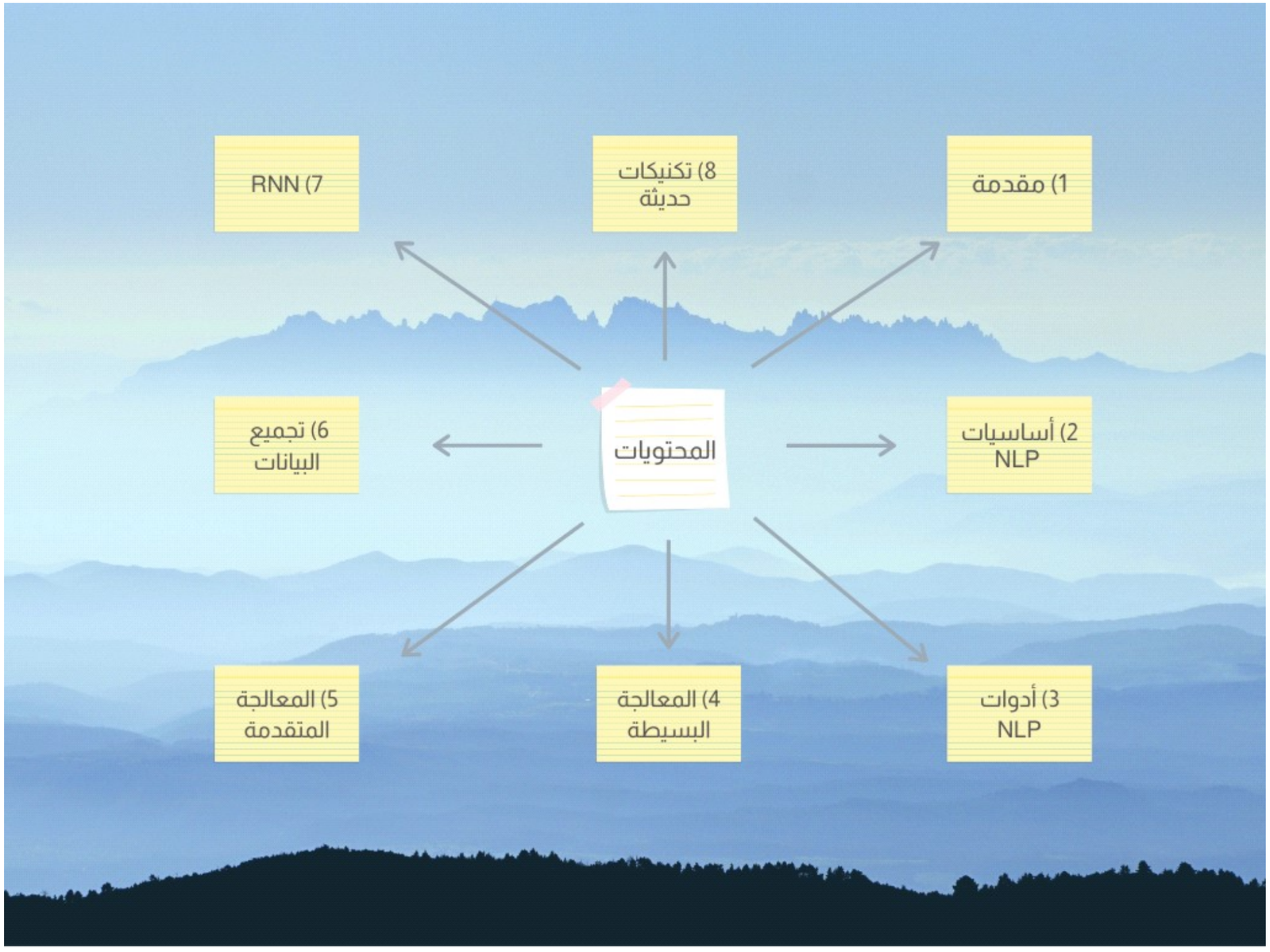
Natural Language Processing المعالجة اللغوية الطبيعية



