Week 9 Homework: Quantitative Corpus Linguistics in R

Vincent Porretta

10/29/2019

This assignment uses music data (artist/song/lyrics/date) that represent Beyonce's Billboard hits. For this assignment you will use the files `beyonce.cvs` and `stoplist_HW.txt` available on Avenue to Learn. 42 points. Due by the start of class on 11/05/2019.

Part 1

1. Read in the data you will require for this assignment, using the code below.

```
beyonce <- read.csv("beyonce.csv", header = T, stringsAsFactors = F)
stoplist <- scan("stoplist_HW.txt", what = character(), sep = "\n")
```

2. Use `count` to make a frequency list from `beyonce` called `beyonce_freq`. Arrange it with the most frequent words on top.
3. Use `beyonce` and `beyonce_freq` to calculate the type-token ratio in Beyonce's lyrics.
4. Remove (i.e., filter out) stopwords (on `stoplist`). Report the top five words that Beyonce uses.
5. Using code from your notes, create bigram frequencies (`beyonce_bigrams`) using the full data set (`beyonce`). What are her top five bigrams?
6. Using code from your notes, make a dataframe of concordance lines (`beyonce_conc`) using the full data set (`beyonce`) that contain the word 'love' with 2 words to the left and 2 words to the right.

Part 2

1. Use the following code to create 2 subcorpora. `beyonce_solo` contains only the songs that beyonce sang alone, while `beyonce_collab` contains only the songs that she sang with another person.

```
beyonce_solo <- beyonce %>%
  filter(is.na(Featuring))
beyonce_collab <- beyonce %>%
  filter(!is.na(Featuring))
```

2. Calculate the frequency of 'oh' and 'baby' in each subcorpus.
3. Use the following code to create a matrix of the frequencies of 'oh' and 'baby' in each subcorpus, by filling in the values you obtained in the previous step.

```
bmat <- matrix(c("solo oh count", "solo baby count",
                 "collab oh count", "collab baby count"), 2, 2,
               dimnames = list(c("Solo", "Collab"), c("oh", "baby")))
```

4. Using the matrix you just created, perform a chi-squared test of association to determine if the distribution of 'oh' and 'baby' are different when she sings alone and when she sings with another person.
5. Use the results to calculate the effect size. Then, visualize the pattern the pattern using `assocplot`. Report what the result means.

6. Calculate the keyness (log likelihood value) of 'me' in Beyonce's solo work `beyonce_solo`, using her collaborative work `beyonce_collab` as a reference corpus. Report what the result means.

7. Calculate the mutual information of 'got me' using frequency information from `beyonce_freq` and `beyonce_bigrams` that you created above. Report what the result means.

8. Run the following code which creates a dataframe of Beyonce's usage of the word 'me' and the ranking the song achieved on the Billboard chart. First make a plot of the relationship. Then calculate the correlation between Beyonce's usage of 'me' and the ranking the song achieved on the Billboard chart (ignore the warning). Report what the result means.

```
ranking <- beyonce %>%
  group_by(Rank, Song) %>%
  count(Word, name = "Freq") %>%
  filter(Word == "me") %>%
  arrange(Rank)
```