Abstract

The second assignment used the same diversity kaleidoscope data as the first assignment. The data was provided to the class for analysis. In this assignment I continue to analyze and compare the terms and phrases used by students and other non-officials with university-wide administration. They are compared over the same genre, when each group is voicing their opinion on the subject of diversity. This includes book reviews, student assignments, newspaper editorials, and personal blogs. As the first assignment covered noun word usage based on total occurrences throughout the entire data set. This assignment will build the ability to analyze words and phrases in individual documents, but we can still look at noun word usage based on total occurrences with the proper queries. However, there will be no stop words and the results would be different. We shall see what the differences are, if any.

Data Description

The source data has not changed from assignment one. 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

In order to obtain the needed data the sequence from the document “Diversity Kaleidoscope Fetching and Converting Documents” was used. Below is the primary sequence that was used for populating the dataset.

We will be using the provided “author-genre-name-key.txt” file to help us match the author and genre with the proper document file.

First for us to filter for Author (B) and Genre (6) we need to run the following:

```
grep "b|6" ../author-genre-name-key.txt | cut -d \| -f 5 | sed 's/\^/cp /g' | sed 's/\$\$' ..//officialconv#g' > official.sh
```

The above command searches for “b|6” in the key text and then extracts the file name. It then creates a shell script that will be run later to copy the extracted files to an official converted folder. A sample line from the script is as follows:
Then we need to copy the files into a folder for Monty Lingua processing:

```
chmod +x official.sh
./official.sh
```

Second we must filter for Author (A) and Genre (7). For this we run the following command:

```
grep "|a|7|" ../author-genre-name-key.txt | cut -d \| -f 5 | sed 's/a/\cat /g' sed 's/\#$/../opinionconv\#/g' > opinion.sh
```

The above command searches for “a|7” in the key text and then extracts the file name. It then creates a shell script that will be run later to copy the extracted files to an opinion converted folder. A sample line from the script is as follows:

```
cp yale-055 ../opinionconv
```

Then we need to copy the files into a folder for Monty Lingua processing:

```
chmod +x opinion.sh
./opinion.sh
```

After the files arrive in their respective folders we must do some pre-processing before we can send them off to MontyLingua. MontyLingua is case sensitive and will treat words with capital letters differently than the same word that is all lowercase. So, in this case the following two words “problem” and “Problem” will be counted and treated as separate term. This may be important in cases where a word is capitalized because it is a part of a title, but for the purposes of this assignment we will not worry about those situations.

To lowercase all of the data in the files a script was created for each set of opinion and official files. The script was populated with commands in the following format:

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

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/ > ~/Private/hw1/opinion/opinion-lc-phrases.txt
python test-6.py ~/Private/hw1/officialconv-lc/ > ~/Private/hw1/official/official-lc-phrases.txt
```

The resulting files then need to be further formatted.

We need to extract only the noun phrases.

```
grep "\bnp\b" opinion-lc-phrases.txt > np-opinion-lc.txt
grep "\bnp\b" official-lc-phrases.txt > np-official-lc.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.txt -f 1,5,6 | awk -F '\t' '{print $1, '\t', $3, '\t', $2}' > np-opinion-lc-fix.txt
```
cut np-official-lc.txt -f 1,5,6 | awk -F '\t' '{print $1, \"\t\", $3, \"\t\", $2}' > np-official-lc-fix.txt

Then we run the stopwords.pl script.

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

Now it's time to populate the SQLite database with makeopiniondb.pl and makeofficialdb.pl.

CREATE TABLE opinionPhrases (phrasekey INTEGER PRIMARY KEY, doc TEXT, phrase TEXT, amount INTEGER)
CREATE TABLE officialPhrases (phrasekey INTEGER PRIMARY KEY, doc TEXT, phrase TEXT, amount INTEGER)

We then want to build the results that were created in homework #1 with phrase count, except this time there are no stop words. The following SQL Query was used.

SELECT phrase, SUM(amount) as TotalAmount FROM opinionphrases GROUP BY phrase ORDER BY TotalAmount DESC, phrase ASC

SELECT phrase, SUM(amount) as TotalAmount FROM officialphrases GROUP BY phrase ORDER BY TotalAmount DESC, phrase ASC

Results

The following two links contain the query results from the noun phrases in both the official capacity and non-official opinion capacity. Those of less than eight (8) occurrences were not kept. I decided to look at the same information as in homework #1. It isn't useful for visualization or comparing two documents, but the data in the database is designed to properly assist in building visualizations.

1.

2.

Results Observations

We looked at the differences in noun word usage between authors of an official capacity writing for official purposes and authors of a non-official capacity writing opinion articles. Both of these data sets come from articles about diversity in the University system.

When we look at the two lists we notice that the two datasets become far more similar when the stop words are eliminated. The top few of each list was remarkably similar. They both use “diversity,” “university,” “students,” “minorities,” and other diversity and people related words. This shows that stop words in the context of opinions are a valuable tool, but also that looking at just how many terms are in an entire set of documents is not a very good measure of comparing document sets. One document can skew the numbers.
However, we if we look more deeply into the list the students also use words as “issues,” “problems.” The word “issues” is also shared with the Official writings, but not the word “problem” or “problems.” As was mentioned in the first assignment it’s hard to draw any conclusions from this other than a varying definition of “issues” between opinion and official sources.