المحتويات

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

القسم الثالث : أدوات NLP

الجزء السادس : Stopwords

==========

وهي الكلمات التي تنتشر كثيرا في الكتابة ولا نريد لها ان يتم عمل تاج لها لانتشارها و قلة تأثيرها في المعني
و هي غالبا ما يتم استخدامها بشكل متكرر و دائم , و لا يكون لها تأثير كبير في المعني فيمكن ان يتم حذفها , مع مراعاة ان بعض الكلمات مهمة لها فرق في المعني مثل الفرق بين and , or و كلمة not

و مكتبة spacy تخزن عددا من الكلمات فيها بالفعل يمكن ان نعرضها هكذا :

```
import spacy
nlp = spacy.load('en_core_web_sm')

print(nlp.Defaults.stop_words)
```

{'alone', "‘ve", 'seem', 'mine', 'he', 'move', 'each', 'among', 'elsewhere', 'might', 'everywhere', 'if', 'here', 'see', 'nine', 'whither', 'last', 'whose', 'about', 'almost', 'fifteen', 'of', 'further', 'latterly', 'herein', 'another', 'ca', 'hereupon', 'several', 'have', 'anywhere', 'same', 'anyone', 'down', 'how', 'or', 'those', 'whoever', 'any', 'part', 'from', 'for', 'say', 'been', 'than', 'who', 'beyond', "n't", 'there', 'well', 'regarding', 'other', 'such', 'though', 'become', "’ve", 'otherwise', 'cannot', 'must', 'show', 'thru', 'once', 'herself', 'put', 'does', 'itself', "‘d", 'never', 'keep', 'anyway', 'every', 'while', 'are', 'afterwards', 'although', 'the', 'us', 'onto', 'along', 'whence', 'him', 'am', 'some', 'get', 'beside', 'however', 'throughout', 'only', 'all', 'amongst', 'they', 'to', 'using', 'many', 'empty', 'yours', 'were', 'due', 'had', "‘s", 'around', 'per', 'front', 'bottom', 'me', 'them', 'himself', 'anyhow', 'whereas', 'n‘t', "'re", 'nor', 'eleven', "'ll", 'even', 'more', 'towards', 'was', 'twenty', 'up', 'during', 'one', 'as', 'did', 'its', "'s", 'meanwhile', 'go', "'d", 'moreover', 'amount', 'sometime', 'nobody', 'could', 'own', 'became', 'hereby', 'twelve', 'again', 'when', 'and', 'toward', 'whereafter', 'three', 'thence', 'therein', 'call', 'so', 'someone', 'something', 'nothing', 'n’t', 'with', 'various', 'out', 'hers', 'hundred', 'yourself', 'what', 'sixty', 'unless', 'therefore', 'through', 'indeed', 'this', 'whom', 'everything', 'without', 'do', 'on', 'she', 'too', 'two', 'none', 'is', 'under', 'wherein', 'rather', "‘ll", 'i', 'these', 'whatever', 'really', 'over', 'thereupon', 'thus', 'most', 'perhaps', 'by', 'few', 'no', 'together', 'four', 'yourselves', "’re", 'a', 'why', 'very', 'which', 'others', 'used', 'whether', 'can', 'his', 'already', 'becomes', 'former', 'much', 'made', 'first', 'formerly', 'upon', 'above', 'but', 'whenever', 'via', 'we', 'least', "'ve", 'now', 'also', 'becoming', 'whereby', 'my', 'thereby', 'should', 'anything', 'an', 'nevertheless', 'else', 'next', 'eight', 'either', 'latter', 'just', 'yet', 'ours', 'her', 'myself', 'at', 'noone', 'that', 'give', 'back', "‘m", 'below', 'being', 'full', 'off', 'across', 'besides', 'in', 'take', 'whereupon', 'make', 'seeming', 'against', 'everyone', 'quite', 'your', 'somewhere', 'into', 'third', 'would', 'hence', 're', 'where', 'always', 'serious', "’d", 'since', "‘re", "’m", 'done', 'nowhere', 'less', 'our', 'between', 'five', "’ll", 'then', 'please', "’s", 'before', 'it', 'whole', "'m", 'fifty', 'behind', 'may', 'be', 'mostly', 'often', 'seemed', 'sometimes', 'after', 'except', 'forty', 'not', 'name', 'within', 'you', 'doing', 'enough', 'still', 'has', 'thereafter', 'six', 'themselves', 'wherever', 'ourselves', 'beforehand', 'hereafter', 'ten', 'because', 'seems', 'their', 'ever', 'both', 'until', 'namely', 'neither', 'side', 'top', 'somehow', 'will'}

