Abstract

In this assignment I take the lowercase data created in Homework 2 and cleanse it. The difference with this assignment and previous assignments is that the data will be heavily cleansed of illegal characters and converted to a more compatible Monty Lingua file. This includes changing all of the commas to periods. Then the data is loaded into a SQLite database for analysis. I analyzed the number of terms used for opinion and official datasets. I then create an R graph to compare these results with the results from Homework 3.

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

The files for Homework 2 were lowercased in the manner showed below. We then take these and use them for the basis of our cleansing:

```
cat <filename> | tr "[:upper:]" "[:lower:]" > ../<currfoldername>-lc/<filename>
```

In the first and second assignment we didn’t cleanse the data that was going into Monty Lingua. We shall do this now to get more accurate results. The files will be cleaned as much as possible before being sent through Monty Lingua. The script cleanfiles.pl was run for this task. This script called two other scripts: stripweirdchars.pl and cleanse.pl

```
perl cleanfiles.pl
```

Finally we must navigate to the Monty Lingua folder and then run the app for the two resulting text files

```
python test-6.py ~/Private/hw1/opinionconv-lc-clean/ > ~/Private/hw5/opinion-lc-clean-phrases.txt
python test-6.py ~/Private/hw1/officialconv-lc-clean/ > ~/Private/hw5/official-lc-clean-phrases.txt
```
Homework 5, SI 618
Thomas Jeffrey Fleszar

The resulting files then need to be further formatted.

We need to extract only the noun phrases.

grep "\bnp\b" opinion-lc-clean-phrases.txt > np-opinion-lc-clean.txt
grep "\bnp\b" official-lc-clean-phrases.txt > np-official-lc-clean.txt

Next we will modify the data by eliminating stop words, but we first need to fix the data so it matches what the stopwords.pl script needs.

cut np-opinion-lc-clean.txt -f 1,5,6 | awk -F \t '{print $1, \"\t\", $3, \"\t\", $2}' > np-opinion-lc-clean-fix.txt

cut np-official-lc-clean.txt -f 1,5,6 | awk -F \t '{print $1, \"\t\", $3, \"\t\", $2}' > np-official-lc-clean-fix.txt

The stopwords.txt file has been changed to include any single characters other than “I,” and the words “gif,” “www,” “http,” and “spacer.”

Then we run the stopwords.pl script.

./stopwords.pl stopwords.txt < np-opinion-lc-clean-fix.txt > np-opinion-lc-nostop.txt
./stopwords.pl stopwords.txt < np-official-lc-clean-fix.txt > np-official-lc-nostop.txt

Then we want to get the frequency of terms from the opinion and official table. We run the following two scripts that access homework5.db: officialTermCount.pl, opinionTermCount.pl

./opinionTermCount.pl > opinion-freq.txt
./officialTermCount.pl > official-freq.txt

A second database called homework5-1.db has been created from these results:

hw5freqmakeopiniondb.pl, hw5freqmakeofficialdb.pl

./hw5freqmakeopiniondb.pl < opinion-freq.txt
./hw5freqmakeofficialdb.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 single digit phrases.

next if ($amount < 8);
next if ($phrase =~ m/^[\d]{1}$/);

The SQL command that populated homework5-1.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.

# load the SQLite library
library(RSQLite)

# define connection:
drv <- dbDriver("SQLite")
conn <- dbConnect(drv, dbname = "homework5-1.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(-1000,3000))

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

# GRAPH LAYOUT
# axis labels
axis(2, at=1:40, labels=data$Phrase, pos=-850, col.axis="black", las=2, tick=FALSE, hadj=1, padj=1, mgp=c(3,0,0))

# legend
Visualization Results

Homework 3 results below

Homework 5 cleansed results below
Results Observations

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.

In assignment three the filtering and stopword lists eliminated single character words, digits, and personal pronouns to look at the similarities between the two datasets outside of their capacity. The new filtering of the current assignment brings back the personal pronouns and the self-identifier “I.” The new image compared to the old image clearly indicates that that the words “I” and “you” are used a great deal in the opinion documents as compared to the official documents. These words are also used
greatly in the official documents, but the proportions in the opinions is much greater compared to the other phrases.

We also see that “diversity,” “university,” “students,” and “people” are the most used words between the two. What we do not see in these images due to filtering are the high counts of single digits. This indicates that the official documents have a lot of numbered lists.