

CSC 411: Assignment #3

Xiangyu Kong
kongxi16

Yun Lu
luyun5

March 12, 2018

Problem 1

The dataset contains headlines including the word “Trump”. The quality of the dataset is generally good since there is a large variety of vocabularies included.

By observing, we can see that the fake news’ headlines are generally longer than the real news’ headlines.

Except for “Donald” and “Trump”, the most frequent words are “to”, “for”, etc, but they do not mean a lot. The top meaningful words in total are “Clinton”, “election” and “president”. This infers that most news are regarding the 2017 presidential election and especially between Donald Trump and Hilary Clinton.

The statistics for the three words are generated through part1 in fake.py and the results are given below in Listing 1. We can see that although “Clinton” has the largest word count, most of them appear in the fake news. Reports that include “election” are more likely to be real news.

Listing 1: statistic results

```
[clinton]:
    real: 83
    fake: 132
    total: 215
[election]:
    real: 87
    fake: 74
    total: 161
[president]:
    real: 66
    fake: 64
    total: 130
```

Problem 2

Problem 3

Problem 4

Problem 5

Problem 6

Problem 7

Problem 8