و يمكن معرفة اذا كانت كلمة معينة من ضمنهم ام لا من هنا :

```
nlp.vocab['myself'].is_stop
nlp.vocab['mystery'].is_stop
```

_

و يمكن اضافة كلمة معينة الي هذه الكلمات , فاولا يتم فحص حالة كلمة btw

```
nlp.vocab['btw'].is_stop
```

ثم نقوم باضافتها هكذا :

```
nlp.Defaults.stop_words.add('btw')
nlp.vocab['btw'].is_stop = True
```

ثم فحصها مرة اخري

```
nlp.vocab['btw'].is_stop
```

كذلك يمكن حذف كلمة منها , بهذه الطريقة

```
nlp.vocab['beyond'].is_stop

nlp.Defaults.stop_words.remove('beyond')
nlp.vocab['beyond'].is_stop = False

nlp.vocab['beyond'].is_stop
```

لكن لاحظ ان الاضافة او الحذف تكون فقط في هذا الكرينيل , فاذا ما تم عمل ريستارت له فلن تكون موجودة و لن يتم حفظها في المكتبة نفسها

_

و في مكتبة nltk هناك ايضا عدد كبير من كلمات التوقف

```
import nltk
nltk.download('stopwords')

from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

stop_words = set(stopwords.words('english'))
print(len(stop_words))
stop_words
```

و هنا يمكن عرض ال tokens بعد حذف كلمات التوقف

```
example_sent = "This is a sample sentence, And This showing off the stop words
filtration."
stop_words = set(stopwords.words('english'))

word_tokens = word_tokenize(example_sent)
word_tokens
```

```
filtered_sentence = [w for w in word_tokens if not w in stop_words]

print(word_tokens)
print('-------------------------------')
print(filtered_sentence)
```

و يمكن ضبط أمر ال small

```
filtered_sentence = [w for w in word_tokens if not w.lower() in stop_words]

print(word_tokens)
print('-------------------------------')
print(filtered_sentence)
```

_

و تدعم مكتبة nltk اللغة العربية بأسلوبين

أولا بعدد 243 كلمة هنا :

```
Ar_SW1 = set(nltk.corpus.stopwords.words("arabic"))
print(len(Ar_SW1))
Ar_SW1
```

'| عليك | اللذين | لعل | بكن | ذلكم | لا | حيثما | لكما | ها | ته | هناك | ليست | هل | تلكم | تلكما | بكما | كأين | له | بها | كيت | وإذ | أف | ذلكما | فيم | لست | ومن | أيها | دون | ذواتي | الذين | لكي | أنتم | أنا | لهن | لولا | هي | إلى | ذه | فمن | عدا | كي | عليه | إنا | ذات | إلا | ذلك | حين | عن | لي | نحو | آه | منذ | ألا | اللتيا | بيد | كم | كيفما | فلا | مما | إذ | هيا | أما | بعض | إن | كلما | على | ولكن | بل | كلاهما | لئن | لك | ذلكن | لهم | اللائي | ذواتا | سوى | فإن | هاتان | هذا | هذه | أنى | إذما | أو | به | غير | هو | بمن | نعم | عسى | أوه | إنه | أين | إيه | اللواتي | لاسيما | كيف | اللتان | ممن | منه | هاته | بي | حبذا | بهن | ريث | لستما | بعد | هن | ذي | لن | كليكما | لم | من | كأنما | الذي | كذا | هم | هاتين | هيهات | ماذا | لهما | لها | هاهنا | هذي | التي | أنتن | آها | إذن | هاتي | هيت | هنا | لوما | أكثر | في | ثم | أينما | أنت | تلك | بين | حيث | بلى | أم | بهم | أولئك | سوف | ذوا | مهما | ثمة | اللاتي | عما | ولو | لستن | كلتا | ذان | لسن | هنالك | أقل | فيه | هؤلاء | فإذا | لكم | وما | أن | إنما | فيما | ليستا | أولاء | لدى | مه | ذو | هذين | لما | مذ | ولا | لكيلا | ليسا | أنتما | ذاك | فيها | نحن | والذين | ليسوا | تينك | عند | وإذا | بهما | إليكما | هلا | ذانك | إما | كأن | لكنما | بخ | منها | ذا | إذا | كأي | لستم | حتى | عل | وهو | اللذان | والذي | ذين | تين | بس | آي | بكم | متى | إليك | هما | بك | بماذا | هذان | يا | بما | اللتين | إليكن | كلا | لنا | حاشا | لكن | مع | هاك | ليت | بنا | تي | خلا | قد | ما | ذينك | كل | أي | وإن | كما | لو | إليكم | إي | كليهما | كذلك | ليس | شتان | هكذا | لسنا |'

