read_me

August 15, 2020

1 SEC TEXT: NLP

1.1 A codebase to allow for search and additional Natural Language Processing work by analysts and developers for SEC 10K and 10Q filings.

The product is built in four files.

The first 2 are written in R and use the edgarWebR library from https://mwaldstein.github.io/edgarWebR/, a well-maintained and popular library for pulling SEC documents and slicing the SEC's unique XBRL approach into readable sections.

sec_nlp_getter.R reads a list of tickers from a local csv file, column named Symbol, and a) retrieves all filings from the SEC for that symbol, b) saves the base HTML document in a file tokenized (split) by sentences c) parses the base document into MDNA and Risk Factor sections d) creates a local file filing_index.csv which stores the location of each document for each ticker.

sec_R_utils.R is the utility file for sec_nlp_getter.

The second 2 files are written in Python and use NLTK and pattern libraries to apply sentiment analysis to the extracted documents.

(sample_workflow.py is a sample file to show the various combinations available to the SECTextNLP class)

sec_text_nlp.py contains the SECTextNLP class.

sec_nlp_utils.py is the utility file for sec_text_nlp.py.

1.2 The following workflow is an example of the use of the SECTextNLP class.

Setup: download git into a directory, unzip the archive.zip file in the git directory. Import py file as below.

Instantiate an SECTextNLP object with a ticker. Select ticker, period date, form name, and type from the df_file_index object to see what documents are available.

[1]:
```
from sec_text_nlp import *

stn = SECTextNLP("AAPL")
stn.df_file_index[['ticker','period_date','form_name','type']].head()
```

[1]:
```
    ticker           period_date  \
139   AAPL  2020-06-27T04:00:00Z
140   AAPL  2020-03-28T04:00:00Z
141   AAPL  2019-12-28T05:00:00Z
142   AAPL  2019-09-28T04:00:00Z
143   AAPL  2019-06-29T04:00:00Z

                                        form_name  type
139           Quarterly report [Sections 13 or 15(d)]  10-Q
140           Quarterly report [Sections 13 or 15(d)]  10-Q
141           Quarterly report [Sections 13 or 15(d)]  10-Q
142  Annual report [Section 13 and 15(d), not S-K I…  10-K
143           Quarterly report [Sections 13 or 15(d)]  10-Q
```

Get the fully concatenated text of the base document Includes the file name, and href is the reference key for use in other joins.

[2]:
```
stn.df_text.head()
```

[2]:
```
  part.name item.name                                      sentence_text  \
0       NaN       NaN  united states securities and exchange commissi…
1       NaN       NaN  20549 form 10-q (mark one)     quarterly repor…
2       NaN       NaN       commission file number: 001-36743 apple inc.
3       NaN       NaN  (exact name of registrant as specified in its …
4       NaN       NaN                         employer identification no.)

                         file  \
0  aapl-20200627_sentences.csv
1  aapl-20200627_sentences.csv
2  aapl-20200627_sentences.csv
3  aapl-20200627_sentences.csv
4  aapl-20200627_sentences.csv

                                                 href
0  https://www.sec.gov/Archives/edgar/data/320193…
1  https://www.sec.gov/Archives/edgar/data/320193…
2  https://www.sec.gov/Archives/edgar/data/320193…
3  https://www.sec.gov/Archives/edgar/data/320193…
4  https://www.sec.gov/Archives/edgar/data/320193…
```

For more reference information, join with the file index on the href key. Now you can see filing_date.

[3]:
```
pd.merge(stn.df_text,stn.df_file_index,how =␣
 ↪'inner',left_on='href',right_on='href')[['ticker','filing_date','sentence_text']].
 ↪head()
#add filing type
```

[3]:
```
  ticker           filing_date  \
0   AAPL  2020-07-31T04:00:00Z
1   AAPL  2020-07-31T04:00:00Z
2   AAPL  2020-07-31T04:00:00Z
3   AAPL  2020-07-31T04:00:00Z
4   AAPL  2020-07-31T04:00:00Z

                                       sentence_text
0  united states securities and exchange commissi…
1  20549 form 10-q (mark one)     quarterly repor…
2       commission file number: 001-36743 apple inc.
3  (exact name of registrant as specified in its …
4                        employer identification no.)
```

Or, if you only want Management Discussion and Analysis: Join with the df_mdna object instead of df_text.

