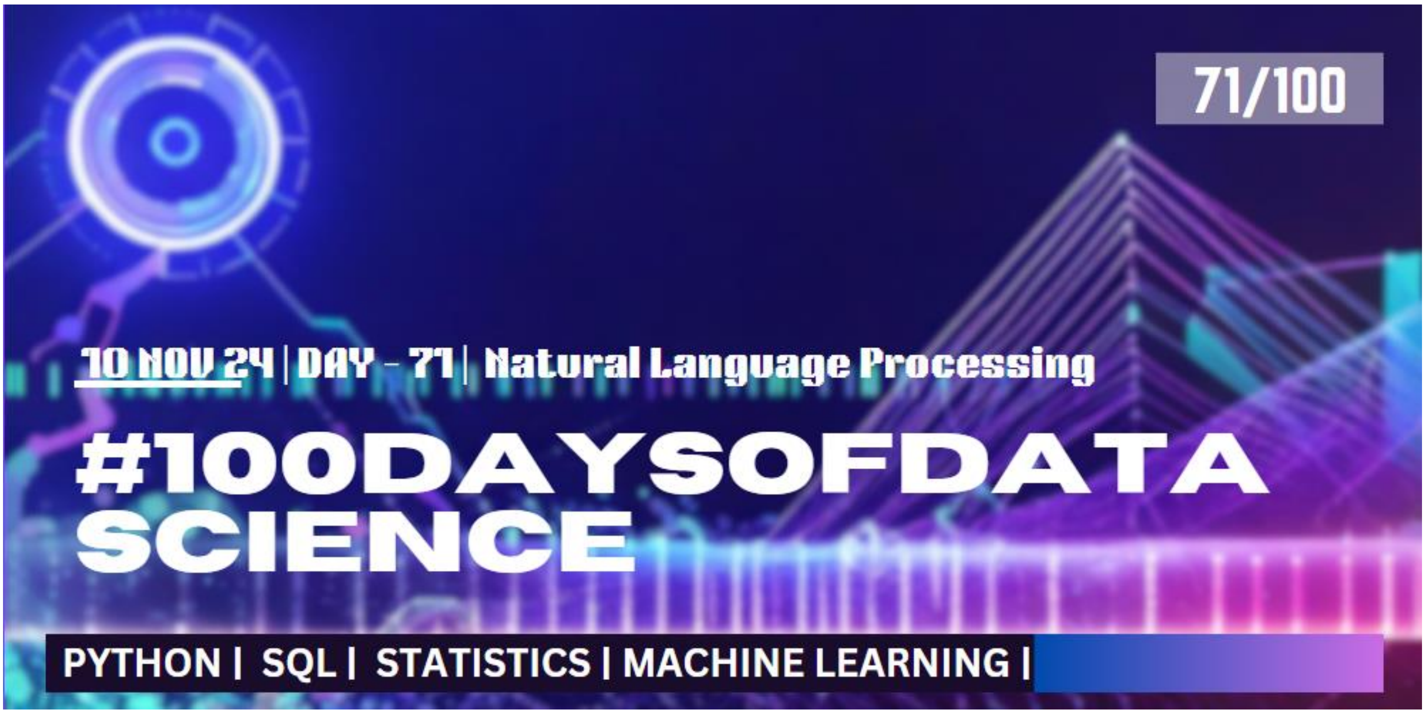

71/100
10 NOV 24 | DAY - 71 | Natural Language Processing
#100DAYSOFDATA SCIENCE
PYTHON | SQL | STATISTICS | MACHINE LEARNING |

Stopwords

What Are Stopwords?

Stopwords are common words in a language, such as "and," "the," "is," "in," and "at," that often appear frequently in text data but add little meaning or context. In NLP, removing stopwords is a standard preprocessing step to help models focus on the main content of the text, making data analysis more efficient and improving the performance of algorithms by eliminating these “filler” words.

