gas_cloud

October 4, 2019

1 Gas Cloud

In this activity you will create your own word clouds using a subset of the Reuters corpus.

[1]:
```
from nltk.corpus import stopwords, reuters
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
from wordcloud import WordCloud
import re
import matplotlib.pyplot as plt

# Code to download wordnet corpora
import nltk
nltk.download('wordnet')

lemmatizer = WordNetLemmatizer()
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     /Users/josearturomorasoto/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

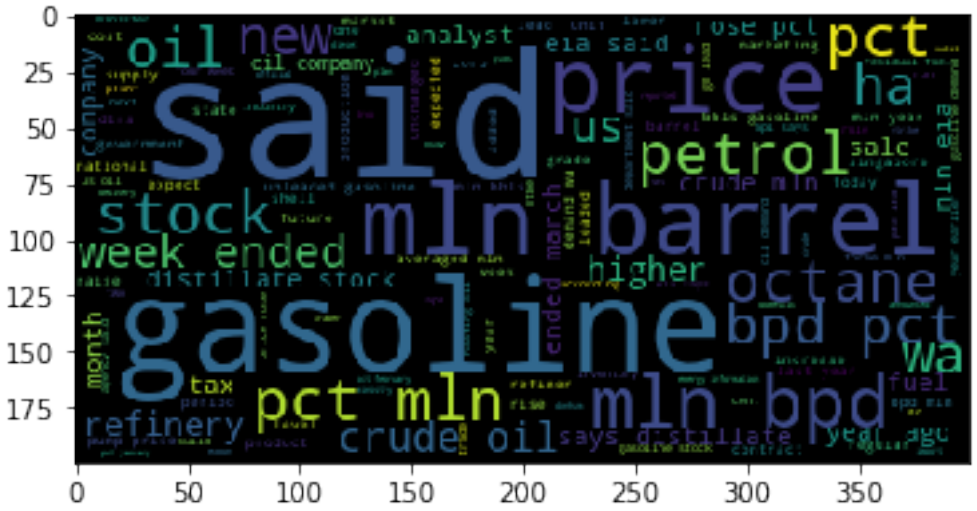
[2]:
```
ids = reuters.fileids(categories='gas')
corpus = [reuters.raw(i) for i in ids]
```

[3]:
```
def process_text(doc):
    sw = set(stopwords.words('english'))
    regex = re.compile("[^a-zA-Z ]")
    re_clean = regex.sub('', doc)
    words = word_tokenize(re_clean)
    lem = [lemmatizer.lemmatize(word) for word in words]
    output = [word.lower() for word in lem if word.lower() not in sw]
    return ' '.join(output)
```

[4]:
```
big_string = ' '.join(corpus)
input_words = process_text(big_string)
```

[5]:
```
wc = WordCloud().generate(input_words)
plt.imshow(wc)
```

[5]: <matplotlib.image.AxesImage at 0x1a21b86dd8>



1.1 Challenge

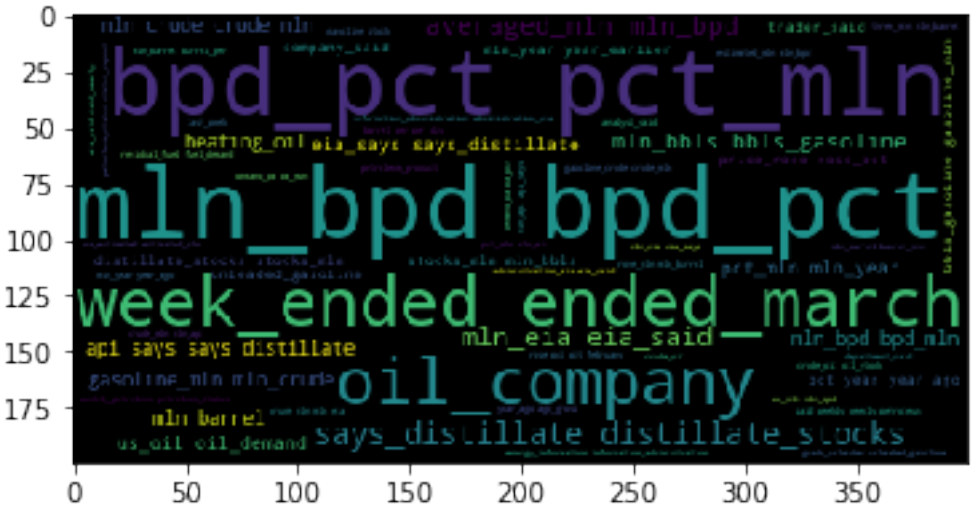
[6]:
```
from nltk.util import ngrams
```

[7]:
```
def process_text_bg(doc):
    sw = set(stopwords.words('english'))
    regex = re.compile("[^a-zA-Z ]")
    re_clean = regex.sub('', doc)
    words = word_tokenize(re_clean)
    lem = [lemmatizer.lemmatize(word) for word in words]
    sw_words = [word.lower() for word in lem if word.lower() not in sw]
    bigrams = ngrams(sw_words, 2)
    output = ['_'.join(i) for i in bigrams]
    return ' '.join(output)
```

[8]:
```
input_bigrams = process_text_bg(big_string)
```

[9]:
```
wc = WordCloud().generate(input_bigrams)
plt.imshow(wc)
```

[9]: <matplotlib.image.AxesImage at 0x1a1cfa1c50>


bpd_pct pct_mln
mln_bpd bpd_pct
week_ended ended_march
oil_company
says_distillate distillate_stocks
api says says distillate
mln_eia eia_said
mln_bpd bpd_mln
mln_hhis hhis_gasoline
mln barrel
us_oil oil_demand
0
25
50
75
100
125
150
175
0
50
100
150
200
250
300
350