[4]:
```
pd.merge(stn.df_mdna,stn.df_file_index,how =␣
↪'inner',left_on='href',right_on='href')[['ticker','filing_date','sentence_text']].
↪head()
```

[4]:
```
  ticker           filing_date  \
0   AAPL  2020-07-31T04:00:00Z
1   AAPL  2020-07-31T04:00:00Z
2   AAPL  2020-07-31T04:00:00Z
3   AAPL  2020-07-31T04:00:00Z
4   AAPL  2020-07-31T04:00:00Z

                                       sentence_text
0                                           item 2.
1  management's discussion and analysis of financ…
2  forward-looking statements provide current exp…
3  for example, statements in this form 10-q rega…
4  forward-looking statements can also be identif…
```

Pass in a topic, and get noun phrases around that topic. Example: Get business segments.

[5]:
```
stn.get_noun_phrases_around_topic(BUSINESS_SEGMENT_LIST)
```

[5]:
```
['americas segment',
 'asia pacific',
 'asia pacific segment',
 'china segment',
 'distribution partners',
 'europe segment',
 'geographic segment',
 'hong kong',
 'retail stores',
```

```
 'software products']
```

Same approach, get products.

[6]:
```
stn.get_noun_phrases_around_topic(PRODUCTS_LIST,nrows=2)
```

[6]:
```
['app store',
 'app store®',
 'apple music',
 'apple music®',
 'apple pay',
 'apple pay®',
 'apple tv',
 'apple tv®',
 'apple watch®',
 'book store',
 'delivers digital content',
 'digital content',
 'icloud backup',
 'icloud drive®',
 'icloud icloud',
 'icloud keychain®',
 'icloud photos',
 'icloud services',
 'ios devices',
 'itunes store',
 'itunes store®',
 'mac app store',
 'multiple ios devices',
 'personal computers',
 'professional software applications',
 'stores music',
 'support offerings',
 'support options',
 'tv app store']
```

Or just get a list of all trademark items:

[7]:
```
stn.get_words_with_trademark(stn.df_mdna)
```

[7]:
```
['airpods ',
 'card ',
 'watch®',
 'air®',
 'beats®',
 'pro®',
 'watchos®',
 'mini®',
```

```
 'homepod ',
 'macos®',
 'ipados®',
 'arcade ',
 'macos ',
 'pay®',
 'pencil®',
 'tvos®',
 'iphone®',
 'music®',
 'ipad®',
 'bar ',
 'store ',
 'ipod®',
 'ipados ',
 'folio ',
 'x®',
 'pro ',
 'imac®',
 'touch®',
 'tv®',
 'retina®',
 'mac®',
 'applecare®',
 'icloud®',
 'macbook®',
 'store®',
 'tvos ']
```

Pair keyword search list with SENTIMENT … Using NLTK sentiment analyzer 'MACRO' is the column name of the new dataframe

[8]:
```
stn.match_keywords(stn.NLTK_sentiment(stn.df_mdna),GLOBAL_SEARCH_LIST,'MACRO')
```

