LING 410X: Language as Data

Semester: Spring '18

Instructor: Sowmya Vajjala

Iowa State University, USA

6 March 2018

Class Outline

- Quick recap of text classification
- Quick recap of what we learnt in R and text analysis so far
- Ideas for final projects, expectations etc.
- Discussion
- Mid-term feedback

Text Classification Review

Text Classification: Review

- We saw how to do text classification using bag-of-words features and SVM classification algorithm

Text Classification: Review

- We saw how to do text classification using bag-of-words features and SVM classification algorithm
- We used a movie review corpus and task was sentiment classification

Text Classification: Review

- We saw how to do text classification using bag-of-words features and SVM classification algorithm
- We used a movie review corpus and task was sentiment classification
- We learnt about using tm.

Text Classification: Review

- We saw how to do text classification using bag-of-words features and SVM classification algorithm
- We used a movie review corpus and task was sentiment classification
- We learnt about using tm.
- We saw how to use a "learned" classification model to predict categories for new data.
- We also saw how to understand whether our classifier is doing well or not.

Text Classification: How to improve a baseline classification approach

- Check different pre-processing settings (e.g., do we need stemming? should I remove stopwords? do I need to keep numbers or remove them? is punctuation important for my classification task?)

Text Classification: How to improve a baseline classification approach

- Check different pre-processing settings (e.g., do we need stemming? should I remove stopwords? do I need to keep numbers or remove them? is punctuation important for my classification task?)
- Look at the number of word features you want to keep. Should you keep every word as a feature or remove some? How to remove?

Text Classification: How to improve a baseline classification approach

- Check different pre-processing settings (e.g., do we need stemming? should I remove stopwords? do I need to keep numbers or remove them? is punctuation important for my classification task?)
- Look at the number of word features you want to keep. Should you keep every word as a feature or remove some? How to remove?
- Is it possible to increase the training data? Is "not having enough data" the main problem?

Text Classification: How to improve a baseline classification approach

- Check different pre-processing settings (e.g., do we need stemming? should I remove stopwords? do I need to keep numbers or remove them? is punctuation important for my classification task?)
- Look at the number of word features you want to keep. Should you keep every word as a feature or remove some? How to remove?
- Is it possible to increase the training data? Is "not having enough data" the main problem?
- Should I try other classification algorithms?

Text Classification: How to improve a baseline classification approach

- Check different pre-processing settings (e.g., do we need stemming? should I remove stopwords? do I need to keep numbers or remove them? is punctuation important for my classification task?)
- Look at the number of word features you want to keep. Should you keep every word as a feature or remove some? How to remove?
- Is it possible to increase the training data? Is "not having enough data" the main problem?
- Should I try other classification algorithms?
- Should I go beyond words ("don't like" as a feature is not the same thing as "don't" and "like" being separate words)

Text Classification: more details

For a little bit of theoretical background, I would suggest reading the following chapter from a standard Natural Language Processing textbook: `https://web.stanford.edu/~jurafsky/slp3/6.pdf`. You will need to know a little bit about probability to understand this.

What we learnt about text analysis and R so far

What we learnt about text analysis so far

1. Reading text content into R
2. Reading a folder of text files into R

What we learnt about text analysis so far

1. Reading text content into R
2. Reading a folder of text files into R
3. Doing some pre-processing (lowercasing, punctuation handling, regular expressions etc)
4. Splitting one file into parts (e.g., by chapter)

What we learnt about text analysis so far

1. Reading text content into R
2. Reading a folder of text files into R
3. Doing some pre-processing (lowercasing, punctuation handling, regular expressions etc)
4. Splitting one file into parts (e.g., by chapter)
5. Counting frequencies, creating basic plots
6. Getting lexical variety measures

What we learnt about text analysis so far

1. Reading text content into R
2. Reading a folder of text files into R
3. Doing some pre-processing (lowercasing, punctuation handling, regular expressions etc)
4. Splitting one file into parts (e.g., by chapter)
5. Counting frequencies, creating basic plots
6. Getting lexical variety measures
7. creating and evaluating text classification models, given a dataset.

What in R was useful for these?

1. Reading content: scan function, different libraries (GuardianR, nytimes etc)

What in R was useful for these?

