```
from google.colab import drive
drive.mount('/content/drive')
```

Data Preprocessing

The cells below preprocess the data.

Prerequisites for the data processing to work.

1) Create a folder in you google drive final-project/raw/

2) Download the zip fles from the following URL https://www.sec.gov/dera/data/financial-statement-data-sets.html

For the purpose of this project I have used the last 4 quarters.

Step 1. Extract the raw data files.

```
import zipfile

with zipfile.ZipFile('/content/drive/My Drive/final-project/raw/2018q3_notes.zip', 'r') as
    zip_ref.extractall('/content/drive/My Drive/final-project/raw/extracted/2018q3')
with zipfile.ZipFile('/content/drive/My Drive/final-project/raw/2018q4_notes.zip', 'r') as
    zip_ref.extractall('/content/drive/My Drive/final-project/raw/extracted/2018q4')
with zipfile.ZipFile('/content/drive/My Drive/final-project/raw/2019q1_notes.zip', 'r') as
    zip_ref.extractall('/content/drive/My Drive/final-project/raw/extracted/2019q1')
with zipfile.ZipFile('/content/drive/My Drive/final-project/raw/2019q2_notes.zip', 'r') as
    zip_ref.extractall('/content/drive/My Drive/final-project/raw/extracted/2019q2')
```

```
with open("/content/drive/My Drive/final-project/raw/extracted/2018q3/txt.tsv") as myfile:
    head = [next(myfile) for x in range(10)]
print(head)
```

```
['adsh\ttag\tversion\tddate\tqtrs\tiprx\tlang\tdcml\tdurp\tdatp\tdimh\tdimn\tcore
```

```
import pandas as pd

#folders = ['2018q3/','2018q4/','2019q1/','2019q2/']
folders = ['2018q3/']
extract_folder = '/content/drive/My Drive/final-project/raw/extracted/'
context_file = 'txt.tsv'

filenames = []
for folder in folders:
  filename = extract_folder + folder + context_file
  filenames.append(filename)
print(filenames)

dfs = pd.concat([pd.read_csv(f, sep='\t') for f in filenames], ignore_index = True)

print(dfs.columns.values)
```

```
['/content/drive/My Drive/final-project/raw/extracted/2018q3/txt.tsv']
['adsh' 'tag' 'version' 'ddate' 'qtrs' 'iprx' 'lang' 'dcml' 'durp' 'datp'
 'dimh' 'dimn' 'coreg' 'escaped' 'srclen' 'txtlen' 'footnote' 'footlen'
 'context' 'value']
```

Step 2 gather some statistics on the data to further cleanse and reduce nois

```
##print(dfs.loc[: , "value"])
#Fetch wordcount for each content value
dfs['word_count'] = dfs['value'].apply(lambda x: len(str(x).split(" ")))
dfs[['value','word_count']].head()
```

	value	word_count
0	BORQS Beijing was qualified for a High and New...	62
1	Yuantel Telecom was qualified for a High and N...	102
2	(I) Impairment of long-lived assets: The Compa...	240
3	For the years ended December 31, 2015 2016 201...	42
4	Three Months Ended March 31, 2017 2018 US$ US$...	34

```
##Descriptive statistics of word counts
dfs.word_count.describe()
```

```
count    455141.000000
mean         96.609286
std         120.663369
min           1.000000
25%           1.000000
50%          40.000000
75%         167.000000
max         666.000000
Name: word_count, dtype: float64
```

```
#Identify common words
import pandas
#Identify common words
freq = pandas.Series(''.join(map(str,dfs['value'])).split()).value_counts()[:20]
freq
```

```
the        1874070
of         1466463
$          1254879
and        1206969
to          698295
in          591267
)           498098
for         440268
2018        361920
a           356613
June        350656
30,         345073
The         343280
as          295407
on          271064
2017        269819
are         249429
Company     245354
is          239585
or          208053
dtype: int64
```

```
#Identify uncommon words
freq1 =  pandas.Series(''.join(map(str,dfs['value'])).split()).value_counts()[:-20]
freq1
```

Pre Processing the text.

Now that we have the basic stats lets do some pre processing to remove noise and normalize the data.Data compo redundant to the core text analytics can be considered as noise.

```
import nltk
import re
nltk.download('wordnet')
from nltk.stem.porter import PorterStemmer
from nltk.stem.wordnet import WordNetLemmatizer
nltk.download('stopwords')
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from nltk.tokenize import RegexpTokenizer
```

```
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Unzipping corpora/wordnet.zip.
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

```
##Creating a list of stop words and adding custom stopwords
stop_words = set(stopwords.words("english"))
##Creating a list of custom stopwords
new_words = ['Jan', 'Janurary', 'Feb', 'February', 'March', 'April', 'May', 'Jun', 'June',
            'Aug', 'August', 'Sept', 'September', 'Oct', 'October', 'Nov', 'November', 'Dec
             'Month', 'Ended', 'Ending', 'Three', 'Period']
stop_words = stop_words.union(new_words)
```

```
#Build the corpus and save it.
corpus = []
for i in range(0, 5000):
    #Remove punctuations
    text = re.sub('[^a-zA-Z]', ' ', str(dfs['value'][i]))

    #Convert to lowercase
    text = text.lower()

    #remove tags
    text=re.sub("&lt;/?.*?&gt;"," &lt;&gt; ",text)

    # remove special characters and digits
    text=re.sub("(\\d|\\W)+"," ",text)

    ##Convert to list from string
    text = text.split()

    ##Stemming
    ps=PorterStemmer()
    #Lemmatisation
    lem = WordNetLemmatizer()
    text = [lem.lemmatize(word) for word in text if not word in
            stop_words]
    text = [ word for word in text if len(word) > 3 ]
    if len(text) > 0:
      text = " ".join(text)
      corpus.append(text)