Key Aspects of Stopwords:

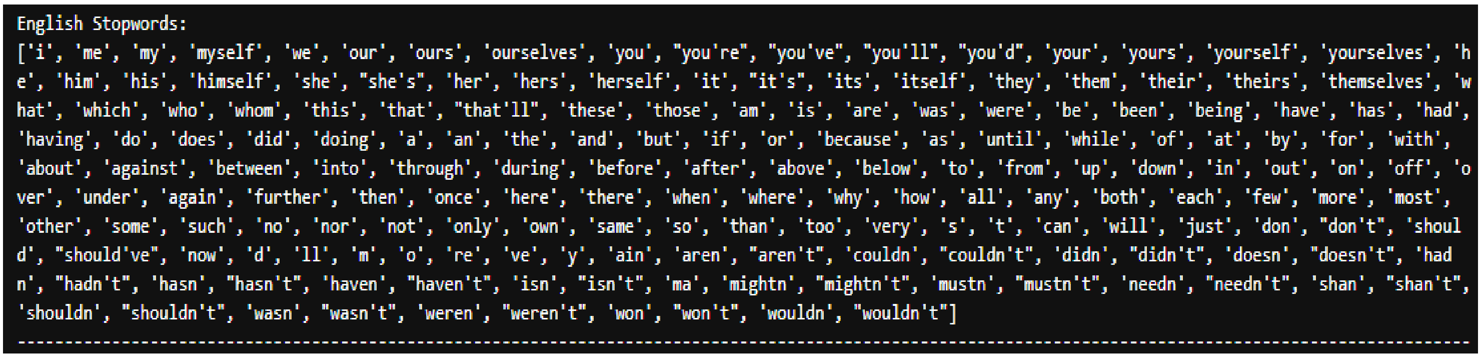
1. **Efficiency**: By removing frequent but uninformative words, text data becomes easier and faster to process.
2. **Focus**: Filtering out stopwords enables models to focus on more meaningful words that contribute to the context and insights.
3. **Use Cases**: Stopwords removal is crucial in tasks like sentiment analysis, search engines, and text summarization, where high-frequency words can dilute the meaningful content.

Common Libraries for Stopword Removal:

1. **NLTK**: Provides a built-in list of stopwords for many languages, allowing you to customize and expand the list for specific needs.
2. **SpaCy**: Includes a customizable set of stopwords that can be easily adjusted by adding or removing words as per project requirements.

