**الكلمات المستبعدة**

# الكلمات الشائعة (stop words) أو الكلمات المستبعدة هي الكلمات التي تستبعد قبل المعالجة اللغوية الآلية للبيانات (نصوص). وهي الكلمات التي تتكرر في النصوص مثل (في، من، إلى،...) ويستحسن تجاهلها عدم فهرستها من أجل تحسين البحث.ويرجع الفضل لهانس بيتر لوهن (من الرواد في استرجاع المعلومات) في استعمال هذا المصطلح والمفهوم في التطوير.

# وتعتبر هذه الطريقة مقاربة سلبية في التعامل مع الأدوات وحروف المعاني في الجمل.

# لا توجد قائمة كلمات محددة للاستبعاد، تستعملها كل المحللات الصرفية وأدوات معالجة اللغة الطبيعية. لا تستعمل كل أدوات المعالجة اللغوية قائمة كلمات شائعة.

# ويمكن للتحليل الصرفي التقليل من استخدام كلمات مستبعدة، لا سيما عند تحليل الجمل.

# الكلمات المستبعدة قد تسبب مشاكل عند استخدام محرك بحث للبحث عن العبارات التي تدخل فيها، لا سيما في أسماء مثل 'المفعول به' ، '' ، أو 'سول من الله"، "وهم".

الكلمات الشائعة في اللغة العربية

يعتمد الكثيرون من الباحثين عن المستبعدات على جردها جردا يدويا، ومن ثم توليد كل الأشكال الممكنة للكلمة وإدراجها في القائمة، مما يزيد في طول القائمة.

لكن يوجد اختلاف في اعتبار الكلمة مستبعدة، فمثلا يمكن أن تعتبر أسماء الأعداد مستبعدة في نص ما، لكنها تصبح ذات قيمة في نص اقتصادي، وكلمات مثل أسماء الشهور، لها أهمية في النصوص التاريخية، لذا توجب أن يتم تصنيف الكلمات، يمكن للمبرمج فيما بعد أن يحدد ما هي الكلمات التي يريدها.

كما قمنا بوضع قائمة أساسية مصنفة للكلمات، ومن ثم توليد كل الأشكال الممكنة، وذلك حسب التصنيف، فمثلا حروف الجر يمكن أن تتصل بالضمائر، مثل ( إلى، إليه، إليهما، إليهم، إليكم...) بينما لا تتصل أسماء الأفعال بالضمائر.

كما يلجأ المبرمجون إلى إدراج الأشكال المنتظمة للكلمات مثل توحيد شكل الهمزة والتاء المربوطة، وألفات، والألف المقصورة، وهذا يمكن أن يتم برمجيا ابتداء من القائمة الأساسية.

**تصنيفات المستبعدات**

|  |  |  |  |
| --- | --- | --- | --- |
| وحدات القياس |  |  | : الرطل، الذراع، الكيلو متر |
| أسماء العملة |  |  |  |
| أسماء الأشهر |  | | العربية والأعجمية والمغربية |
| الأعداد والترتيب |  |  |  |
| ظروف المكان والزمان |  |  |  |
| الحروف العربية وأسمائها |  |  |  |
| الضمائر | 23 | أسماء |  |
| أسماء الإشارة | 26 | أسماء |  |
| أسماء الموصول | 17 | أسماء |  |
| أسماء الشرط | 10 | أسماء |  |
| أسماء الاستفهام | 11 | أسماء |  |
| أسماء الكناية | 8 | أسماء |  |
| أسماء الأفعال | 24 | أسماء |  |
| الظروف والجهات الستة | 48 | أسماء |  |
| أسماء الإضافة | 20 | أسماء |  |
| الأسماء الخمسة | 5 | أسماء |  |
| أسماء الأعداد | 14 | أسماء |  |
| أسماء التوكيد | 8 | أسماء |  |
| أسماء الاستثناء | 4 | أسماء |  |
| المنصوبات | 28 | أسماء |  |
| أفعال القلوب | 21 | الأفعال |  |
| المتعدية إلى مفعولين | 7 | الأفعال |  |
| المتعدية إلى ثلاث مفاعيل | 8 | الأفعال |  |
| أفعال التعجب | 2 | الأفعال |  |
| أفعال المدح والذم | 4 | الأفعال |  |
| الأفعال الجامدة | 2 | الأفعال |  |
| كان وأخواتها | 24 | النواسخ |  |
| كاد وأخواتها | 17 | النواسخ |  |
| الحروف المشبهة بليس | 4 | النواسخ |  |
| إنّ وأخواتها | 6 | النواسخ |  |
| لا النافية للجنس | 1 | النواسخ |  |
| الحروف حروف | 57 | الحروف | |
| حروف وأسماء | 19 | الحروف | |
| حروف وأفعال | 4 | الحروف | |

