LING 410X: Language as Data

Semester: Spring '18

Instructor: Sowmya Vajjala

Iowa State University, USA

12 Apr 2018

Class Outline

- Plan, Schedule for final project presentations (24th, 26th April)
- General practice
- Reminder: Submit Assignment 6 on time!

Final project presentations

- 5% of your total grade
- 10-12 min maximum + 3-5min for questions (per team, if you are doing together with someone)
- What I expect:
  1. Describe what you are trying to do, how is it relevant to outside world, to your own discipline etc.
  2. Show us what you have so far (some figures, if you have any, is good)
  3. Talk about how you got those results you show

Presentations: schedule for 24th April

In this order:

1. Carlos Eduardo Back Vianna
2. Su-Yeon Cho
3. Brody Dingel and John Piegors
4. Helena Hansen
5. Gage Williams

Presentations: schedule for 26th April

In this order:

1. Lauren Didesch
2. Kori Ralston
3. Xiaochi Jin and Nicole Richwine

Three exercises are described in the next few slides. You can choose whichever is closer to your final project idea and work on that. You can also work on your final projects in the class if you want. But work on only stuff related to this class!

Write a summary of what you did today in the discussion forum for today.

Exercise on Text Clustering/Visualization

- There are two zip files-stylocorpus.zip and rollingdelta.zip - learn to work with Stylo library in R, and using these datasets. Follow the stylo tutorials from last time.
- Note down your observations.

Exercise on Text Classification

- There is a zip file prof-classify.zip containing the data. primary_set folder contains about 700 text files, and there are two categories: A2 and B2. Ignore all other information in the file name. Consider this as your training data. There are some test instances in secondary_set.
- Source of data: `https://goo.gl/v0r6et`
  (we are using a subset of the full data today)
- Description of data: It is a dataset of essays written by English learners from China, Japan and Korea, and there are two levels of proficiency (A2 and B2)
- Task: learn to distinguish between A2 and B2 - use any classification library you want (including stylo)
- Data in stylo format is in: forStylo folder. General data, which you can use with tm or anything is in the other folder.
- Note down your observations.

Exercise on Topic Modeling

- Use any collection of files you want, practice doing topic models
- Note down your observations.

Next week

- Attendance for today: Write about what you did today.
- Topics for next week: Revision.
- Post on the discussion board in a thread called "Topics for Revision" if you want any topics to be discussed in better detail in the coming week.