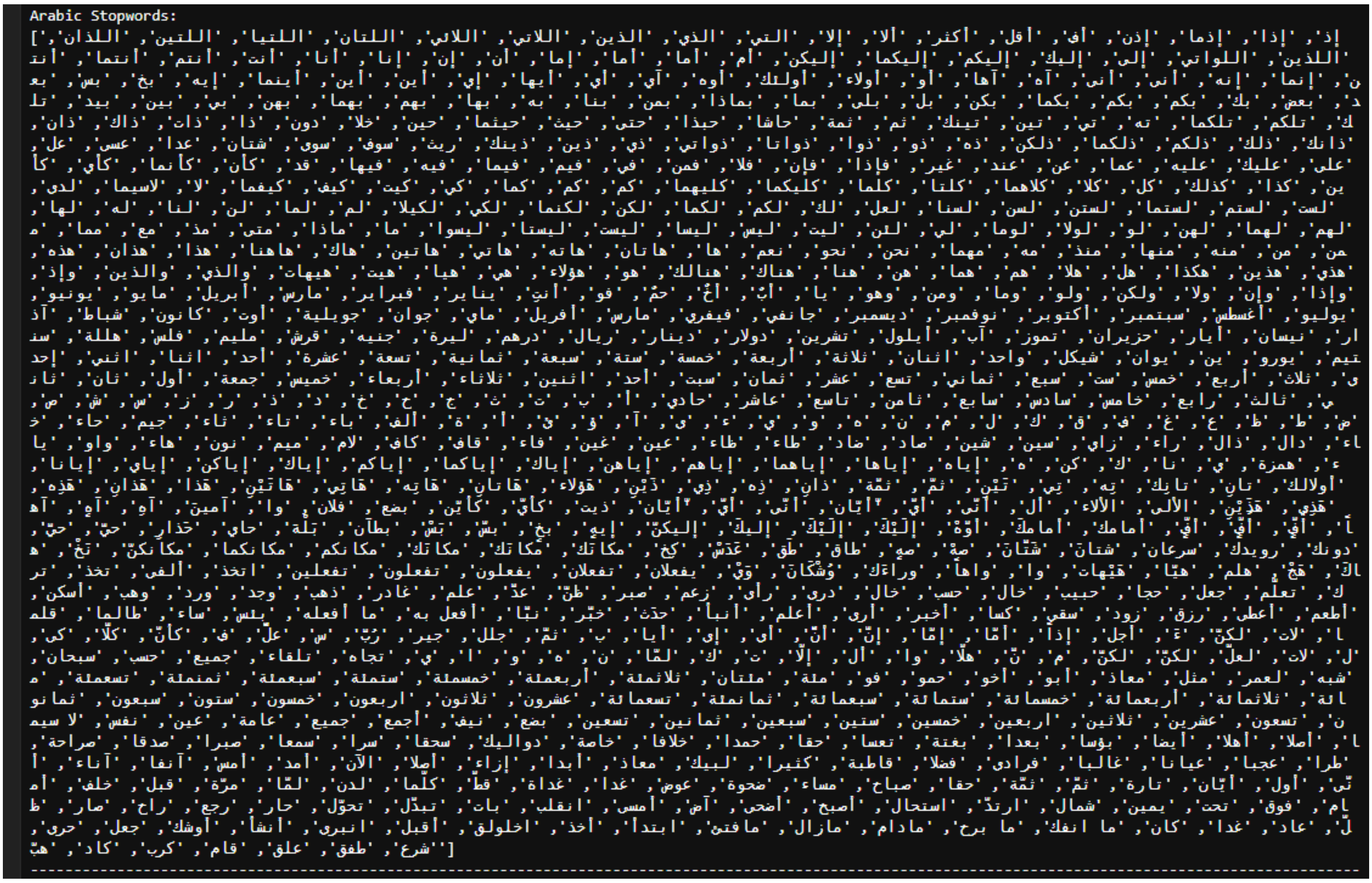
Sample List of Common Stopwords in English:

- **Domain-Specific Stopwords**: Some NLP projects may define custom stopwords, such as "data," "science," or "technology," depending on the analysis requirements.



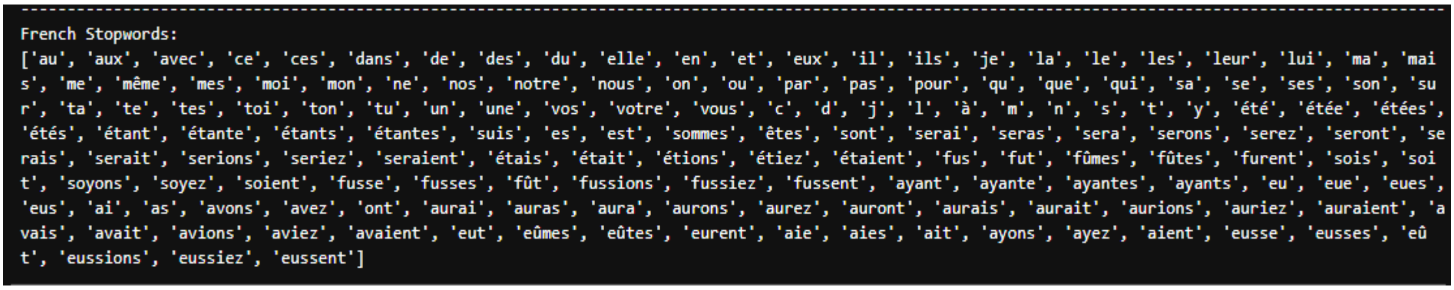
```
English Stopwords:
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'h
e', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'w
hat', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had',
'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with',
'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'o
ver', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most',
'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'shoul
d', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'had
n', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't",
'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
------------------------------------------------------------------------------------------------------------------------------------------------------
```