#View corpus item
corpus[:100]
```

```
['borqs beijing qualified high technology enterprise hnte since eligible preferen
 'yuantel telecom qualified high technology enterprise hnte since eligible prefer
 'impairment long lived asset company periodically review estimated useful life o
 'year ended december cost revenue sale marketing expense general administrative
 'three month ended march cost revenue sale marketing expense general administrat
 'expire',
 'expire',
 'expire',
 'expire',
 'stock plan march company adopted equity incentive plan plan plan replaces compa
 'note investment hong kong joint venture company hold interest joint venture hon
 'prepaid expense current asset prepaid expense current asset consist principally
 'obsolescence allowance inventory follows year ended march balance beginning yea
 'segment information company reportable segment plastic injection molding electr
 'allowance doubtful account company regularly monitor risk collecting amount owe
 'march finished good computer component total inventory',
 'unaudited selected quarterly result operation follows thousand except share amo
 'merger agreement provides zebra commence offer promptly practicable event later
 'closing offer subject following condition merger agreement terminated accordanc
 'smaller reporting company',
 'restatement previously reported condensed consolidated quarterly financial stat
 'june ownership structure expressed percentage general limited partnership inter
 'analysis significant assumption used income approach projected cash flow follow
 'commitment contingency time time company party ordinary routine litigation inci
 'cost capitalized software related accumulated amortization follows million capi
 'receivables receivables allowance doubtful account represent estimated realizab
 'following table summarizes final allocation purchase price million preliminary
 'core laboratory',
 'note stock option december board director company adopted stock option stock bo
 'basis presentation accompanying unaudited financial statement prepared accordan
 'earnings loss share basic earnings loss share represents income loss period sha
 'following table present financial asset liability company record fair value rec
 'clcs',
 'weighted weighted average average remaining number exercise life intrinsic warr
 'following table provides summary change fair value including transfer level lia
 'debt june december senior note principal amount senior note principal amount se
 'goodwill goodwill excess acquisition price fair value tangible identifiable int
 'month ended june loss attributable community health system stockholder transfer
 'stock based compensation company estimate fair value share based payment date g
 'deferred asset deferred expense property equipment capitalized start expense de
 'earnings common share income common share determined follows three month ended
 'amount reclassified accumulated comprehensive income loss recognized consolidat
 'table forth accretable yield activity firm consumer loan three month ended june
 'expense firm consolidated statement income included following three month ended
 'following table show impact single notch notch downgrade long term issuer ratin
 'false',
 'earnings common share basic earnings share amount earnings adjusted dividend de
 'following table summarize gain loss derivative designated hedging instrument th
 'mortgage servicing asset originate periodically sell commercial residential mor
 'false',
 'long term obligation credit agreement january company entered credit agreement
 'earnings share million except share data three month ended june month ended jun
 'following table provides amount cash cash equivalent presented condensed consol
 'stockholder equity change accumulated comprehensive income loss table present c
 'fair value asset liability june december follows thousand june december fair va
 'three month ended june nine month ended june operating income loss access equip
 'business segment information company organized four reportable segment based in
 'three month ended june month ended june combined condensed statement operation
 'total expended property plant equipment follows thousand nine month ended june
 'unit exchangeable share common stock',
 'following table summarize unrealized gain loss related short term investment de
 'related party transaction included amount related party affiliate april project
 'property held sale company classifies property held sale management commits sel
 'following forth tnmps recent interim transmission cost rate increase effective
 'june parent guarantor guarantor elims consolidated current asset cash cash equi
 'supplemental guarantor condensed consolidating financial statement rule regulat
 'summary change carrying value goodwill company reportable segment presented mil
 'following table present component periodic benefit cost plan pension benefit th
 'segment company five reportable segment comprised four individual operating sub
 'following table discussion present detail prior year claim claim adjustment exp
 'component temporary impairment otti loss recognized earnings asset type follows
 'inventory consisted following june december thousand material work progress fin
 'following table provides information revenue differentiated based product categ
 'performance percentage ranging greater provide awarding share ranging share ori
 'weighted average yield based amortized cost june dollar thousand weighted amort
 'following table present security sold agreement repurchase related weighted ave
 'following table present recorded investment residential consumer loan based pay
 'accrued liability consist following thousand june december accrued clinical mar
 'following table summarizes activity market condition award plan related informa
 'restaurant portfolio optimization fiscal company initiated plan review restaura
 'basis presentation consolidation accompanying unaudited condensed consolidated
 'lease agreement office space hong kong duration year',
 'lease agreement office space hangzhou china duration year',
 'following table present related party transaction avon affiliate cerberus insti
 'segment information determine segment profit deducting related cost expense seg
 'unfunded letter credit lending commitment june december presented thousand june
 'following table present asset measured fair value recurring basis date indicate
 'note subsequent event pending acquisition financial holding july synovus entere
 'following schedule present additional information regarding impaired loan class
 'following schedule summarizes information debt security available sale held mat
 'summary corporation loan follows dollar thousand originated acquired total loan
 'following table present assumption utilized determining fair value loan servici
 'summary reserve representation warranty corporation follows three month ended j
 'large accelerated filer',
 'long term debt long term debt consisted following july december series class no
 'trust service investment management trust service investment management include
 'following table present unrealized gain loss period relates equity security sti
 'following table provides information balance sheet classification accrual milli
 'following table provides component inventory million dollar july december finis
 'total gain loss security reported consolidated statement income comprehensive i
```

Write corpus to a file.

```
## Write data cleansed to the file.
with open('/content/drive/My Drive/final-project/corpus/corpus.txt', 'w+') as f:
    for line in corpus:
        f.write("%s\n" % line)
```