1. Reading content: scan function, different libraries (GuardianR, nytimes etc)
2. Pre-processing: functions like tolower, paste, functions in stringr library, regular expressions, etc.

What in R was useful for these?

1. Reading content: scan function, different libraries (GuardianR, nytimes etc)
2. Pre-processing: functions like tolower, paste, functions in stringr library, regular expressions, etc.
3. Splitting a file into parts: which() function
4. Organizing data in different ways: vectors, lists, data frames, matrices

What in R was useful for these?

1. Reading content: scan function, different libraries (GuardianR, nytimes etc)
2. Pre-processing: functions like tolower, paste, functions in stringr library, regular expressions, etc.
3. Splitting a file into parts: which() function
4. Organizing data in different ways: vectors, lists, data frames, matrices
5. Counting frequencies: sort, table
6. Creating basic plots: plot

What in R was useful for these?

1. Reading content: scan function, different libraries (GuardianR, nytimes etc)
2. Pre-processing: functions like tolower, paste, functions in stringr library, regular expressions, etc.
3. Splitting a file into parts: which() function
4. Organizing data in different ways: vectors, lists, data frames, matrices
5. Counting frequencies: sort, table
6. Creating basic plots: plot
7. lexical variety: sum, length functions
8. others: using rbind, cbind, sapply, lapply etc

Other useful things

1. writing our own R functions
2. writing a for loop
3. R markdown

So many of them - how to keep track?

- Attend classes regularly. Maintain notes.
- Spend some time with lecture slides/tutorials; Have a DIY attitude
- Use R outside classroom, and not only for doing assignments
- Be organized - have a folder structure in your computer. Keep all code in one place.
- Participate in the class, discuss in the forums, meet during office hours.

So many of them - how to keep track?

- Attend classes regularly. Maintain notes.
- Spend some time with lecture slides/tutorials; Have a DIY attitude
- Use R outside classroom, and not only for doing assignments
- Be organized - have a folder structure in your computer. Keep all code in one place.
- Participate in the class, discuss in the forums, meet during office hours.

"Education is the only business where customers pay more and expect less" - one day, I hope that will be proven wrong!

Final Projects - Discussion

Final Projects for the course: Expectations

- grade weightage: 25%
- Individual or group projects (Group projects are preferred, with group size being 2 or 3 max).
- Aim: choose some text dataset, explore a micro/meso/macro text analysis problem, and use some visualizations to summarize information from text.
- initial report (explaining what dataset you will use, what you will do with it, how you plan to do it): Due on 7th April - 5%
- presentation in the class in the last week of classes - 5%
- submission of your report about the project (with visuals, relevant R code etc) in exams week - 15%

Some ideas for project tasks

My final projects example descriptions document on Canvas.

Some Datasets

- Some data repositories for classification problems - look for text data in these
  - https://goo.gl/UUkNZ1
  - https://goo.gl/3nKyAQ
- For topic modeling:
  - Topic modeling datasets for humanities: https://de.dariah.eu/tatom/datasets.html
  - Clinton-Trump tweets dataset: https://www.kaggle.com/benhamner/clinton-trump-tweets
  - Congressional speech data: http://www.cs.cornell.edu/home/llee/data/convote.html
  - Presidential speeches transcripts from Miller Center https://millercenter.org/the-presidency/presidential-speeches. A project that is related to this: https://github.com/BBischof/speaksLike

Rest of this class

- Think about some ideas for this course project (Take a look at Canvas document!)
- Talk to others, see if you want to form groups (strongly encouraged)
- try to connect what you learn about in your own disciplines to this course and formulate some project ideas that will be relevant for you in future (in course work, in job applications in future etc).
- Present your ideas in class on Thursday (Don't miss the class!).
- You don't need to prepare slides (you can, if you want). The idea is to discuss some ideas, and get some feedback from others.
- Keep in mind: there is only a limited amount of time. Don't think about impossible ideas.

Mid-term Feedback

- Please fill up the mid-term feedback.
- It is primarily for me to get some feedback, as there is still enough time to get better.
- It is also for you to think about how you are doing, and how you can improve.

Next class

- Discussion about your project ideas
- Today's attendance question: Try to look around, and, explain what a do.call() function does in R, with an example.
- There will be time alloted to do Assignment 4 in the class.