



المحتويات

| | | | | | | | | | |
|-----------------|--------------------------|-----------------|-----------------------|-------------|------------------------|------------|----------------|-------------------|----------------------|
| | | | | التطبيقات | العقبات و التحديات | تاريخ NLP | ما هو NLP | المحتويات | 1) مقدمة |
| | | | | | البحث في النصوص | ملفات pdf | الملفات النصية | المكتبات | 2) أساسيات NLP |
| T.Visualization | Syntactic Struc. | Matchers | Stopwords | NER | Stem & Lemm | POS | Sent. Segm. | Tokenization | 3) أدوات NLP |
| | Dist. Similarity | Text Similarity | TF-IDF | BOW | Word2Vec | T. Vectors | Word embed | Word Meaning | 4) المعالجة البسيطة |
| T. Generation | NGrams | Lexicons | GloVe | L. Modeling | NMF | LDA | T. Clustering | T. Classification | 5) المعالجة المتقدمة |
| | Summarization & Snippets | | Ans. Questions | | Auto Correct | Vader | Naïve Bayes | Sent. Analysis | |
| Search Engine | Relative Extraction | | Information Retrieval | | Information Extraction | | Data Scraping | Tweet Collecting | 6) تجميع البيانات |
| | | | | | Rec NN\TNN | GRU | LSTM | Seq to Seq | 7) RNN |
| Chat Bot | Gensim | FastText | Bert | Transformer | Attention Model | T. Forcing | CNN | Word Cloud | 8) تكنيكات حديثة |

القسم الثاني : أساسيات NLP

الجزء الثالث : التعامل مع ملفات pdf

لقراءة ملفات الـ pdf عليك أولا ان تقوم بتحميل مكتبة PyPDF2 عبر كتابة هذه الامر في النوتبوك

```
pip install PyPDF2
```

بعدها نقوم باستدعاء المكتبة

```
import PyPDF2
```

و هناك ثلاث دوال هامة في المكتبة , للقراءة , وللمرج , وللكتابة

```
PdfFileMerger, PdfFileReader, PdfFileWriter
```

و يتم قراءة الملف عبر كتابة هذين السطرين

```
f = open('US_Declaration.pdf','rb')  
pdf_reader = PyPDF2.PdfFileReader(f)
```

و حرفي rb يدل على read binary

و لمعرفة عدد الصفحات

```
pdf_reader.numPages
```

و لقراءة صفحة معينة يتم استدعائها أولاً ثم عرضها

```
page_one = pdf_reader.getPage(0)  
page_one_text = page_one.extractText()  
print(page_one_text)
```


و بالنسبة لكتابة ملف pdf فالطريقة الحالية هي قراءة صفحات من ملف معين و اعادة كتابتها في ملف اخر , وبالتالي يمكن عمل ملف فيه تجميعات من صفحات كثيرة

و يبدأ الامر بقراءة صفحة من ملف معين

```
f = open('US_Declaration.pdf','rb')  
pdf_reader = PyPDF2.PdfFileReader(f)
```

```
first_page = pdf_reader.getPage(0)
```

ثم عمل الكائن الخاص بكتابة الملف

```
pdf_writer = PyPDF2.PdfFileWriter()
```

ثم اضافة الصفحة المقروءة فيه

```
pdf_writer.addPage(first_page)
```

ثم عمل الملف الذي سيكتب فيه


```
merger = PdfFileMerger()
```

```
merger.append(PdfFileReader(open(filename1, 'rb')))
```

```
merger.append(PdfFileReader(open(filename2, 'rb')))
```

```
merger.write("merged file 1.pdf")
```

او يمكن اختيار صفحات محددة هكذا :

```
file1 = PdfFileReader(open(filename1, "rb"))
```

```
file2 = PdfFileReader(open(filename2, "rb"))
```

```
output = PdfFileWriter()
```

```
output.addPage(file1.getPage(5))
```

```
output.addPage(file2.getPage(3))
```

```
outputStream = open("merged file 2.pdf", "wb")
```

```
output.write(outputStream)
```

```
outputStream.close()
```