[8]:
```
                                part.name  \
0    PART I   -   FINANCIAL INFORMATION
1    PART I   -   FINANCIAL INFORMATION
2    PART I   -   FINANCIAL INFORMATION
3    PART I   -   FINANCIAL INFORMATION
4    PART I   -   FINANCIAL INFORMATION
..                                    …
267                            PART II
268                            PART II
269                            PART II
270                            PART II
271                            PART II
```

```
                                              item.name  \
0    Item 2. Management's Discussion and Analysis o…
1    Item 2. Management's Discussion and Analysis o…
2    Item 2. Management's Discussion and Analysis o…
3    Item 2. Management's Discussion and Analysis o…
4    Item 2. Management's Discussion and Analysis o…
..                                                 …
267  Item 7. Management's Discussion and Analysis o…
268  Item 7. Management's Discussion and Analysis o…
269  Item 7. Management's Discussion and Analysis o…
270  Item 7. Management's Discussion and Analysis o…
271  Item 7. Management's Discussion and Analysis o…

                                       sentence_text     section  \
0                                           item 2.  discussion
1    management's discussion and analysis of financ…  discussion
2    forward-looking statements provide current exp…  discussion
3    for example, statements in this form 10-q rega…  discussion
4    forward-looking statements can also be identif…  discussion
..                                                 …         …
267  in the opinion of management, there was not at…  discussion
268  however, the outcome of legal proceedings and …  discussion
269  therefore, although management considers the l…  discussion
270                                      apple inc.  discussion
271           | 2015 form 10-k | 35 table of contents  discussion

                     file  \
0    aapl-20200627_mdna.csv
1    aapl-20200627_mdna.csv
2    aapl-20200627_mdna.csv
3    aapl-20200627_mdna.csv
4    aapl-20200627_mdna.csv
..                      …
267    d17062d10k_mdna.csv
268    d17062d10k_mdna.csv
269    d17062d10k_mdna.csv
270    d17062d10k_mdna.csv
271    d17062d10k_mdna.csv

                                                  href    neg    neu    pos  \
0    https://www.sec.gov/Archives/edgar/data/320193…  0.000  1.000  0.000
1    https://www.sec.gov/Archives/edgar/data/320193…  0.173  0.767  0.060
2    https://www.sec.gov/Archives/edgar/data/320193…  0.000  0.884  0.116
3    https://www.sec.gov/Archives/edgar/data/320193…  0.000  1.000  0.000
4    https://www.sec.gov/Archives/edgar/data/320193…  0.000  1.000  0.000
..                                                 …    …    …    …
267  https://www.sec.gov/Archives/edgar/data/320193…  0.251  0.582  0.167
```

```
268  https://www.sec.gov/Archives/edgar/data/320193…  0.189  0.551  0.260
269  https://www.sec.gov/Archives/edgar/data/320193…  0.104  0.765  0.131
270  https://www.sec.gov/Archives/edgar/data/320193…  0.000  1.000  0.000
271  https://www.sec.gov/Archives/edgar/data/320193…  0.000  1.000  0.000

     compound      MACRO  MACRO_number
0      0.0000        NaN           NaN
1     -0.4767        NaN           NaN
2      0.2732        NaN           NaN
3      0.0000  ['covid']           1.0
4      0.0000        NaN           NaN
..        …         …             …
267   -0.3182        NaN           NaN
268   -0.0258        NaN           NaN
269   -0.0258        NaN           NaN
270    0.0000        NaN           NaN
271    0.0000        NaN           NaN

[4100 rows x 12 columns]
```

Get trademark list, and get the sentiment for each word from the MDNA text.

[9]:
```
list_trademarks = stn.get_words_with_trademark(stn.df_mdna)
df = stn.match_keywords(stn.NLTK_sentiment(stn.
 ↪df_mdna),list_trademarks,'trademarks')
df.dropna()
```

