HW8

Guangya Wan

March 29, 2019

```
## /home/guangya/anaconda3/bin/python:1: SettingWithCopyWarning:
## A value is trying to be set on a copy of a slice from a DataFrame.
## Try using .loc[row_indexer,col_indexer] = value instead
##
## See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#inde
```

```
x = np.array([2,3])
print(x)
```

```
## [2 3]
```

```
for i in data['text']:
    s.append(i)
rank_word = sorted(np.sum(np.array(vector.fit_transform(s).toarray()),axis = 0))[::-1]
plt.scatter(x = range(len(rank_word)),y = rank_word,label = 'counts')
plt.xlabel('Wordrank')
plt.ylabel('Wordcount')
plt.title('word frequency')
plt.legend()
count_vec = np.sum(np.array(vector.fit_transform(s).toarray()),axis = 0)
# I decide to make stop word as the top5 word
remove_index = []
for i in rank_word[:10]:
    remove_index.append(np.argwhere(count_vec == i)[0][0])
stop_words = []
for k,v in vector.vocabulary_.items():
    if(v in remove_index):
        print(k) # stop words
        stop_words.append(k)
```

```
## for
## the
## in
## to
## was
## my
## is
## and
## of
## it
```

```
vector = CountVectorizer(stop_words=stop_words, max_df=0.95,min_df= 5)
rank_word = sorted(np.sum(np.array(vector.fit_transform(s).toarray()),axis = 0))[::-1]
```