CMPS 142 Machine Learning
Spring 2018, Homework #3

Aaron Steele, atsteele@ucsc.edu
Tommy Tran, ttran56@ucsc.edu

1.1: Preprocessing the Training set

1.1.2 (a)

Distinct tokens: len(set(allTokens))
Output: Distinct tokens: 8703

1.1.5 (a)

Output:

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

1.1.6 (a)

Total number of distinct tokens in the token set:
Vocabulary: 1184

1.1.7

(a)

HW3_steele_train.csv

(b)

Yes, it does

(c)

Yes, it does

(d)

Line number 2 sum = 7

(e)

1.3

HW3_steele_code.zip