Novel Processing for Stop Words

Aeesha S. Shaheen
Computer Science Department,
College of Computer Science and Mathematics,
University of Mosul, Mosul, Iraq

Hadia Salih, PhD
Computer Science Department,
College of Computer Science and Mathematics,
University of Mosul, Mosul, Iraq

Amera Ismail Melhum
Computer Science Department,
College of Computer Science,
University of Dohuk, Dohuk, Iraq

ABSTRACT
One of the most difficult tasks that face scholars in a branch of linguistics known as Natural Language (NLP) is how to deal with stop words – those ubiquitous words which add little to the meaning of a text yet tend to be in it. In the most traditional approaches, these are even deleted in order to facilitate analyses of the contents. Some of these explain how these unwanted 'filler' words should be applied using a novel concept of color coding them. Rather than only substituting the 'cut-out' stop word with the appropriate symbol, we replace the symbol with a colored one representing that specific stop word. This revolutionary technique, allows improved visual of text and opens new avenues in text condensation, affective evaluation and even keywords selection. Thanks to the standard distribution of color-coded stop words, it is possible to illustrate visually how each stop word is distributed.

Keywords
stop wors, natural language processing, removal stop words, color-coding, text retravel.

1. INTRODUCTION
Natural Language Processing (NLP) focuses on the confluence of linguistics, computer science and artificial intelligence in understanding and producing human languages by machines. The notion of stop words is one example that embodies this idea. It refers to such terms which are used frequently in communication that many terms such as 'the', 'and,' or 'is,' which even if grammatical terms, are of little significance in the analysis. For the most cases, these are usually silenced in order to concentrate on more meaningful content. However, cutting them out needs can also cut out necessary linguistics order which can lead to losing some important cohesion links. [5,11]

In this paper, we offer an alternative approach to the problem concerning stop words by creating hue-stop words. Rather than omitting them, we rather replace each stop word with a symbol which is in a different color, hence visually illustrating their presence and relevance in the text. This not only preserves the grammatical correctness of the passage but enhances stop word analysis focusing on its distribution and overall function in different NLP processes. Applying state-of-the-art natural language processing tools and techniques, we demonstrate the purpose of employing textual color coding in relation to stop words in warts like text summarization, sentiment analysis and keyword extraction among others that are not always appreciated by most scholars. [2,7]

As an extension, we update the current stop word list of 2020 and can ensure that sources and depositor evidenced accuracy come from multiple lists. This list is used to develop our color-coding procedure, thus enhancing internal consistency in the procedure irrespective of the text and application context. Our findings suggest that learning color coded stop words can enhance data mining, pattern recognition, and decision makings in NLP, but more interestingly also adds a novel way of teaching language processing basics.

While the concept of utilizing color-coding in the use of stop words is new and develops leaves room for improvement, we do not encourage the complete exclusion of other relevant methods used in NLP such as stop word removal. The aspiration we have is to augment opposed to supplant these common practices with ways in which one is able to grasp the complex world of natural language. In the following sections, we show how the color-coding method of stop words is implemented and the limitations that maybe there with regards to them and providing tips on how to use them in real life settings. [2,4]

2. RELATED RESEARCHES:
The subject of how stop words are dealt with in natural language processing is well documented with quite a number of works looking into their impact on several activities per tasks. Major observations in this regard include:

- Social Media Analysis: social media has also been the focus of earlier studies, and stop words assist in opinion countering, information diffusion, and sentiment analysis of micro-blogging sites characterized by language usage distinct from all other forms of writing (Li et al., 2014).
- Multilingual NLP: The basic task of cross-language NLP is diagnosis of the difference in content and linguistic systems of various tongues and comprehension of stop words are variable of course. (Molina-González & Ananiadou, 2016).
- Stop Word List Quality: It has been studied how well the various lists of stop words currently available with the NLP toolkits and packages have been assessed. Certain research in the use of these lists established cases of lack of compliance to certain dictations besides underutilization of certain words to optimal levels (Nothman et al. , 2018).
- Visualization Techniques: Recent studies have also focused on the other novel approaches in visualizing stop words like the word clouds in which the stop words have been either underlined or presented in different colors. These visualizations can help find out main themes and most frequent stop words in the text corpus (He et al. , 2021).

Recent works in the context of pretraining and fine-tuning of the neural language models such as BERT and GPT-3 has also focus on use of stop words. It has been found out in other works by (Dodge et al., 2020), for instance that using fine-tuning with

stop words can enhance the model's performance in other downstream tasks.

