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

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.m
```

```
!pip install textstat
!pip install PyPDF2
!pip install -U textblob
import os
import glob
import textstat
import csv
import PyPDF2
from textblob import TextBlob
```

```
Requirement already satisfied: textstat in /usr/local/lib/python3.7/dist-packages (0
Requirement already satisfied: pyphen in /usr/local/lib/python3.7/dist-packages (from
Requirement already satisfied: PyPDF2 in /usr/local/lib/python3.7/dist-packages (1.26
Requirement already up-to-date: textblob in /usr/local/lib/python3.7/dist-packages (0
Requirement already satisfied, skipping upgrade: nltk>=3.1 in /usr/local/lib/python3
Requirement already satisfied, skipping upgrade: six in /usr/local/lib/python3.7/dist
```

```
TARGET_FILES = r'/content/drive/My Drive/Control sample/*.pdf'
file_list = glob.glob(TARGET_FILES)
```

```
new_list=[]
name=[]
size=[]
pagen=[]
for file in file_list:
  print(file)
  n=(file.split('.pdf',0))
  name.append(n)
  size.append((os.stat(file).st_size)/1000000)


  if file.endswith(".pdf"):
    with open (file,'rb') as pdfobject:
      pdfreader=PyPDF2.PdfFileReader(pdfobject)
      page_numbers=pdfreader.numPages
      pagen.append(page_numbers)
      if pdfreader.isEncrypted == True:
        pass
      else:
          currentpage = 0
          text = ""
          while (currentpage <page_numbers):
            page=pdfreader.getPage(currentpage)
            try:
              text=text+page.extractText()
            except:
                print(currentpage) #to print the error page
            currentpage +=1
          new_list.append(text)


```

```
gf=[]
kg=[]
ke=[]
dc=[]
sc=[]
dif=[]
st=[]
ct=[]
sep=[]
ses=[]
for test in new_list:
  #print(textstat.gunning_fog(test))
  #print(textstat.lexicon_count(test, removepunct=True))
  gf.append(textstat.gunning_fog(test))
  kg.append(textstat.flesch_kincaid_grade(test))
  ke.append(textstat.flesch_reading_ease(test))
  dc.append(textstat.dale_chall_readability_score(test))
  sc.append(textstat.sentence_count(test))
  dif.append(textstat.difficult_words(test))
  st.append(textstat.text_standard(test))
  ct.append(textstat.lexicon_count(test, removepunct=True))
  sep.append(TextBlob(test).sentiment.polarity)
  ses.append(TextBlob(test).sentiment.subjectivity)
```

```
ticker=[]
year=[]
for file in file_list:
  ticker.append((file.rsplit('/content/drive/My Drive/Control sample/')[1][0:3]))
  year.append((file.rsplit('/content/drive/My Drive/Control sample/')[1][4:8]))
```

```
from itertools import zip_longest
data = [ticker, year, gf, kg,ke,dc,sc,dif,st,name,ct, pagen, size,sep,ses]
export_data = zip_longest(*data, fillvalue = '')
with open('/content/drive/My Drive/MSFT/READABILITY_CONTROLS.csv', 'w', encoding="ISO-8
      write = csv.writer(file)
      write.writerow(("TICKER", "YEAR", "GUNNING FOG", "FLESCH-KINCAID GRADE","FLESCH-R
      write.writerows(export_data)
```