Statistical Language Models 2019 Week 2 part 1

Dr. Dave Campbell
davecampbell@math.carleton.ca

Event

Hockey Hack Day in Canada

March 27 & 28, 2020:   Vancouver, Toronto, Ottawa

https://hockeyhackday.ca

Approximate Course Outline

Week 1: ShinyApps and Dashboarding

Week 2: TidyText & obtaining data, dealing with time events

Week 3: Regular Expressions; Word co-occurrence explorations

Week 4: Sentiment Analysis; Stochastic process models

Week 5: Exponential models for time between events.

Week 6: Bayesian Basics; Author attribution models; hierarchical models

Week 7: MCMC Diagnostics

Week 8: Embeddings and Word2Vec; Cryptography

Week 9: Clustering; Latent Dirichlet Allocation and topic models.

Week 10: Variational Inference

Week 11: Getting Fancier with Language Models

Week 12: Student projects and presentations

Maggie’s Farm
Bob Dylan {https://www.youtube.com/watch?v=DFv3sRnmHB0}

```
Maggie = c("I ain't gonna work on Maggie's farm no more
No, I ain't gonna work on Maggie's farm no more
Well, I wake up in the morning, fold my hands and pray for rain
I got a head full of ideas that are drivin' me insane
It's a shame the way she makes me scrub the floor
I ain't gonna work on Maggie's farm no more
I ain't gonna work for Maggie's brother no more
No, I ain't gonna work for Maggie's brother no more
Well, he hands you a nickel, he hands you a dime
He asks you with a grin if you're havin' a good time
Then he fines you every time you slam the door
I ain't gonna work for Maggie's brother no more
I ain't gonna work for Maggie's pa no more
No, I ain't gonna work for Maggie's pa no more
Well, he puts his cigar out in your face just for kicks
His bedroom window, it is made out of bricks
The national guard stands around his door
Ah, I ain't gonna work for Maggie's pa no more
I ain't gonna work for Maggie's ma no more
No, I ain't gonna work for Maggie's ma no more
Well, she talks to all the servants about man and God and law
Everybody says, she's the brains behind pa
She's sixty-eight but she says she's fifty-four
I ain't gonna work for Maggie's ma no more
I ain't gonna work on Maggie's farm no more
I ain't gonna work on Maggie's farm no more
Well, I try my best to be just like I am
But everybody wants you to be just like them
They say sing while you slave and I just get bored
Ah, I ain't gonna work on Maggie's farm no more")
```

Regular Expression detour

```
# find which vector elements have the pattern

grep(pattern = Pattern2Find,   x = VectorOfStrings2Search)

grep(pattern = "and",  x = MaggieText)

#return the actual value:

grep(pattern = Pattern2Find,   x = VectorOfStrings2Search, value = TRUE)

grep(pattern = "and",  x = MaggieText, value = TRUE)

# Logical search, return trues or falses

grepl(pattern = Pattern2Find,  x = VectorOfStrings2Search)
grepl(pattern = "and",  x = MaggieText)
stringr::str_detect(string = VectorOfStrings2Search, pattern = Pattern2Find)
stringr::str_detect(string = MaggieText, pattern = "and")
```

Regular Expression detour

Replace the first occurrence (in each of the strings in the vector):

```
sub(pattern = Pattern2Find ,replacement= String2Sub, x = String2Alter)
```

```
sub(pattern = "and",replacement="***", x = MaggieText)
```

```
stringr::str_replace(string = String2Alter, pattern = Pattern2Find ,replacement= String2Sub)
```

```
stringr::str_replace(string = MaggieText, pattern = "and" ,replacement= "***")
```

Regular Expression detour

Replace all occurrences:

```
gsub(pattern = Pattern2Find ,replacement= String2Substitute, string = String2Alter)

gsub(pattern = "and",replacement="***", x = MaggieText)

stringr::str_replace_all(string = String2Alter, pattern = Pattern2Find ,replacement= String2Sub)

stringr::str_replace_all(string = MaggieText, pattern = "and" ,replacement= "***")
```

Options and Bracketing

```
grep(MaggieText,pattern = "(ma)|(pa)\\s",value = TRUE)

grep(MaggieText,pattern = "((ma)|(pa))\\s",value = TRUE)
```

Back to Macbeth Acts

```
Macbeth = gutenberg_download(1533)

Macbeth %>%

  unnest_tokens(      output = ngrams,input = text, token =
"ngrams",n=2)  %>%

  count(ngrams) %>%  filter(grepl("act (i|x|v|l|c|d|m)",
ngrams))
```

```
#finding acts with a roman numeral followed by anything
but the letter n

Macbeth %>%

  unnest_tokens(      output = ngrams,input = text, token =
"ngrams",n=2)  %>%

  count(ngrams) %>%   filter(grepl("act (i|x|v|l|c|d|m)[^n]",
ngrams))
```

```
#finding acts with any number of roman numerals only up
to the end of the word

Macbeth %>%

  unnest_tokens(      output = ngrams,input = text, token =
"ngrams",n=2)  %>%

  count(ngrams) %>%   filter(grepl("act (i|x|v|l|c|d|m)+\\>",
ngrams))
```

Find Maggie’s family and things

```
MaggieText = unlist(strsplit(Maggie,split="\n"))
```

Return song lines mentioning "and" but not "hand"

Return song lines mentioning Maggie’s family members

Replace pronouns with "they" but ignores possessive pronouns and contractions. Make sure that grammar is correct.

