CSC 411: Assignment #3

Xiangyu Kong **Yun Lu**
kongxi16 **luyun5**

March 14, 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

use set instead of list to avoid situation where word appear twice in headline making the probability negative and the log of it will be negative
First tried m from 1 to 10 and p_hat from 0.05 - 1, but best value is m = 9 and p = 0.95, then tried m from 1 to 20 and p from 0.05 - 1, then m = 15, p = 0.95

Problem 3

The results are produced by part3 in fake.py and are listed below.

$P(c|word) = \frac{P(word|c) \times P(c)}{P(words)}$ where
$P(c) = \frac{count(c)}{count(total)}$, $P(word|c) = \frac{count(word\ in\ c)}{count(c)}$ and $P(word) = \frac{count(word\ in\ c)}{count(total)}$

$P(c|notword)$ follows the similar calculations.

The most important presence for predicting a class means $P(c|word)$ must be high and the most important absence for predicting a class means $P(c|notword)$ must be high.

```
a:
Real:
top 10 important presence:
        ['be', 'pardon', 'felt', 'four', 'protest', 'asian',
        'aides', 'liar', 'hate', 'marching']
top 10 important absence:
        ['hats', 'pide', 'hath', 'sleep', 'deri', 'hating',
        'captain', 'saved', 'assembled', 'kommonsentsjane']
Fake:
top 10 important presence:
        ['nobel', 'hats', 'pide', 'colleges', 'four', 'hath',
        'sleep', 'deri', 'hating', 'captain']
top 10 important absence:
        ['manaforts', 'asian', 'aides', 'marching', 'stinks',
         'protections', 'kidman', 'sorry', 'whack', 'softener']

b:
Real:
top 10 important presence:
        ['pardon', 'felt', 'protest', 'asian', 'aides', 'liar',
        'hate', 'marching', 'stinks', 'votes']
Fake:
top 10 important presence:
        ['nobel', 'hats', 'pide', 'colleges', 'hath', 'sleep',
        'deri', 'hating', 'captain', 'hate']
```

Problem 4

Problem 5

Problem 6

Problem 7

Problem 8