AF3214

Week 12. Textual Analysis and Readability

Learn how to process PDF files using Python

Read https://www.geeksforgeeks.org/working-with-pdf-files-in-python/

Obtain PDF files of academic papers

Manually download the following 4 academic papers written by the author of this Jupyter Notebook and save these 4 PDF files in your local computer:
(1) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3694637
(2) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3410538
(3) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3209449
(4) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2625975

Obtain number of pages from PDF files

In [ ]:
```
!pip install PyPDF2==2.12.1
```

In [ ]:
```
from PyPDF2 import PdfReader
```

In [ ]:
```
pdf1 = PdfReader("SSRN-id3694637.pdf")
pdf1_pages = len(pdf1.pages)
print("Number of Pages: " + str(pdf1_pages))
```

In [ ]:
```
pdf2 = PdfReader("SSRN-id3410538.pdf")
pdf2_pages = len(pdf2.pages)
print("Number of Pages: " + str(pdf2_pages))
```

In [ ]:
```
pdf3 = PdfReader("SSRN-id3209449.pdf")
pdf3_pages = len(pdf3.pages)
print("Number of Pages: " + str(pdf3_pages))
```

In [ ]:
```
pdf4 = PdfReader("SSRN-id2625975.pdf")
pdf4_pages = len(pdf4.pages)
print("Number of Pages: " + str(pdf4_pages))
```

Obtain text from PDF files

In [ ]:
```
# if we want to do the readbility for more than 1 page, use the loop functic
pdf1_page1 = pdf1.pages[0]
pdf1_page1_text = pdf1_page1.extract_text()
print("Text on Page 1: " + pdf1_page1_text)
```

In [ ]:
```
pdf2_page1 = pdf2.pages[0]
pdf2_page1_text = pdf2_page1.extract_text()
print("Text on Page 1: " + pdf2_page1_text)
```

In [ ]:
```
pdf3_page1 = pdf3.pages[0]
pdf3_page1_text = pdf3_page1.extract_text()
print("Text on Page 1: " + pdf3_page1_text)
```

In [ ]:
```
pdf4_page1 = pdf4.pages[0]
pdf4_page1_text = pdf4_page1.extract_text()
print("Text on Page 1: " + pdf4_page1_text)
```

Compute readability

Read https://pypi.org/project/textstat/ and
https://en.wikipedia.org/wiki/Gunning_fog_index

In [ ]:
```
!pip install textstat
```

In [ ]:
```
import textstat
```

In [ ]:
```
readability = textstat.gunning_fog(pdf1_page1_text)
readability
```

In [ ]:
```
readability = textstat.gunning_fog(pdf2_page1_text)
readability
```

In [ ]:
```
readability = textstat.gunning_fog(pdf3_page1_text)
readability
```

In [ ]:
```
readability = textstat.gunning_fog(pdf4_page1_text)
readability
```

Try another Python package to process PDF files

Read https://pdfminersix.readthedocs.io/en/latest/index.html

Converting a PDF file to text

Read
https://pdfminersix.readthedocs.io/en/latest/topic/converting_pdf_to_text.html

In [ ]:
```
!pip install pdfminer.six
```

extract_text

The most simple way to extract text from a PDF is to use extract_text:

In [ ]:
```
from pdfminer.high_level import extract_text
```

In [ ]:
```
pdf1 = extract_text("SSRN-id2625975.pdf")
pdf1
```

In [ ]:
```
readability = textstat.gunning_fog(pdf1)
readability
```

Additional Coding: use loops to obtain all pages in PDF files

Readability of Text

Annual report

Please download the most recent annual report in the PDF format filed by Apple at https://investor.apple.com/sec-filings/default.aspx, show the total number of pages, compute the readability of all pages of this PDF file, and show the readability result.

In [ ]:
```
from PyPDF2 import PdfReader

apple = PdfReader("aapl2024.pdf")
number_of_pages = len(apple.pages)

# We use loops to obtain all pages and put them in a list
apple_allpages = []
for i in range(number_of_pages):
    page = apple.pages[i]
    apple_allpages.append(page.extractText())
print(apple_allpages)

# Using join() method to concatenate items in a list to a single string
apple_allpages2 = " ".join(apple_allpages)
apple_allpages2
```

In [ ]:
```
# Compute readability for all pages

import textstat

readability = textstat.gunning_fog(apple_allpages2)
readability
```