wordcloud

October 4, 2019

1 Word Clouds

[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]   Unzipping corpora/wordnet.zip.
```

[2]:

```
ids = reuters.fileids(categories='gold')
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]:

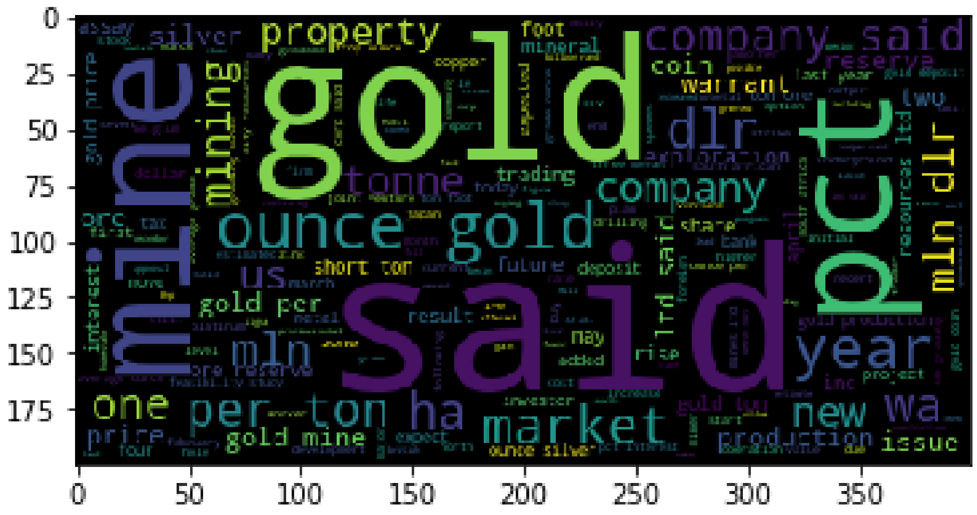
```
# Process text for wordcloud creation
big_string = ' '.join(corpus)
input_text = process_text(big_string)
```

[5]:

```
wc = WordCloud().generate(input_text)
plt.imshow(wc)
```

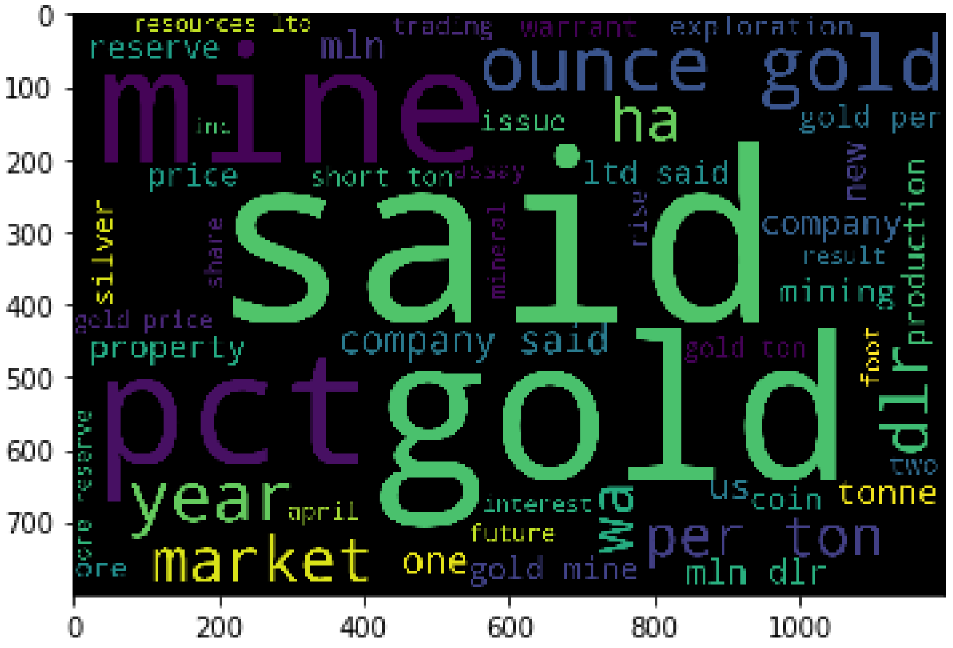
[5]:

```
<matplotlib.image.AxesImage at 0x1a22e11240>
```



```
[6]: wc = WordCloud(width=1200, height=800, max_words=50).generate(input_text)
     plt.imshow(wc)
```

```
[6]: <matplotlib.image.AxesImage at 0x1a22e94550>
```


resources ltd
trading
warrant
exploration
reserve
mln
ounce gold
mine
ha
issue
gold per
inc
price
short ton
assay
ltd said
new
said
company
silver
share
mineral
rise
result
production
mining
gold price
property
company said
gold ton
foot
pct
gold
dlr
reserve
two
us
coin
tonne
year
april
interest
wa
future
per ton
ore
market
one
gold mine
mln dlr