Some exploratory descriptives
Bob Dylan {https://www.youtube.com/watch?v=DFv3sRnmHB0}

```
Maggie = c("I ain't gonna work on Maggie's farm no more
No, I ain't gonna work on Maggie's farm no more
Well, I wake up in the morning, fold my hands and pray for rain
I got a head full of ideas that are drivin' me insane
It's a shame the way she makes me scrub the floor
I ain't gonna work on Maggie's farm no more
I ain't gonna work for Maggie's brother no more
No, I ain't gonna work for Maggie's brother no more
Well, he hands you a nickel, he hands you a dime
He asks you with a grin if you're havin' a good time
Then he fines you every time you slam the door
I ain't gonna work for Maggie's brother no more
I ain't gonna work for Maggie's pa no more
No, I ain't gonna work for Maggie's pa no more
Well, he puts his cigar out in your face just for kicks
His bedroom window, it is made out of bricks
The national guard stands around his door
Ah, I ain't gonna work for Maggie's pa no more
I ain't gonna work for Maggie's ma no more
No, I ain't gonna work for Maggie's ma no more
Well, she talks to all the servants about man and God and law
Everybody says, she's the brains behind pa
She's sixty-eight but she says she's fifty-four
I ain't gonna work for Maggie's ma no more
I ain't gonna work on Maggie's farm no more
I ain't gonna work on Maggie's farm no more
Well, I try my best to be just like I am
But everybody wants you to be just like them
They say sing while you slave and I just get bored
Ah, I ain't gonna work on Maggie's farm no more")
```

```
tibble(line=1:length(MaggieText),  text = MaggieText) %>%

  unnest_tokens(input=text,output = words)%>%

   count(words,sort=TRUE) %>%

   mutate(words = reorder(words,n)) %>%

   filter(n>3)%>%

   # start the plotting part

   ggplot(aes(words,n),size = 7) +   # this says that we are using data words and n

   geom_col() +                      # this is the plot type

   xlab(NULL) +                      # this turns off the x label since it makes things look messy

   theme_minimal() +                  # style choice

   coord_flip()                     # rotate the plot
```

Twitter data

It is useful to look through the help documentation for the rtweet::auth.
Twitter data is somewhat more advanced to obtain.

```
library(ROAuth); library(rtweet)

# Declare Twitter API Credentials

consumer_key = "YOUR API KEY"

consumer_secret = "YOUR API SECRET"

access_token = "YOUR ACCESS TOKEN"

access_secret = "ACCESS TOKEN SECRET"



requestURL = "https://api.twitter.com/oauth/
request_token"

accessURL ="https://api.twitter.com/oauth/
access_token"

authURL ="https://api.twitter.com/oauth/authorize"
```

```
my_oauth =
OAuthFactory$new(consumerKey=consumer_key, # still
need your key

               consumerSecret=consumer_secret, #
still need your secret

               requestURL=requestURL,
accessURL=accessURL, authURL=authURL)

my_oauth$handshake()  # Send R to requested site to
authenticate



#From there you can just use this from now on:

create_token(consumer_key=consumer_key,
consumer_secret=consumer_secret,

       access_token=access_token,
access_secret=access_secret)
```

```
library(rtweet); source("TwitterAuth.R")

Hashtag = "#datascience"

DS = search_tweets(Hashtag,n = 10000) #Maximum number of tweets returned from a single token is 18,000.

SN = search_tweets("#snowmaggedon2020 OR #VancouverSnow", n = 10000)

CU = get_timeline("Carleton_U",n=5000)
```

ggplot basics

We can also look into the number of retweets per tweet for tweets that were retweeted:

```
ggplot(DS,    # data to use,

     aes(retweet_retweet_count)) +  # state which columns to use& makes blank graph

      geom_bar(show.legend = FALSE)   #graph type and its options
```

Or the number of followers that users have:

```
ggplot(DS,    # data to use

    aes(friends_count)) +

 stat_density(show.legend = FALSE)
```

```
#Sort of like Matlab, you can name the plot and add to it later
CoolPlot = ggplot(DS,     # data to use
       aes(friends_count)) +
  stat_density(show.legend = FALSE)
CoolPlot
CoolPlot + xlim(c(0,10000))
CoolerPlot = CoolPlot + xlim(c(0,10000))
plot(CoolerPlot)
```

Data transformations

```
ggplot(DS,     # data to use

            aes(log10(friends_count))) +

               stat_density(show.legend = FALSE, na.rm = TRUE)
```

Coordinate transformations

```
DSno0 = DS[DS$friends_count>0,]

ggplot(DSno0,     # data to use

           aes(friends_count)) +

              stat_density(show.legend = FALSE, na.rm = TRUE) +

               coord_trans(x = "log",y = "identity")
```