LING 410: Language as Data

Semester: Spring '18

Instructor: Sowmya Vajjala

Iowa State University, USA

17 Apr 2018

Class Outline

- Few quick announcements
- Assignment 6 discussion
- General revision/discussion (in the form of lot of questions)

Course Evaluations

- Course evaluations messages must have come- please complete the evaluations.
- I will give time in one of the lab sessions remaining, if needed.

Next Class

- Time for preparing for your project presentations.
- I will leave very early (around 10 am) as I have to attend a LAS event
- But I strongly recommend you to work on your projects in this time.

Project Presentations - Schedule and Guidelines

Project presentations schedule-1

24th April

- Carlos Eduardo Back Vianna
- Su-Yeon Cho
- Brody Dingel and John Piegors
- Helena Hansen
- Gage Williams

(max 15 min each, including questions)

Project presentations schedule-2

26th April

- Lauren Didesch
- Kori Ralston
- Xiaochi Jin and Nicole Richwine

(max 15 min each, including questions)

Presentation Guidelines

- grade: 5%
- Each team gets 10 (+/- 2 minutes) to present, and 2-3 more minutes for questions and discussion.
- Stick to the time, and practice accordingly.
- You can use your laptop or send the presentations to me before the class.
- All presentations need to be uploaded on Canvas in FinalProject-Presentation by the end of the class on 26th.
- You have time up to 3rd May midnight to upload your final submission (report + R code). I cannot accept late submissions.
- More on format for final report next week.

Presentation Guidelines-2

- Introduce yourselves, introduce your project, talk about your data, your methods, your results, and conclusions so far.
- Mention what is left for final submission, and how what you did can be improved upon.
- Be professional, take questions in a good spirit.
- If you have no problems with privacy etc, you can share your presentations in the discussion forum too.

Assignment 6 discussion

Assignment 6 - process discussion

- Starting point: use tm to build a term-document matrix (rows are words, columns are texts)
- i.e., use VCorpus or Corpus(VectorSource(.....)) or Corpus(DirSource(...)) or any other tm functions, to read in a folder of .txt files (2 files in our case)
- use tm_map() function to do some pre-processing
- Use the post-pre-processed version to build the term-doc. matrix

Assignment 6 - process discussion

To build wordclouds for each text

- You should re-order the matrix such that more frequent words come on top.
- use the wordcloud function on this above matrix

Assignment 6 - process discussion

To build wordclouds for each text

- You should re-order the matrix such that more frequent words come on top.
- use the wordcloud function on this above matrix

To build commonality cloud, you use the whole tdm instead of only one column each time. It shows only those words that appear in both documents, and the size of the word represents combined frequency.

Assignment 6 - process discussion

Comparison cloud

- aim: to compare relative occurrence of words in two or more documents (yes, we can do with multiple documents).
- If a word appears more frequently in one document, than the other, it is seen in that side of the cloud.

`https://www.rdocumentation.org/packages/wordcloud/versions/2.5/topics/comparison.cloud`

analyzing SOTU speeches with wordcloud

Some online blog posts:

- `http://blog.fellstat.com/?p=101`
- `https://rpubs.com/brandonkopp/creating-word-clouds-in-r`

When to go for which method

- Small number of documents, and you have a general idea about them (1-5):

When to go for which method

- Small number of documents, and you have a general idea about them (1-5): wordclouds are perhaps sufficient.
- Lot of documents, and you have little idea about what each of them are about:

When to go for which method

- Small number of documents, and you have a general idea about them (1-5): wordclouds are perhaps sufficient.
- Lot of documents, and you have little idea about what each of them are about: topic models
- Small number or lot of documents, you are looking to put them into groups based on some characteristics:

When to go for which method

- Small number of documents, and you have a general idea about them (1-5): wordclouds are perhaps sufficient.
- Lot of documents, and you have little idea about what each of them are about: topic models
- Small number or lot of documents, you are looking to put them into groups based on some characteristics: clustering
- small number or lot of documents, you know each has to be in one of the given categories, you should predict that category:

When to go for which method

- Small number of documents, and you have a general idea about them (1-5): wordclouds are perhaps sufficient.
- Lot of documents, and you have little idea about what each of them are about: topic models
- Small number or lot of documents, you are looking to put them into groups based on some characteristics: clustering
- small number or lot of documents, you know each has to be in one of the given categories, you should predict that category: classification

In the following slides, I will describe some problem scenario involving textual data, and you should discuss with your neighbor or other classmate (don't work individually!!) and share your ideas after some time.

For several of these, there is no concrete, single answer. The goal of this exercise is to make you think about new scenarios, give you some practice with approaching a problem without much guidance.

Take this scenario: authorship question

- Federalist papers are a collection of 85 essays written under a pseudonym "Publius" by Alexander Hamilton, James Madison, and John Jay
- Authorship of 73 of them is fairly certain.
- For the other 12, scholars debate about authorship.
- If someone comes to you and asks you to suggest a solution for knowing the authorship, what will you suggest?

(Discuss with your neighbor. We can pool responses after 5minutes)

Take this scenario: vocabulary richness

- I want to know whether English language learners use a richer vocabulary as they get better in terms of language proficiency.
- How can I answer this question - what do I need, how do I proceed, based on what you learnt in this class so far?

(Discuss with your neighbor. We can pool responses after 5minutes)

Take this scenario: analyzing word usage

- Let us say I have this scenario where I want to study:
  - What are the different ways in which people use the word "apple" in newspapers
  - How are English swear words used in different English speaking countries

  - how can what you learnt in this class help you solve these two questions?

(Discuss with your neighbor. We can pool responses after 5minutes)

Take this scenario: authorship question, again

- Let us say I have 500 articles written by a columnist A, 500 articles written by columnist B, and 500 by columnist C in a newspaper (say NYT).
- Someone gave me 100 anonymous articles, and we are 100% sure the author is either A or B or C.
- If someone comes to you and asks you to suggest a solution for knowing the authorship, what will you suggest?

(Discuss with your neighbor. We can pool responses after 5minutes)

Take this scenario: politics and media

- How will you approach the question: what are the main issues of interest for democrats vs republicans?

(Discuss with your neighbor. We can pool responses after 5minutes)

Take this scenario:pre-processing

- Can you think of some scenarios, where we don't have to (or perhaps should not) remove punctuation, do lowercasing, remove stopwords etc?

(Discuss with your neighbor. We can pool responses after 5minutes)

Take this scenario: pre-processing

If you compare a collection of news articles vs tweets

- How is pre-processing different?
- Are there issues that needs to be handled in one, but not the other kind of data?
- What issues exist if we want to do, say, classification of a news article into one of the possible 4 groups vs classifying a tweet into one of the 4 possible groups?

(Discuss with your neighbor. We can pool responses after 5minutes)

Challenges

- In all these scenarios, what are some primary challenges?

(Discuss with your neighbor. We can pool responses after 5minutes)

Steps

- Can you think of some common steps involved in working with such questions?

(Discuss with your neighbor. We can pool responses after 5minutes)

Next Class

- Time for preparing for your project presentations.
- I will leave early as I have to attend a LAS event -you can continue working in the lab.