**توليد أشكال المستبعدات**

وضعنا تصنيفا آخر للزيادات الممكنة لكل كلمة مستبعدة، تمكن المبرمج من إضافة الزوائد للحصول على الأشكال المطلوبة.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **العطف** | **التعريف** | **الجر** | **الإضافة** | **التصريف** |
| رمز القابلية | ع | ت | ج | ض | ص |
| مثال | وأول | الأول | بأول | أولهم | كاد، يكاد |
| غير قابل | \* | \* | \* | \* | \* |
| مثال | بل، ثم | الأفعال، الحروف | الأفعال، الحروف | الأفعال | الأسماء |

**تطبيقات المستبعدات**

يمكن استعمال المستبعدات في تطبيقات المعالجة الآلية للغة، ومن بينها التدقيق الإملائي، إذ لشيوع هذه الكلمات، يمكن أن نضعها مباشرة في القاموس، دون حاجة لتحليلها.

**إحصائيات المستبعدات**

للعمل

|  |  |  |  |
| --- | --- | --- | --- |
| عدد الكلمات | الزوائد |  |  |
|  |  |  |  |
|  |  |  |  |

# Stop words[[1]](#footnote-1)

**Stop words** is the name given to words which are filtered out prior to, or after, processing of natural language data (text). [Hans Peter Luhn](http://en.wikipedia.org/wiki/Hans_Peter_Luhn), one of the pioneers in [information retrieval](http://en.wikipedia.org/wiki/Information_retrieval), is credited with coining the phrase and using the concept in his design. It is controlled by human input and not automated. This is sometimes seen as a negative approach to the natural articles of speech as mentioned above.

There is no definite list of stop words which all natural language processing tools incorporate. Not all NLP tools use a stoplist. Some tools specifically avoid using them to support [phrase searching](http://en.wikipedia.org/w/index.php?title=Phrase_searching&action=edit&redlink=1). The use of a [stemming](http://en.wikipedia.org/wiki/Stemming) algorithm may reduce part of the rationale or dependence on a stoplist to filter out words.[[*citation needed*](http://en.wikipedia.org/wiki/Wikipedia:Citation_needed)]

Stop words can cause problems when using a [search engine](http://en.wikipedia.org/wiki/Search_engine) to search for phrases that include them, particularly in names such as '[The Who](http://en.wikipedia.org/wiki/The_Who)', '[The The](http://en.wikipedia.org/wiki/The_The)', or '[Take That](http://en.wikipedia.org/wiki/Take_That)'.

# Mot vide[[2]](#footnote-2)

Les **mots vides** (ou *stop words*, en anglais) sont des [mots](http://fr.wikipedia.org/wiki/Mot) qui sont tellement [communs](http://fr.wikipedia.org/w/index.php?title=Commun&action=edit&redlink=1) qu'il est inutile de les indexer ou de les utiliser dans une recherche. En français, des mots vides évidents pourraient être « [le](http://fr.wikipedia.org/wiki/Article_d%C3%A9fini) », « [la](http://fr.wikipedia.org/wiki/Article_d%C3%A9fini) », « [de](http://fr.wikipedia.org/wiki/D%C3%A9terminant_%28grammaire%29) », « [du](http://fr.wikipedia.org/wiki/D%C3%A9terminant_%28grammaire%29) », « [ce](http://fr.wikipedia.org/wiki/D%C3%A9terminant_%28grammaire%29) », « [ça](http://fr.wiktionary.org/wiki/%C3%A7a) », ...