و بعدد 248 كلمة هنا :

```
from nltk.corpus import stopwords
Ar_SW2 = stopwords.words('arabic')
print(len(Ar_SW2))
' | '.join(Ar_SW2)
```

' إذ | إذا | إذما | إذن | أف | أقل | أكثر | ألا | إلا | التي | الذي | الذين | اللاتي | اللائي | اللتان | اللتيا | اللتين | اللذان | اللذين | اللواتي | إلى | إليك | إليكم | إليكما | إليكن | أم | أما | أما | إما | أن | إن | إنا | أنا | أنت | أنتم | أنتما | أنتن | إنما | إنه | أنى | أنى | آه | آها | أو | أولاء | أولئك | أوه | آي | أي | أيها | إي | أين | أينما | إيه | بخ | بس | بعد | بعض | بك | بكم | بكم | بكما | بكن | بل | بلى | بما | بماذا | بمن | بنا | به | بها | بهم | بهما | بهن | بي | بين | بيد | تلك | تلكم | تلكما | ته | تي | تين | تينك | ثم | ثمة | حاشا | حبذا | حتى | حيث | حيثما | حين | خلا | دون | ذا | ذات | ذاك | ذان | ذانك | ذلك | ذلكم | ذلكما | ذلكن | ذه | ذو | ذوا | ذواتا | ذواتي | ذي | ذين | ذينك | ريث | سوف | سوى | شتان | عدا | عسى | عل | على | عليك | عليه | عما | عن | عند | غير | فإذا | فإن | فلا | فمن | في | فيم | فيما | فيه | فيها | قد | كأن | كأنما | كأي | كأين | كذا | كذلك | كل | كلا | كلاهما | كلتا | كلما | كليكما | كليهما | كم | كم | كما | كي | كيت | كيف | كيفما | لا | لاسيما | لدى | لست | لستم | لستما | لستن | لسن | لسنا | لعل | لك | لكم | لكما | لكن | لكنما | لكي | لكيلا | لم | لما | لن | لنا | له | لها | لهم | لهما | لهن | لو | لولا | لوما | لي | لئن | ليت | ليس | ليسا | ليست | ليستا | ليسوا | ما | ماذا | متى | مذ | مع | مما | ممن | من | منه | منها | منذ | مه | مهما | نحن | نحو | نعم | ها | هاتان | هاته | هاتي | هاتين | هاك | هاهنا | هذا | هذان | هذه | هذي | هذين | هكذا | هل | هلا | هم | هما | هن | هنا | هناك | هنالك | هو | هؤلاء | هي | هيا | هيت | هيهات | والذي | والذين | وإذ | وإذا | وإن | ولا | ولكن | ولو | وما | ومن | وهو | يا '

كما ان هناك مكتبة خاصة بكلمات التوقف للغة العربية , ويتم تحميلها من هنا :

```
pip install arabicstopwords
```