[9]:
```
                              part.name  \
22  PART I   -   FINANCIAL INFORMATION
25  PART I   -   FINANCIAL INFORMATION
27  PART I   -   FINANCIAL INFORMATION
28  PART I   -   FINANCIAL INFORMATION
32  PART I   -   FINANCIAL INFORMATION
..                                   …
50                             PART II
51                             PART II
52                             PART II
53                             PART II
54                             PART II

                                              item.name  \
22  Item 2. Management's Discussion and Analysis o…
25  Item 2. Management's Discussion and Analysis o…
27  Item 2. Management's Discussion and Analysis o…
28  Item 2. Management's Discussion and Analysis o…
32  Item 2. Management's Discussion and Analysis o…
..                                                 …
```

```
50  Item 7. Management's Discussion and Analysis o…
51  Item 7. Management's Discussion and Analysis o…
52  Item 7. Management's Discussion and Analysis o…
53  Item 7. Management's Discussion and Analysis o…
54  Item 7. Management's Discussion and Analysis o…

                                       sentence_text     section  \
22  the covid-19 pandemic has significantly curtai…  discussion
25  the company is working on safely re-opening it…  discussion
27  the most pronounced impact occurred in april 2…  discussion
28  the full extent of the future impact of the co…  discussion
32  third quarter fiscal 2020 highlights total net…  discussion
..                                                …         …
50  the year-over-year growth in mac net sales and…  discussion
51  mac net sales and unit sales increased in all …  discussion
52  mac asps decreased during 2014 compared to 201…  discussion
53                                      apple inc.   discussion
54  | 2015 form 10-k | 25 table of contents servic…  discussion

                     file                                                href  \
22  aapl-20200627_mdna.csv  https://www.sec.gov/Archives/edgar/data/320193…
25  aapl-20200627_mdna.csv  https://www.sec.gov/Archives/edgar/data/320193…
27  aapl-20200627_mdna.csv  https://www.sec.gov/Archives/edgar/data/320193…
28  aapl-20200627_mdna.csv  https://www.sec.gov/Archives/edgar/data/320193…
32  aapl-20200627_mdna.csv  https://www.sec.gov/Archives/edgar/data/320193…
..                     …                                                 …
50    d17062d10k_mdna.csv  https://www.sec.gov/Archives/edgar/data/320193…
51    d17062d10k_mdna.csv  https://www.sec.gov/Archives/edgar/data/320193…
52    d17062d10k_mdna.csv  https://www.sec.gov/Archives/edgar/data/320193…
53    d17062d10k_mdna.csv  https://www.sec.gov/Archives/edgar/data/320193…
54    d17062d10k_mdna.csv  https://www.sec.gov/Archives/edgar/data/320193…

      neg    neu    pos  compound      MACRO  MACRO_number  \
22  0.000  0.864  0.136    0.2960  ['china']           1.0
25  0.000  1.000  0.000    0.0000  ['china']           1.0
27  0.142  0.651  0.207    0.4019  ['china']           1.0
28  0.000  0.851  0.149    0.2732  ['china']           1.0
32  0.000  1.000  0.000    0.0000  ['china']           1.0
..    …    …    …       …        …           …
50  0.000  0.773  0.227    0.5719  ['china']           1.0
51  0.000  0.811  0.189    0.2732  ['china']           1.0
52  0.000  0.896  0.104    0.2732  ['china']           1.0
53  0.000  1.000  0.000    0.0000  ['china']           1.0
54  0.000  0.906  0.094    0.5994  ['china']           1.0

                trademarks  trademarks_number
22              ['store®']                1.0
```

```
25          ['pro®', 'pro ']                2.0
27          ['air®', 'pro®']                2.0
28  ['applecare®', 'store®']                2.0
32                 ['card ']                1.0
..                       …                …
50               ['beats®']                1.0
51                 ['pro®']                1.0
52           ['applecare®']                1.0
53               ['beats®']                1.0
54                 ['pro®']                1.0

[480 rows x 14 columns]
```

Follow the same process, but use the pattern library sentiment analyzer.

[10]:
```
list_trademarks = stn.get_words_with_trademark(stn.df_mdna)
df = stn.match_keywords(stn.pattern_sentiment(stn.
 ↪df_mdna),list_trademarks,'trademarks')
df = df.
 ↪dropna()[['href','sentence_text','pattern_sentiment','pattern_subjectivity','trademarks','t
df
```

[10]:
```
                                                 href  \
22  https://www.sec.gov/Archives/edgar/data/320193…
25  https://www.sec.gov/Archives/edgar/data/320193…
27  https://www.sec.gov/Archives/edgar/data/320193…
28  https://www.sec.gov/Archives/edgar/data/320193…
32  https://www.sec.gov/Archives/edgar/data/320193…
..                                                …
50  https://www.sec.gov/Archives/edgar/data/320193…
51  https://www.sec.gov/Archives/edgar/data/320193…
52  https://www.sec.gov/Archives/edgar/data/320193…
53  https://www.sec.gov/Archives/edgar/data/320193…
54  https://www.sec.gov/Archives/edgar/data/320193…

                                        sentence_text  pattern_sentiment  \
22  the covid-19 pandemic has significantly curtai…          -0.200000
25  the company is working on safely re-opening it…           0.000000
27  the most pronounced impact occurred in april 2…           0.102778
28  the full extent of the future impact of the co…          -0.062500
32  third quarter fiscal 2020 highlights total net…           0.000000
..                                                …                …
50  the year-over-year growth in mac net sales and…           0.200000
51  mac net sales and unit sales increased in all …           0.000000
52  mac asps decreased during 2014 compared to 201…          -0.103571
53                                       apple inc.           0.000000
54  | 2015 form 10-k | 25 table of contents servic…          -0.017857
```

```
    pattern_subjectivity                trademarks  trademarks_number
22              0.300000               ['store®']                1.0
25              0.000000         ['pro®', 'pro ']                2.0
27              0.369444         ['air®', 'pro®']                2.0
28              0.187500 ['applecare®', 'store®']                2.0
32              0.000000                ['card ']                1.0
..                   …                       …                …
50              0.250000               ['beats®']                1.0
51              0.000000                 ['pro®']                1.0
52              0.548810           ['applecare®']                1.0
53              0.000000               ['beats®']                1.0
54              0.175000                 ['pro®']                1.0

[480 rows x 6 columns]
```

