

Homework 8

Ryan Kingery

STAT 5014
Dr. Settlage

1 Problem 2

Since everyone loves a word cloud, why not do that? The following Python code loads in the file `survey_data.txt` as a string `doc`. Some minor text processing is then done to make the word cloud more representative. After this, a word cloud object is generated for `doc` using the Python wordcloud package, and displayed using the `matplotlib.pyplot.imshow` function. The code is shown below.

```
import matplotlib.pyplot as plt
import textmining as tm
from wordcloud import WordCloud

DIR = '~/STAT_5014/08_text_mining_Rnotebooks_bash_sed_awk/survey_data.txt'

def process_doc(doc):
    doc = doc.lower()
    doc = doc.replace('intermediate', 'int')
    doc = doc.replace('intermed', 'int')
    doc = doc.replace('int', 'intermediate')
    doc = doc.replace('beginner', 'beg')
    doc = doc.replace('beg', 'beginner')
    doc = doc.replace('master', 'ms')
    doc = doc.replace('obj-c', 'objC')
    doc = doc.replace('c++', 'cpp')
    doc = doc.replace('/', ' ')
    doc = doc.replace('-', ' ')
    for stopword in tm.stopwords:
        doc = doc.replace(' '+stopword+' ', ' ')
    return doc

doc = open(DIR).read()
doc = process_doc(doc)
```