و هي فيها نوعين من كلمات التوقف , جذور الكلمات , و 507 كلمة , تظهر هنا

***

```
len(stp.classed_stopwords_list())
stp.classed_stopwords_list()
```

```
['آنفا', 'ل', 'عسى', 'هب', 'هج', 'الازلتم', 'غرب', 'ش', 'ها', 'لدن', 'إيه', 'ؤ', 'أجمع', 'عوض', 'بنا', 'زمان', 'لازلنا', 'حي', 'ذواتا', 'جنوب', 'حم',
'الازلتن', 'هن', 'هم', 'هل', 'هاؤم', 'ارتد', 'هي', 'هو', 'لازالتا', 'تلكما', 'وقت', 'أولئكم', 'نحو', 'حسب', 'نحن', 'لئن', 'اللذين', 'أب', 'أخ', 'قبل', 'بدون',
'مادامتا', 'بئس', 'د', 'ذا', 'لعمر', 'هاهنا', 'كما', 'لستم', 'لستن', 'حتى', 'لدى', 'ذه', 'ذي', 'ذو', 'أي', 'أو', 'أف', 'أن', 'أم', 'نعما', 'هيت', 'هيا', 'مابرح', 'حينما',
'هلا', 'إنما', 'جعل', 'كخ', 'ت', 'عما', 'بكن', 'بكم', 'مازلنا', 'غ', 'رويدك', 'دون', 'أولئك', 'كي', 'هؤلاء', 'لها', 'ي', 'مكانكما', 'كم', 'كل', 'مافتئتن',
'ماانفككتما', 'لعل', 'إنا', 'لازال', 'متى', 'مابرحوا', 'خلال', 'مازلتن', 'مازلتم', 'ثنا', 'راح', 'هلم', 'لولا', 'مابرحتما', 'مافتئت', 'مافتئتم', 'تان', 'مرة', 'أصلا',
'وشكان', 'كأنما', 'إياهما', 'أسفل', 'ص', 'أمامك', 'م', 'مادمتما', 'إياهم', 'ماانفكوا', 'استحال', 'إليك', 'أمسى', 'انبرى', 'مافتئتا', 'أنتم', 'لازالوا', 'أنتن', 'ليستا',
'ماانفككن', 'أوشك', 'طفق', 'لازالت', 'صه', 'لازالا', 'هناك', 'ليس', 'هاذين', 'مازالوا', 'إلا', 'سوف', 'ب', 'آمين', 'آناء', 'هاتين', 'ليت', 'و', 'أكثر', 'إلى',
'أثناء', 'ابن', 'بعدئذ', 'فلان', 'ليل', 'تين', 'أصبح', 'لكما', 'كأين', 'قد', 'ماانفككت', 'قط', 'لحظة', 'شمال', 'ثم', 'أجل', 'عليك', 'دونك', 'ذلكما', 'حما', 'أنت',
'التي', 'أوه', 'أنا', 'مثل', 'هذين', 'س', 'لازلتما', 'ك', 'أنى', 'حمو', 'حمي', 'هما', 'جير', 'أبدا', 'اللذان', 'عند', 'ظل', 'سرعان', 'خلف', 'تحول', 'هاته', 'هاتي',
'حوالى', 'جلل', 'تينك', 'أولاء', 'آنذاك', 'بعد', 'حين', 'ذيت', 'أيان', 'قلما', 'مازالا', 'خ', 'إزاء', 'بعض', 'فيما', 'لنا', 'مكانكن', 'تلكم', 'ذين', 'اللائي', 'تلقاء',
'خلا', 'حيث', 'مافتئوا', 'صار', 'أينما', 'يمين', 'رب', 'ماذا', 'لست', 'مافتئن', 'ذان', 'إياه', 'ذاك', 'ة', 'ن', 'حذار', 'ماانفكتا', 'ع', 'مادمتم', 'مادمتن', 'مابرحنا',
'ى', 'ذات', 'واها', 'أما', 'جدا', 'أمد', 'اللواتي', 'أمس', 'لسن', 'معاذ', 'الذين', 'حيثما', 'تحت', 'غير', 'حاشا', 'مافتئ', 'مابرحتن', 'مابرحتم', 'أمام', 'ج',
'أضحى', 'اللتيا', 'هاتان', 'هكذا', 'مابرحتا', 'كلتا', 'بضع', 'انقلب', 'اللتين', 'ساء', 'لستما', 'إليكن', 'إليكم', 'غدا', 'حاي', 'كلما', 'تبدل', 'مكانكم', 'كذلك', 'شتان',
'الآن', 'أولالك', 'إياي', 'ممن', 'إياك', 'كيت', 'بعدما', 'اخلولق', 'حرى', 'مادمنا', 'ط', 'ه', 'عاد', 'كيف', 'إياكم', 'إياكن', 'سبحان', 'هذان', 'أقبل', 'الألاء', 'مما',
'ليسوا', 'لما', 'ضحوة', 'مابرحا', 'آها', 'ماداموا', 'آ', 'كاد', 'عل', 'عن', 'مافتئتما', 'ز', 'لازلت', 'بيد', 'ق', 'ظ', 'ماداما', 'كان', 'كلاهما', 'بين', 'هيهات', 'إذا',
```