Un mot vide est un mot non significatif figurant dans un texte. On l'oppose à [mot plein](http://fr.wikipedia.org/wiki/Mot_plein). La signification d'un mot s'évalue à partir de sa [distribution](http://fr.wikipedia.org/wiki/Loi_de_probabilit%C3%A9) (au sens [statistique](http://fr.wikipedia.org/wiki/Statistique)) dans une [collection](http://fr.wikipedia.org/wiki/Collection_%28%C3%A9dition%29) de textes. Un mot dont la distribution est uniforme sur les textes de la collection est dit "vide". En d'autres termes, un mot qui apparaît avec une [fréquence](http://fr.wikipedia.org/wiki/Fr%C3%A9quence) semblable dans chacun des textes de la collection n'est pas discriminant, ne permet pas de distinguer les textes les uns par rapport aux autres.

Lorsque tous les textes de la collection sont rédigés dans une même [langue](http://fr.wikipedia.org/wiki/Langue), les mots vides sont principalement des mots caractéristiques de cette langue comme les [prépositions](http://fr.wikipedia.org/wiki/Pr%C3%A9position), les [articles](http://fr.wikipedia.org/wiki/D%C3%A9terminant_%28grammaire%29), les [pronoms](http://fr.wikipedia.org/wiki/Pronom). D'où l'assimilation courante entre *mots vides* et *mots grammaticaux* et partant, entre *mots pleins* et *mots lexicaux* ([noms](http://fr.wikipedia.org/wiki/Nom), [verbes](http://fr.wikipedia.org/wiki/Verbe), [adjectifs](http://fr.wikipedia.org/wiki/Adjectif)). Les listes pré-établies, dites de mots vides, utilisées par les [moteurs de recherche](http://fr.wikipedia.org/wiki/Moteur_de_recherche) sont ainsi des listes de mots grammaticaux.

Cependant dans une collection de textes réunis autour d'un [thème](http://fr.wikipedia.org/wiki/Th%C3%A8me_%28linguistique%29) commun, certains mots peuvent respecter une distribution uniforme. Ce sont alors des mots vides pour cette collection bien qu'ils ne soient pas des mots grammaticaux. D'autre part, certains mots grammaticaux sont assez rares pour constituer des mots pleins (ex. : *via*).

### الاستبعاد التلقائي للكلمات الشائعة[[3]](#footnote-3)

يتجاهل بحث Google الكلمات والرموز الشائعة مثل "أين" و"كيف"، وكذلك الأرقام المفردة والحروف المفردة (باللغة الإنجليزية)، لأنها تؤدي إلى تقليل سرعة بحثك دون تحسين النتائج.

وإذا كانت هذه الكلمة الشائعة أساسيةً في الحصول على النتائج التي تريدها، يمكنك تضمينها بوضع علامة "+" أمامها. (تأكد من ترك مسافة قبل علامة "+".)

وهناك طريقة أخرى لفعل ذلك وهي إجراء بحث باستخدام عبارات وهو ما يعني ببساطة وضع علاماتي اقتباس حول كلمتين أو أكثر. الكلمات الشائعة في بحث العبارات (مثل، "أين أنت") يتم تضمينها في البحث.

# المراجع:

1. أنطوان الدحداح، معجم قواعد اللغة العربية،في جداول ولوحات، مكتبة لبنان ناشرون، 2008.
2. Leah S. et al Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis, Univ. of Massachusetts, Dept. of Computer Science, 2002
3. Ibrahim Abu El-Khair, Effect of stop words elimination for Arabic information retrieval: a comparative study, Dept of library and information science, faculty of arts, Minia University-Egypt, [iabuelkhair@gmail.com](mailto:iabuelkhair@gmail.com), international journal of computing & information sciences, vol. 4 No. 3 December 2006, On-Line.

1. <http://en.wikipedia.org/wiki/Stop_words> [↑](#footnote-ref-1)
2. http://fr.wikipedia.org/wiki/Mot\_vide [↑](#footnote-ref-2)
3. <http://www.google.com/support/websearch/bin/answer.py?answer=35889> [↑](#footnote-ref-3)