Sample List of Common Stopwords in Arabic:



```
Arabic Stopwords:
['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', 'أنت
ن', 'إنما', 'إنه', 'أنى', 'أنى', 'آه', 'آها', 'أو', 'أولاء', 'أولئك', 'أوه', 'آي', 'أي', 'أيها', 'إي', 'أين', 'أين', 'أينما', 'إيه', 'بخ', 'بس', 'بع
د', 'بعض', 'بك', 'بكم', 'بكم', 'بكما', 'بكن', 'بل', 'بلى', 'بما', 'بماذا', 'بمن', 'بنا', 'به', 'بها', 'بهم', 'بهما', 'بهن', 'بي', 'بين', 'بيد', 'تل
ك', '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', 'كأ
ين', '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', 'م
من', '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', 'أبٌ', 'أخٌ', 'حمٌ', 'فو', 'أنتِ', 'يناير', 'فبراير', 'مارس', 'أبريل', 'مايو', 'يونيو',
'يوليو', 'أغسطس', 'سبتمبر', 'أكتوبر', 'نوفمبر', 'ديسمبر', 'جانفي', 'فيفري', 'مارس', 'أفريل', 'ماي', 'جوان', 'جويلية', 'أوت', 'كانون', 'شباط', 'آذ
ار', 'نيسان', 'أيار', 'حزيران', 'تموز', 'آب', 'أيلول', 'تشرين', 'دولار', 'دينار', 'ريال', 'درهم', 'ليرة', 'جنيه', 'قرش', 'مليم', 'فلس', 'هللة', 'سن
تيم', 'يورو', 'ين', 'يوان', 'شيكل', 'واحد', 'اثنان', 'ثلاثة', 'أربعة', 'خمسة', 'ستة', 'سبعة', 'ثمانية', 'تسعة', 'عشرة', 'أحد', 'اثنا', 'اثني', 'إحد
ى', 'ثلاث', 'أربع', 'خمس', 'ست', 'سبع', 'ثماني', 'تسع', 'عشر', 'ثمان', 'سبت', 'أحد', 'اثنين', 'ثلاثاء', 'أربعاء', 'خميس', 'جمعة', 'أول', 'ثان', 'ثان
ي', 'ثالث', 'رابع', 'خامس', 'سادس', 'سابع', 'ثامن', 'تاسع', 'عاشر', 'حادي', 'أ', 'ب', 'ت', 'ث', 'ج', 'ح', 'خ', 'د', 'ذ', 'ر', 'ز', 'س', 'ش', 'ص',
'ض', 'ط', 'ظ', 'ع', 'غ', 'ف', 'ق', 'ك', 'ل', 'م', 'ن', 'ه', 'و', 'ي', 'ء', 'ى', 'آ', 'ؤ', 'ئ', 'أ', 'ة', 'ألف', 'باء', 'تاء', 'ثاء', 'جيم', 'حاء', 'خ
اء', 'دال', 'ذال', 'راء', 'زاي', 'سين', 'شين', 'صاد', 'ضاد', 'طاء', 'ظاء', 'عين', 'غين', 'فاء', 'قاف', 'كاف', 'لام', 'ميم', 'نون', 'هاء', 'واو', 'يا
ء', 'همزة', 'ي', 'نا', 'ك', 'كن', 'ه', 'إياه', 'إياها', 'إياهما', 'إياهم', 'إياهن', 'إياك', 'إياكما', 'إياكم', 'إياك', 'إياكن', 'إياي', 'إيانا',
'أولالك', 'تانِ', 'تانِك', 'تِه', 'تِي', 'تَيْنِ', 'ثمّ', 'ثمّة', 'ذانِ', 'ذِه', 'ذِي', 'ذَيْنِ', 'هَؤلاء', 'هَاتانِ', 'هَاتِه', 'هَاتِي', 'هَاتَيْنِ', 'هَذا', 'هَذانِ', 'هَذِه',
'هَذِي', 'هَذَيْنِ', 'الألى', 'الألاء', 'ال', 'أنّى', 'أيّ', 'ّأيّان', 'أنّى', 'أيّ', 'ّأيّان', 'ذيت', 'كأيّ', 'كأيّن', 'بضع', 'فلان', 'وا', 'آمينَ', 'آهِ', 'آهٍ', 'آه
اً', 'أفٍّ', 'أفٍّ', 'أفٍّ', 'أمامك', 'أمامكَ', 'أوّهْ', 'إلَيْكَ', 'إلَيْكَ', 'إليكَ', 'إليكنّ', 'إيهٍ', 'بخٍ', 'بسّ', 'بَسْ', 'بطآن', 'بَلْهَ', 'حاي', 'حَذارِ', 'حيَّ', 'حيَّ',
