CSCI 4152/6509

P1

Ziyu Qiu, B00791470

Cooper Gagnon, B00767631

Matthew Moore, B00767194

March 22, 2021

Project Title:

Sentiment Analysis of Named Entities Taken From Live RSS News Feeds

Problem Statement:

Record average sentiment of given named entities within a large news dataset. This custom dataset will be created from scraping several thousand news RSS feeds. Potentially research how sentiment of named entities changes overtime.

1 Project Plan:

1.1 Acquire Testing Data

As for an existing dataset, we are planning to use a data set containing 1 million news headlines [2]. We can use this as our initial training set. After we have successfully implemented a script that handles average sentiment analysis of named entities on this data set, we will then move onto testing on 'live data' (Covered in Create Custom Dataset).

1.2 Implement Named Entity Extraction

Using the NLTK library we will tag all the words using a POS tagger from NLTK. We will then extract all named entities within a given sentence which we will use later to link to a sentiment analysis.

1.3 Implement Average Named Entity Sentiment Analysis

We need to take each sentence and get its score using NLTK's sentiment analysis library [3]. We will link these sentiment values to the named entities of the given sentence. We will then record all the sentiments as an average in a hash array using named entities as keys.

1.4 Create Custom Dataset

Using a text list of all news websites used by Google News we need to scrape all the websites to find their RSS feeds [1]. We then need to scrape all of the RSS feeds in order to create a live data set that is relevant to the current date.

1.5 Create User Interface for Searching Named Entities

We will create a basic interface by which we can search for a given named entity and receive its average sentiment analysis over the entire dataset.

2 Relevant Work and Approaches

- `https://www.ijcaonline.org/archives/volume178/number46/oswal-2019-ijca-919367.pdf`
    - Average sentiment analysis of named entities found on twitter.

References

[1] Agarwal Amit, 'Which Sites and Blogs are Indexed by Google News?', July. 2011 `https://www.labnol.org/internet/sites-indexed-in-google-news/19323/`

[2] Kulkarni Rohit 'A Million News Headlines' Kaggle, Feb. 2021, `https://www.kaggle.com/therohk/million-headlines`

[3] Sentiment Analysis `https://www.nltk.org/howto/sentiment.html`

[4] Newsboat. "Newsboat RSS feed reader." GitHub, `github.com/newsboat/newsboat`