Abstract

The third assignment we use the R graphics library to plot the noun phrases used in an official capacity and the noun phrases used in a non-official capacity. This is done through connecting the SQLite database that was created in the previous assignment with R graphics system. With R a double-sided histogram is created comparing the same phrases between the contexts in which they are used.

Data Description

The source data has not changed from the previous assignment except that it is now in a database. The provided Diversity Kaleidoscope data was used. I specifically chose to compare the following authors: [b. University-wide admin, university relations, office of registrar, human resources, etc.] and [a. Students and other individuals (in non-official capacity or not part of university)] These authors were compared in the context of genre #7 for author a.: “When voicing our opinions, e.g., book review, student assignments, newspaper editorials, personal web pages, personal blogs.” And genre #6 for author b.: “Officially / politically, e.g., pages concerning university history, reports, speeches, mission statements, university press releases, non-discrimination statements, initiatives.”

The filtering produced 1024 documents for author (b) genre (6) and 122 documents for author (a) genre (7). This resulted in a total of 1146 documents for analysis.

Process Diary

This data visualization will be using the data from the second homework’s database. However, this data is rather large and it takes quite some time to query it. For the purposes of this assignment we shall use the query results from assignment two. The SQL statements and the data sets can be found on the following page: Homework 2

A second database called homework3.db has been created from these results:

```
hw3freqmakeopiniondb.pl, hw3freqmakeofficialdb.pl

./hw3freqmakeopiniondb.pl < opinion-freq.txt
./hw3freqmakeofficialdb.pl < official-freq.txt
```

The above two Perl scripts have the following data cleansing elements. The first eliminates all words with frequency less than 8. The second eliminates all phrases that are numbers or contain characters that are not in the alphabet. And the third eliminates single letter phrases.

```
next if ($amount < 8);
next if ($phrase !~ m/[A-Za-z]+/);
next if ($phrase =~ m/^([A-Za-z])[1]+$/);
```

The SQL command that populated homework3.db.

```
CREATE TABLE officialPhrases (phrasekey INTEGER PRIMARY KEY, phrase TEXT, amount INTEGER)
CREATE TABLE opinionPhrases (phrasekey INTEGER PRIMARY KEY, phrase TEXT, amount INTEGER)
```
Load the data into the R visualization engine. A script file, load_data.R, was written in the R language to connect to the SQLite database and load the data.

```r
# load the SQLite library
library(RSQLite)

# define connection:
drv <- dbDriver("SQLite")
conn <- dbConnect(drv, dbname = "homework3.db")

# create an object which contains the SQL query:
query <- dbSendQuery(conn, "SELECT official.amount AS OfficialFreq, official.phrase AS Phrase,
opinion.amount AS OpinionFreq FROM (SELECT phrase, amount FROM opinionphrases) as opinion,
(SELECT phrase, amount FROM officialphrases) as official WHERE opinion.phrase = official.phrase
ORDER BY official.amount DESC LIMIT 40")

# fetch data according to query:
data <- fetch(query)

# clear query
dbClearResult(query)

# disconnect
dbDisconnect(conn)
```

The SQL statement above has been duplicated below for readability.

```
SELECT official.amount AS OfficialFreq, official.phrase AS Phrase, opinion.amount AS OpinionFreq
FROM (SELECT phrase, amount FROM opinionphrases) as opinion,
(SELECT phrase, amount FROM officialphrases) as official
WHERE opinion.phrase = official.phrase
ORDER BY official.amount DESC
LIMIT 40
```

A second R script file, graph_data.R, was created to convert the results from the load_data script into visualization. I would like to thank Mark for helping me understand the R commands necessary to create this plot.

```r
# PARAMETERS
# background color
par("bg"="#ccd7d7")

# margins
par("mar"=c(5,4,4,2))

# font size
par("ps"=12)

# foreground color
par("fg"="#303333")

# HISTOGRAMS
# right, official phrases
barplot(data$OfficialFreq, width=1, horiz=TRUE, space=0, col="#FF0000", xlim=c(-1000,6000))

# left, opinion phrases
barplot(-data$OpinionFreq, width=1, horiz=TRUE, space=0, col="#0000FF", add=TRUE)
```
Visualization Results

Much of the word observations were done in the previous two assignments. This assignment allows us to compare the most used similar words between documents written in an official capacity and
documents written primarily as opinions. Due to the significant difference in document output the numbers are heavily skewed in favor of official publications. In the previous assignments we concentrated on the words used by the opinion documents. This visualization concentrates on Official phrases.

We can see immediately which documents are those of opinion form the use of “I” that jumps out at you. When we visualize the results it is far easier to see relationships than just looking at numbers.

There has been further cleansing of the data before reaching this point as stated in the process diary above. All of the stop words are gone as well as some single letters and digits. These were not in the stopwords list, but will be added in a later assignment. It's prudent to possibly cleanse these prior to Monty Lingua.

This ended up leaving “diversity,” “university,” “students,” and “people” as the most used words between the two. It’s also interesting to see that the word “time” near the top of the graph has significant use in the opinion column. It could mean anything, but a reasonable assumption could be that there is a sense of urgency that isn’t recognized by the official organizations.