'دونك', 'رويدك', 'سرعان', 'شتانَ', 'شَتَّانَ', 'صهْ', 'صهٍ', 'طاق', 'طَق', 'عَدَسْ', 'كِخ', 'مكانَك', 'مكانَك', 'مكانَك', 'مكانكم', 'مكانكما', 'مكانكنّ', 'نَخْ', 'ه
اكَ', 'هَجْ', 'هلم', 'هيّا', 'هَيْهات', 'وا', 'واهاً', 'وراءَك', 'وُشْكَانَ', 'وَيْ', 'يفعلان', 'تفعلان', 'يفعلون', 'تفعلون', 'تفعلين', 'اتخذ', 'ألفى', 'تخذ', 'تر
ك', 'تعلَّم', 'جعل', 'حجا', 'حبيب', 'خال', 'حسب', 'خال', 'درى', 'رأى', 'زعم', 'صبر', 'ظنَّ', 'عدَّ', 'علم', 'غادر', 'ذهب', 'وجد', 'ورد', 'وهب', 'أسكن',
'أطعم', 'أعطى', 'رزق', 'زود', 'سقى', 'كسا', 'أخبر', 'أرى', 'أعلم', 'أنبأ', 'حدَث', 'خبَّر', 'نبَّا', 'أفعل به', 'ما أفعله', 'بئس', 'ساء', 'طالما', 'قلم
ا', 'لات', 'لكنَّ', 'ءَ', 'أجل', 'إذاً', 'أمّا', 'إمّا', 'إنَّ', 'أنَّ', 'أى', 'إى', 'أيا', 'ب', 'ثمَّ', 'جلل', 'جير', 'رُبَّ', 'س', 'علًّ', 'ف', 'كأنّ', 'كلَّا', 'كى',
'ل', 'لات', 'لعلَّ', 'لكنَّ', 'لكنَّ', 'م', 'نَّ', 'هلّا', 'وا', 'أل', 'إلّا', 'ت', 'ك', 'لمّا', 'ن', 'ه', 'و', 'ا', 'ي', 'تجاه', 'تلقاء', 'جميع', 'حسب', 'سبحان',
'شبه', 'لعمر', 'مثل', 'معاذ', 'أبو', 'أخو', 'حمو', 'فو', 'مئة', 'مئتان', 'ثلاثمئة', 'أربعمئة', 'خمسمئة', 'ستمئة', 'سبعمئة', 'ثمنمئة', 'تسعمئة', 'م
ائة', 'ثلاثمائة', 'أربعمائة', 'خمسمائة', 'ستمائة', 'سبعمائة', 'ثمانمئة', 'تسعمائة', 'عشرون', 'ثلاثون', 'اربعون', 'خمسون', 'ستون', 'سبعون', 'ثمانو
ن', 'تسعون', 'عشرين', 'ثلاثين', 'اربعين', 'خمسين', 'ستين', 'سبعين', 'ثمانين', 'تسعين', 'بضع', 'نيف', 'أجمع', 'جميع', 'عامة', 'عين', 'نفس', 'لا سيم
ا', 'أصلا', 'أهلا', 'أيضا', 'بؤسا', 'بعدا', 'بغتة', 'تعسا', 'حقا', 'حمدا', 'خلافا', 'خاصة', 'دواليك', 'سحقا', 'سرا', 'سمعا', 'صبرا', 'صدقا', 'صراحة',
'طرا', 'عجبا', 'عيانا', 'غالبا', 'فرادى', 'فضلا', 'قاطبة', 'كثيرا', 'لبيك', 'معاذ', 'أبدا', 'إزاء', 'أصلا', 'الآن', 'أمد', 'أمس', 'آنفا', 'آناء', 'أ
نّى', 'أول', 'أيّان', 'تارة', 'ثمّ', 'ثمّة', 'حقا', 'صباح', 'مساء', 'ضحوة', 'عوض', 'غدا', 'غداة', 'قطّ', 'كلّما', 'لدن', 'لمّا', 'مرّة', 'قبل', 'خلف', 'أم
ام', 'فوق', 'تحت', 'يمين', 'شمال', 'ارتدّ', 'استحال', 'أصبح', 'أضحى', 'آض', 'أمسى', 'انقلب', 'بات', 'تبدّل', 'تحوّل', 'حار', 'رجع', 'راح', 'صار', 'ظ
لّ', 'عاد', 'غدا', 'كان', 'ما انفك', 'ما برح', 'مادام', 'مازال', 'مافتئ', 'ابتدأ', 'أخذ', 'اخلولق', 'أقبل', 'انبرى', 'أنشأ', 'أوشك', 'جعل', 'حرى',
'شرع', 'طفق', 'علق', 'قام', 'كرب', 'كاد', 'هبّ']
--------------------------------------------------------------------------------------------------------------------------------
```

