SEC TEXT: NLP

A codebase to allow for search and Natural Language Processing 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.

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
from sec_text_nlp import *

#create an SECTextNLP object for ticker 'AAPL'
stn = SECTextNLP("AAPL")

#read the file index for a list of available documents
stn.df_file_index[['ticker','period_date','form_name','type']].head()
```

	ticker	period_date	form_name	type
139	AAPL	2020-06-27T04:00:00Z	Quarterly report [Sections 13 or 15(d)]	10-Q
140	AAPL	2020-03-28T04:00:00Z	Quarterly report [Sections 13 or 15(d)]	10-Q
141	AAPL	2019-12-28T05:00:00Z	Quarterly report [Sections 13 or 15(d)]	10-Q
142	AAPL	2019-09-28T04:00:00Z	Annual report [Section 13 and 15(d), not S-K I...	10-K
143	AAPL	2019-06-29T04:00:00Z	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 against othe

stn.df_text.head()
```

	part.name	item.name	sentence_text	file	
0	NaN	NaN	united states securities and exchange commissi...	aapl-20200627_sentences.csv	https://www.sec.gov/
1	NaN	NaN	20549 form 10-q (mark one) ☒ quarterly repor...	aapl-20200627_sentences.csv	https://www.sec.gov/
2	NaN	NaN	commission file number: 001-36743 apple inc.	aapl-20200627_sentences.csv	https://www.sec.gov/
3	NaN	NaN	(exact name of registrant as specified in its ...	aapl-20200627_sentences.csv	https://www.sec.gov/
4	NaN	NaN	employer identification no.)	aapl-20200627_sentences.csv	https://www.sec.gov/

```
#for more reference information, join with the file index on the href key.
#now you can see filing_date
pd.merge(stn.df_text,stn.df_file_index,how = 'inner',left_on='href',right_
```

	ticker	filing_date	sentence_text
0	AAPL	2020-07-31T04:00:00Z	united states securities and exchange commissi...
1	AAPL	2020-07-31T04:00:00Z	20549 form 10-q (mark one) ☒ quarterly repor...
2	AAPL	2020-07-31T04:00:00Z	commission file number: 001-36743 apple inc.
3	AAPL	2020-07-31T04:00:00Z	(exact name of registrant as specified in its ...
4	AAPL	2020-07-31T04:00:00Z	employer identification no.)

```
#OR, if you only want Management Discussion and Analysis,
#join with the df_mdna object instead of df_text
pd.merge(stn.df_mdna,stn.df_file_index,how = 'inner',left_on='href',right_
```

	ticker	filing_date	sentence_text
0	AAPL	2020-07-31T04:00:00Z	item 2.
1	AAPL	2020-07-31T04:00:00Z	management's discussion and analysis of financ...
2	AAPL	2020-07-31T04:00:00Z	forward-looking statements provide current exp...
3	AAPL	2020-07-31T04:00:00Z	for example, statements in this form 10-q rega...
4	AAPL	2020-07-31T04:00:00Z	forward-looking statements can also be identif...

```
#you can also pass in a topic, and get noun phrases around that topic.
#get business segments?
stn.get_noun_phrases_around_topic(BUSINESS_SEGMENT_LIST)
```

```
['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
stn.get_noun_phrases_around_topic(PRODUCTS_LIST,nrows=2)
```

```
['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
stn.get_words_with_trademark(stn.df_mdna)
```

```
['ipad®',
 'iphone®',
 'ipados®',
 'watch®',
 'macbook®',
 'store™',
 'arcade™',
 'retina®',
 'applecare®',
 'imac®',
 'tvos®',
 'pencil®',
 'pro™',
 'bar™',
 'touch®',
 'macos™',
 'pro®',
 'card™',
 'ipod®',
 'mini®',
 'watchos®',
 'music®',
 'air®',
 'beats®',
 'mac®',
 'homepod™',
 'folio™',
 'macos®',
 'airpods™',
 'x®',
 'store®',
 'pay®',
 'tvos™',
 'ipados™',
 'icloud®',
 'tv®']
```

```
#pair keyword search list with SENTIMENT ...
#using NLTK sentiment analyzer
#'MACRO' is the column name of the new dataframe
stn.match_keywords(stn.NLTK_sentiment(stn.df_mdna),GLOBAL_SEARCH_LIST,'MACR
```

	part.name	item.name	sentence_text	section	file	
0	PART I - FINANCIAL INFORMATION	Item 2. Management's Discussion and Analysis o...	item 2.	discussion	aapl-20200627_mdna.csv	https://v
1	PART I - FINANCIAL INFORMATION	Item 2. Management's Discussion and Analysis o...	management's discussion and analysis of financ...	discussion	aapl-20200627_mdna.csv	https://v
2	PART I - FINANCIAL INFORMATION	Item 2. Management's Discussion and Analysis o...	forward-looking statements provide current exp...	discussion	aapl-20200627_mdna.csv	https://v
3	PART I - FINANCIAL INFORMATION	Item 2. Management's Discussion and Analysis o...	for example, statements in this form 10-q rega...	discussion	aapl-20200627_mdna.csv	https://v
4	PART I - FINANCIAL INFORMATION	Item 2. Management's Discussion and Analysis o...	forward-looking statements can also be identif...	discussion	aapl-20200627_mdna.csv	https://v

```
#take the trademark list, and get the sentiment from the MDNA text
list_trademarks = stn.get_words_with_trademark(stn.df_mdna)
stn.match_keywords(stn.NLTK_sentiment(stn.df_mdna),list_trademarks,'tradema
```

	part.name	item.name	sentence_text	section	file	
0	PART I - FINANCIAL INFORMATION	Item 2. Management's Discussion and Analysis o...	item 2.	discussion	aapl-20200627_mdna.csv	https://v
1	PART I - FINANCIAL INFORMATION	Item 2. Management's Discussion and Analysis o...	management's discussion and analysis of financ...	discussion	aapl-20200627_mdna.csv	https://v
2	PART I - FINANCIAL INFORMATION	Item 2. Management's Discussion and Analysis o...	forward-looking statements provide current exp...	discussion	aapl-20200627_mdna.csv	https://v
3	PART I - FINANCIAL INFORMATION	Item 2. Management's Discussion and Analysis o...	for example, statements in this form 10-q rega...	discussion	aapl-20200627_mdna.csv	https://v
4	PART I - FINANCIAL INFORMATION	Item 2. Management's Discussion and Analysis o...	forward-looking statements can also be identif...	discussion	aapl-20200627_mdna.csv	https://v

```
#follow the same process, but use the pattern library sentiment analyzer ir
df = stn.match_keywords(stn.pattern_sentiment(stn.df_mdna),list_trademarks,
df.dropna()[['sentence_text','pattern_sentiment','pattern_subjectivity','tr
```

	sentence_text	pattern_sentiment	pattern_subjectivity	trademarks	trademar
22	the covid-19 pandemic has significantly curtai...	-0.200000	0.300000	['store®']	1.0
25	the company is working on safely re-opening it...	0.000000	0.000000	['pro™', 'pro®']	2.0
27	the most pronounced impact occurred in april 2...	0.102778	0.369444	['pro®', 'air®']	2.0
28	the full extent of the future impact of the co...	-0.062500	0.187500	['applecare®', 'store®']	2.0
32	third quarter fiscal 2020 highlights total net...	0.000000	0.000000	['card™']	1.0