Christopher Kruger
EM 624
15 November 2023

Analysis of Two Articles – Narrative

The two articles I chose covered the tragedy of Adam Johnson's death and the emerging news of the manslaughter charges. I purposefully chose the BBC and NPR articles as each national magazine covers the perspective from two different countries (UK vs. USA) on the topic. Using the beautiful soup python library for parsing through the html data for both websites, I was able to print the titles of each article, get all body text, the bigrams, sentiment distributions, and the wordcloud for the most common words.

The program begins by displaying the title of the two articles, the amount of words used in each article, as well as the raw text of the article as a list for further processing. For reference, the BBC article is "article 1" and the NPR article is "article 2".

```
 ---- ORIGINAL ARTICLES ----

 The title of the first article is:
  Adam Johnson: Manslaughter arrest over ice hockey player's death - BBC News

 Full text for article 1 ( 705 words):

 ['A', 'man', 'has', 'been', 'arrested', 'on', 'suspicion', 'of', 'manslaughter', 'over',

The title of the second article is:
 Ex-NHL player Adam Johnson skate death: Man released on bail : NPR

Full text for article 2 ( 418 words):

['By', 'Ayana', 'Archie', 'Former', 'Pittsburgh', 'Penguins', 'forward', 'Adam', 'Johnson', 'in',
```

The program then cleans up the text by removing all stop words, words with less than three characters, and additional words, namely "adam", "johnson", and "hockey" which would are used repeatedly between each article.

```
---- CLEAN ARTICILES ----

Clean article 1:

['arrested', 'suspicion', 'manslaughter', 'death', 'player', 'neck', 'match', 'nottingham', 'panthers', '

Clean article 2:

['ayana', 'archie', 'pittsburgh', 'penguins', 'forward', 'action', 'game', 'columbus', 'ohio', 'police',
```

The program then calculates the bigrams for each article. The bigrams show the user which two words, as a consecutive pair, are written the most. The bigrams shown are in the format of "word1_word2" for ease of visibility for the user and only shows the bigrams that are used more than three times.

```
The bigrams used more than 3 times (and formatted to x_y) in article 1 are:
 ['neck_guards', 'south_yorkshire', 'yorkshire_police']
The bigrams used more than 3 times (and formatted to x_y) in article 2 are:
 ['connection_death', 'pittsburgh_penguins', 'south_yorkshire', 'suspicion_manslaughter']
```

For further processing, the bigrams and total article text are combined as one larger list.

```
The combined total article and bigrams for article 1 is:
 ['neck_guards', 'south_yorkshire', 'yorkshire_police', 'arrested', 'suspicion', 'manslaughter', 'death', 'player', 'neck',

The combined total article and bigrams for article 2 is:
 ['connection_death', 'pittsburgh_penguins', 'south_yorkshire', 'suspicion_manslaughter', 'ayana', 'archie', 'pittsburgh',
```

Sentiment distributions are then calculated for each article. This tells the user if the word choices were positive, negative, or neutral in tone. The BBC article was calculated first, with a positive ratio of 9.7%, negative ratio of 16.1%, and a neutral ratio of 74.2%. Next, the NPR article was calculated, with values of 2%, 32.6%, and 65.4% respectively. I found it interesting that the article topic should ideally be mostly equal in neutral tone for stating facts, but the NPR article had double the amount of the negative word ratio, with less positive and neutral tonnage of words.

```
The following is the distribution of the sentiment for article 1:


It is positive for 9.7%

It is negative for 16.1%

It is neutral for 74.2%

The following is the distribution of the sentiment for article 2:


It is positive for 2.0%

It is negative for 32.6%

It is neutral for 65.4%
```

Wordclouds were then generated for each article. The wordcloud for article 1 showed why it was more neutral in tone, as the most common words (outside of the stopwords provided) were “neck”, “player”, “death”, and “league”. In comparison to article 2, the most common words were “police”, “death”, “game”, and “connection.” In both wordclouds, it is clear

too that both articles discussed common themes such as “panthers” (which was the team he played for), “Sheffield” (the town that represents the panthers), as well as other general words such as “investigation”, “hospital,” and “England.”

It is important to note though that article 1 mistakenly also recorded words for the sideline articles on BBC such as topics on the latest conflict in Israel. I was unable to remove these, and do not represent the two article’s themes which are shown in the wordclouds.

Article 1 (BBC)

Article 2 (NPR)