getting it into a two column format (id, text).
3. Using mallet package in R to build a topic model.
4. use mallet.import() function to convert a dataset into mallet format
5. use MalletLDA() function to create a topic model, setting its parameters.
6. observe and analyze the output in different ways.

Tuesday's class

- I followed this step by step, with movie reviews corpus (without considering positive/negative sentiment information)
- Code for that is: anothertopicmodelexample.R on Canvas.
- We also discussed about how a topic model learns (conceptually), and how are they typically evaluated.
- Today: it is more about building your own topic models.

Exercise

- open mallet-29marlab.R - you have some lines of code, along with comments and questions.
- Walk through that step by step (dataset is provided) and answer the questions I posed there.
- About the corpus: common core standards exemplar texts - from US school textbooks/reading materials.
- Different topics, different genres - literature, informational texts, speeches etc.
- You may recognize some titles.

Post your observations from the exercise, addressing my questions in the comments inside R code. Post in the forum with today's date.

Exercise: Alternate option

- Topic modeling with News articles.
- Get news articles on "USA" from Guardian for two time periods. Build one topic model per time period based on article content.
- Compare the topics discussed about USA between different time periods.
- Same questions as before remain about pre-processing, choosing number of topics etc.

General Remarks for next week

- We completed all except two chapters from the textbook (chapter 5, chapter 11).
- Next week: text visualization (covers ideas in Chapter 11) - read this chapter.
- Reminder: Submit Assignment 5
- Regarding Assignment 6: You can either use mallet library or use topicmodels library as described in the tutorial.
- Start thinking about your final projects now - conceptually, everything is pretty much done.