Sample List of Common Stopwords in French:



```
--------------------------------------------------------------------------------------------------------------------------------
French Stopwords:
['au', 'aux', 'avec', 'ce', 'ces', 'dans', 'de', 'des', 'du', 'elle', 'en', 'et', 'eux', 'il', 'ils', 'je', 'la', 'le', 'les', 'leur', 'lui', 'ma', 'mai
s', 'me', 'même', 'mes', 'moi', 'mon', 'ne', 'nos', 'notre', 'nous', 'on', 'ou', 'par', 'pas', 'pour', 'qu', 'que', 'qui', 'sa', 'se', 'ses', 'son', 'su
r', 'ta', 'te', 'tes', 'toi', 'ton', 'tu', 'un', 'une', 'vos', 'votre', 'vous', 'c', 'd', 'j', 'l', 'à', 'm', 'n', 's', 't', 'y', 'été', 'étée', 'étées',
'étés', 'étant', 'étante', 'étants', 'étantes', 'suis', 'es', 'est', 'sommes', 'êtes', 'sont', 'serai', 'seras', 'sera', 'serons', 'serez', 'seront', 'se
rais', 'serait', 'serions', 'seriez', 'seraient', 'étais', 'était', 'étions', 'étiez', 'étaient', 'fus', 'fut', 'fûmes', 'fûtes', 'furent', 'sois', 'soi
t', 'soyons', 'soyez', 'soient', 'fusse', 'fusses', 'fût', 'fussions', 'fussiez', 'fussent', 'ayant', 'ayante', 'ayantes', 'ayants', 'eu', 'eue', 'eues',
'eus', 'ai', 'as', 'avons', 'avez', 'ont', 'aurai', 'auras', 'aura', 'aurons', 'aurez', 'auront', 'aurais', 'aurait', 'aurions', 'auriez', 'auraient', 'a
vais', 'avait', 'avions', 'aviez', 'avaient', 'eut', 'eûmes', 'eûtes', 'eurent', 'aie', 'aies', 'ait', 'ayons', 'ayez', 'aient', 'eusse', 'eusses', 'eû
t', 'eussions', 'eussiez', 'eussent']
```

Types of Stopwords and Their Impact:

1. **High-Frequency Words**: Words that occur frequently across documents but add minimal meaning. Removing these helps reduce text noise.
2. **Domain-Specific Stopwords**: Some stopwords may be tailored for specific industries or contexts, such as excluding "data" in data science projects, which can improve focus on more relevant terms.

When to Use Stopword Removal:

- **Ideal For**: Tasks where meaningful keywords are essential, like **topic modeling** or **document classification**.
- **Avoid For**: Applications where sentence structure is important, like **text generation** or **language translation**, where stopwords contribute to the natural flow.

Selecting Stopwords: Choosing the right stopwords list depends on the NLP project's context. For instance, legal documents might require a custom list, while social media texts might include common slang and abbreviations.