LING 410X: Language as Data

Semester: Spring '18

Instructor: Sowmya Vajjala

Iowa State University, USA

15 February 2018

Class Outline

- Announcements etc
- Assignment 2 discussion
- Discussion on the Technology Review article
- Some notes on R syntax
- Quick review of KWIC and n-grams
- Practice exercises.

Extra tutorial session - optional

- I want to hold an additional session on some evening (preferably thursday or friday) next week (say: 4-6pm) for general discussion and questions.
- This is optional and there is no set agenda
- The goal is to do some revision, and give additional clarifications on topics some of you are not clear about.
- I will book a lab and let you know - you can walk in and join for as long as you want.
- I would prefer if questions are sent to me apriori (I will setup a discussion forum)
- How many of you will be interested in this? (It is okay if there are only 1 or 2)

Assignment 2 Discussion

Assignment 2 discussion: Question 1

```
htm <- readLines("2446.htm")
txt <- readLines("2446.txt")
length(htm) #gives 8139
length(txt) #gives 5081
```

Questions:

- Why is the length different?
- Which one of these formats is easy to process forR in your opinion? Why?
- Which one may give more interesting information?
- readline() vs readLines() functions
- readLi nes() vs scan() functions

Assignment 2 discussion: Question 2

```
guardian_key <- "XXXXXXX"
step 1: results <- get_guardian("justin+trudeau", section= "world", api.key = guardian_key,
                            to.date = "2018-01-15", from.date = "2018-01-01")
#section is optional.
step 2: nrow(results) #gives number of results
step 3: names(results) #gives column names
step 3:
my_df <- data.frame(results["id"],results["wordcount"])
#or whichever columns you want
step 4: copy-paste to spreadsheet, or use a new library to create spreadsheets from R or
do this: write.csv(my_df, file = "results.csv")
- If you double-click this file you created, it usually will open in MS Excel
or such software as a spreadsheet.
```

Assignment 2 discussion: Question 2

```
guardian_key <- "XXXXXXX"
step 1: results <- get_guardian("justin+trudeau", section= "world", api.key = guardian_key,
                            to.date = "2018-01-15", from.date = "2018-01-01")
#section is optional.
step 2: nrow(results) #gives number of results
step 3: names(results) #gives column names
step 3:
my_df <- data.frame(results["id"],results["wordcount"])
#or whichever columns you want
step 4: copy-paste to spreadsheet, or use a new library to create spreadsheets from R or
do this: write.csv(my_df, file = "results.csv")
- If you double-click this file you created, it usually will open in MS Excel
or such software as a spreadsheet.
```

Oh, btw, I know many of your guardian API keys!!

Article Discussion

Overview

- What is already known to the authors: ISIS uses social media, especially Twitter, to spread its ideas.

Overview

- What is already known to the authors: ISIS uses social media, especially Twitter, to spread its ideas.
- Their question: What do ISIS and followers/sympathizers talk on twitter? Why (how?) do such messages spread?

Overview

- What is already known to the authors: ISIS uses social media, especially Twitter, to spread its ideas.
- Their question: What do ISIS and followers/sympathizers talk on twitter? Why (how?) do such messages spread?
- Data: about 2 million Arabic language tweets posted by 25K ISIS members over a period of: Jan-June 2015.

Overview

- What is already known to the authors: ISIS uses social media, especially Twitter, to spread its ideas.
- Their question: What do ISIS and followers/sympathizers talk on twitter? Why (how?) do such messages spread?
- Data: about 2 million Arabic language tweets posted by 25K ISIS members over a period of: Jan-June 2015.
- Pre-processing: performed tokenization and stemming. Removed non-Arabic tweets. compiled 100 most popular stems.
- Further pre-processing: removed stems that are not related to their analysis. Left with 34 stems. Grouped them into four categories: violence, theological, sectarian, names.

Overview

- What is already known to the authors: ISIS uses social media, especially Twitter, to spread its ideas.
- Their question: What do ISIS and followers/sympathizers talk on twitter? Why (how?) do such messages spread?
- Data: about 2 million Arabic language tweets posted by 25K ISIS members over a period of: Jan-June 2015.
- Pre-processing: performed tokenization and stemming. Removed non-Arabic tweets. compiled 100 most popular stems.
- Further pre-processing: removed stems that are not related to their analysis. Left with 34 stems. Grouped them into four categories: violence, theological, sectarian, names.

Analysis

- Analysis: if a tweet has majority stems from one category, categorize the tweet as that (main point. there is more).
- They report on percentage of tweets belonging to different categories
- They have plots showing the tweets of different categories over a period of time, and correlating them with news items.

Note: I am not going to discuss their conclusions and implications and so on. That is not relevant for our course.

Quick notes on R syntax

collapse vs sep in paste function

What will be output of all these???

```
paste("1st", "2nd", "3rd", collapse = ", ")
paste("1st", "2nd", "3rd", sep = ", ")
paste("1st", "2nd", "3rd", collapse = ", ", sep = ":")
vec1 <- c("1st", "2nd", "3rd")
vec2 <- c("4th", "5th", "6th")
paste(vec1, collapse = ":: ")
paste(vec1, sep = ":: " )
paste(vec1, vec2, sep = "::")
paste(vec1, vec2, sep = "::", collapse = "--")
```

Interesting R feature we need to be aware of

```
paste(c('v1','v2'),collapes='+')
paste(c('v1','v2'),whatever='+')
```

-What do you think happens in these two cases?

Interesting R feature we need to be aware of

```
paste(c('v1','v2'),collapes='+')
paste(c('v1','v2'),whatever='+')
```

-What do you think happens in these two cases?
paste interprets this as this: you want to paste each element of first vector with a variable collapes or whatever. Will not throw an error!!!

Tuesday class review

Tuesday's class

- Getting a word's occurrence, in context
- KWIC.R, ModifiedKWIC.R - on canvas. You should figure out what the differences are. Both work.
- We discussed about ngram package in R
  - `https://cran.r-project.org/web/packages/ngram/ngram.pdf`
  - `https://cran.r-project.org/web/packages/ngram/vignettes/ngram-guide.pdf`

Practice Exercises

Exercise - 1: KWIC

- Modify the KWIC.R code from Tuesday to take the following information from a user: a folder/directory path, a word, context size.
- Using this information, print to the R console the word and its context for all .txt files in the folder, one by one.
- Note 1: create a small folder with 2 or 3 files. Don't try with 100 files directly.
- Note 2: Don't test using a common word like "The" or something. You will see a lot of output! Try with some very infrequent word.
- Share your code on the forum for today.

Exercise - 2: Ngrams

Take any text file, get top-10 uni/bi/tri/4 ngrams for this file. Create four 4 plots where x-axis is 1–10, y-axis has the ngram frequency (i.e., plot the frequencies of ngrams in descending order). Do this with R markdown and post your html/doc/pdf on discussion forum titled 15th Feb 2018.

Additional Exercise, if interested

Modify the KWIC program to suit variable contexts on left and right (e.g., 2 words left, 3 words on right)

Next week

- Topics: overview of text analysis and pattern extraction from collections of documents - text classification, clustering and topic modeling
- Introduction to text classification
- To do for you: read this article: `https://goo.gl/qhT3u4` and we will discuss about this next week on Tuesday during the class.
- Keep this in mind: We are moving from straight forward counting analyses to predictive analysis from next week!