'إياكما', 'الذي', 'بل', 'نخ', 'طاق', 'أبو', 'أبي', 'هذا', 'لاسيما', 'هنا', 'ساءما', 'ذينك', 'آه', 'ساعة', 'ح', 'ابتدأ', 'صباح', 'ضمن', 'لكيلا', 'بمن', 'كيفما', 'سوى',
'إياها', 'تانك', 'أقل', 'إليكما', 'لئلا', 'كذا', 'آض', 'بما', 'ماانفكا', 'فيم', 'إياهن', 'فا', 'هذي', 'أبا', 'هذه', 'تجاه', 'شطر', 'حينئذ', 'أيضا', 'شرق', 'رجع', 'عامة',
'إذن', 'غروب', 'عندما', 'وراءك', 'إذ', 'ألا', 'لكي', 'مابرحت', 'لكن', 'لكم', 'يومئذ', 'لكنما', 'قبلما', 'بي', 'ذ', 'ماانفككنا', 'بك', 'بن', 'به', 'اللتان', 'ماانفككتم',
'تارة', 'بس', 'مابرحن', 'فقط', 'أنشأ', 'ماانفككتن', 'بخ', 'مثلما', 'شبه', 'بها', 'مازلتما', 'فوق', 'شهر', 'ماانفكت', 'إي', 'اللاتي', 'مازال', 'إن', 'كأن', 'كليهما',
'مافتئا', 'كليكما', 'شرع', 'ليسا', 'مكانك', 'ليست', 'وي', 'أعلى', 'لسنا', 'ثمة', 'لات', 'حسبما', 'ث', 'طالما', 'ذوا', 'مع', 'ماانفك', 'نفس', 'مذ', 'بكما', 'ذوي',
'ذوو', 'مه', 'من', 'حبذا', 'مادام', 'عدا', 'بينما', 'وراء', 'بات', 'عدس', 'وا', 'كلا', 'على', 'علق', 'مازلن', 'حار', 'وراءكن', 'مازالتا', 'وراءكم', 'مادمت',
'أخي', 'أخو', 'قام', 'ئ', 'إذما', 'أين', 'بطآن', 'الألى', 'ض', 'ذلكم', 'ذلكن', 'أية', 'نعم', 'أيا', 'هاك', 'ذواتي', 'مازلت', 'آنئذ', 'أخذ', 'ذلك', 'عندئذ', 'أخا',
'ويكأن', 'إيانا', 'كرب', 'لي', 'لو', 'لن', 'له', 'لم', 'لك', 'بلا', 'مادمن', 'ذانك', 'إما', 'هنالك', 'مادامت', 'ء', 'ريث', 'طق', 'جميع', 'حول', 'ر', 'مهما', 'تي',
["كأي", 'لوما', 'ته', 'ف', 'فو', 'في', 'أ', 'أول', 'مافتئنا', 'مساء', 'بماذا', 'عين', 'لا', 'جنب', 'تلك', 'بله', 'منذ', 'أنتما', 'وراءكما', 'لازلن', 'ما', 'بلى

و 13 الف كلمة للكلمات بتصريفاتها المختلفة

```
len(stp.stopwords_list())
stp.stopwords_list()
```

و يمكن فحص الكلمة هل هي من كلمات التوقف ام لا من هنا :

```
stp.is_stop(u'ممكن')
stp.is_stop(u'منكم')
```

او يمكن اظهار جميع تصريفات كلمة معينة من هنا

```
len(stp.stopword_forms(u"على"))
stp.stopword_forms(u"على")
```

بالإضافة الي هذا , هناك ملف به 750 كلمة باللغة العربية , قام به محمد طاهر الرفاعي , هنا :

https://github.com/mohataher/arabic-stop-words

و أيضا ملف excel فيه العديد من التصنيفات و المعلومات عن كلمات توقف باللغة العربية هنا :

https://sourceforge.net/projects/arabicstopwords/files/