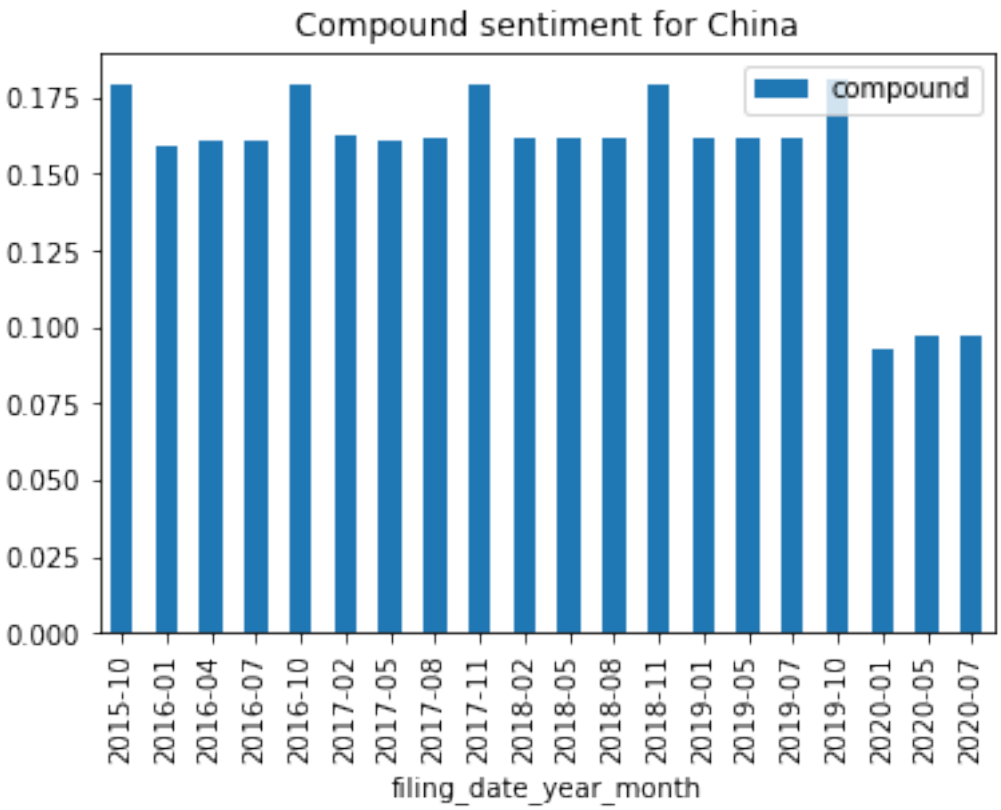
Lastly, read from a list of pre-extracted csv files. global_macro = ['covid', 'recession', 'global', 'virus', 'coronavirus', 'china', 'economy', 'gdp']

products_sentiment segments_sentiment trademarks_sentiment

using the SECTextNLP function read_from_csv, and pass in one of the above names, as below. Then you can filter by an item, and plot the sentiment. The following shows sentiment for AAPL around the word 'China'

[11]:
```
df = stn.read_from_csv(csv_file_name='global_macro')
df = df.loc[df['global_macro_long']=='china']
df = stn.year_month(df,'filing_date')
df = df.groupby('filing_date_year_month').mean().reset_index()
df.plot(kind='bar',x='filing_date_year_month',y='compound',title='Compound␣
↪sentiment for China')
```

[11]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x13e22e050>
```


Compound sentiment for China
compound
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
2015-10
2016-01
2016-04
2016-07
2016-10
2017-02
2017-05
2017-08
2017-11
2018-02
2018-05
2018-08
2018-11
2019-01
2019-05
2019-07
2019-10
2020-01
2020-05
2020-07
filing_date_year_month