CMPS142 – Spring 2018 – Homework 3

Kevin Romero – kromero8@ucsc.edu – 1635745
Spencer Peterson – spjpeter@ucsc.edu – 1544868

Handed out: May 20th, 2018
Due: May 26th, 2018

Problem 1

1. Processing the training set
   (a) The total number of distinct token types in our training set after step two is 8,701.
   (b) The list of tokens we get for the string "Ive been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing at all times." after step five is as follows:

```
["'ve", "search", "right", "word", "thank", "breather", "promis", "wont",
"take", "help", "grant", "fulfil", "promis", "wonder", "bless", "time"]
```

   (c) The total number of tokens we have after step 6, our $|V|$, is 1,363.
2. Step 7: Feature vector representation
   (a) See attached 'HW3_Romero_train.csv'.
   (b) There is no column named "yellow".
   (c) There is a column named "music".
   (d) The sum of row two is 15.
   (e) No columns sum to 0.
3. Processing the test set
   (a) The SMS text that starts with "Hi. Wk been ok" was turned into the following tokens:

```
["hi", "wk", "ok", "ye", "bit", "run", "forgot", "need", "get", "home",
        "n", "caus", "u", "''"]
```

   (b) See attached 'HW3_Romero_test.csv'.
   (c) There are 1116 rows in the test data CSV.
   (d) There are 1364 columns, including the label.
   (e) The first five columns in the train file are "photo", "code", "forget", "bun", "celebr". The first five columns in the test file are "photo", "code", "forget", "bun", "celebr".
   (f) The headers do contain the token "head".
   (g) 884 columns sum to 0 in the test data.
   (h) It is important to have train and test data represented in the same feature space because otherwise the model generated on the training data wouldn't apply to the test data. Whatever model we have learn will likely learn weights that only inherently have meaning when applied to feature vectors that are identical to the training data. Otherwise, the dot products of the weight vector and arbitrary feature vectors would be meaningless and not give us anything close to accurate results.
4. Submitting code: The zip file containing our code is named 'HW3_Romero_code.zip'.