All in all, as for the previous papers, they have contributed much to the understanding of the concept of stop words with regards to the effectiveness of different NLP tasks and the peculiarities of the stop word handling. These findings have shaped the future of the best approach and practice that are used in natural language processing and the discovery of another related research. [5,8]

3. ABOUT STOP WORDS

Linguistic filters are one of the basic prerequisites of Natural Language Processing (NLP) and text analysis. It refers to that vocabulary which forms the background of a particular language and which recurrently appears in a language but does not seem to convey any particular meaning while interpreting the text. Stop words in English include: the, and, is, in, of, a, and that Whereas, words like the, and, is, in, of, a and that are supposed to be ignored because they are very common irrespective of the content of the text in question. [1,12]

The importance of stop words can be seen in terms of the effect that they have on NLP jobs and the computation associated with text processing. Whereas, in numerous NLP techniques like text classification, sentiment analysis, information retrieval, and keyword extraction stop words act as noise and hamper the derivation of better and improved results. That is why deleting stop words from the text is the initial data preparation step used to enhance the effectiveness of different NLP procedures. [10]

However, the variances described above should be noted and advised that the process of stop words removal is not always suitable for all NLP tasks. Therefore, in language modeling and treating some other information retrieval tasks, stop words can really be indispensable to retain the syntactic structures and meaning of the text.

With regards to stop words, NLP definitely mandates a special trite given that it is one of the key factors that determine the fastness and accuracy of numerous text analyses. Stop words can either completely remove the context or eliminate noise and a fine line needs to be drawn, color-coding approach is also very useful and can add a lot to text analysis. [6,7,9]

4. STOP WORDS AND NLP:

It is necessary to explain that stop words are one of the basic notions in Natural Language Processing (NLP) and are involved in a number of NLP tasks. This topic of understanding has already been discussed earlier; in Natural language text, stop words are words that don't have a high informational value in the context of a given task and are frequent. Examples of stop words include; articles such as 'the', 'a', 'an', conjunctions like 'and', 'or', prepositions like 'in', 'on', ' of', and some auxiliary verbs 'is', 'am', 'are', 'was'.

In NLP, stop words' functionality is mainly connected with text preparation and feature selection steps. In the preprocessing step, the stop words are usually omitted from the text which helps in the lowering of the number and hence making the processing of the text faster. Stop words are words

that do not offer much meaning in a phrase and thus using them helps in compounding the effects through removing them as it helps in enhancing the output of various NLP tasks such as text classification, sentiment analysis as well as information retrieval. [1,3,13]

5. HOW STOP WORDS IMPACT NLP

1. Computational Efficiency: Therefore, principles of coherent writing state that it is possible to exclude stop words from the text because they just increase the word count and necessary time and overall memory for the text analysis. This is especially the case when it is applied to a text, or text corpus, and real time processing and analysis.
2. Feature Selection: In the tasks such as text classification, the words that are irrelevant to the difference between the classes are used. Thus, removing them the beneficial features for classification of the given task may be fitted into the model.
3. Improved Semantics: In some NLP tasks, if not excluded, stop words may affect the meaning of the text as may be understood by a reader. For instance, if we are working on a sentiment analysis model a negation word such as 'not' might significantly alter the position of the sentiment of a given a sentence. There are certain words like 'not' which if excluded makes the analysis more precise.
4. Reduced Noise: Some terms are often very frequent seeing in most document hence if not handled well they could dominate the term frequency list. They are excluded in the presumption that they only add unnecessary noise to the data and do not impact the model's learning enough.
5. Text Summarization: In cases of text summarization, the stop words are inconsequential and therefore can be eliminated in order to make the summaries accurate and complete.

This makes the removal of stop words a common practice although it is noteworthy that there exist certain cases where it is useful to retain some of the stop words. For example, specific information retrieval problems, can utilize the existence of stop words to define the user's Searching State as accurately as possible.

In general, stop words are an important part of NLP preprocessing as they facilitate enhancing the efficiency of the NLP models and unpacking valuable information from text. Nonetheless, the modulation of stop words to be either eliminated or kept has to be made depending on the task or necessity that is being in use. [5,11,14]

6. PRACTICAL PART

It has led to expanding interest in identifying new ways to manage stop words in the NLP. One of the possibilities is using coloring, where stop words within the text are colored to show the way they are distributed and how they influence the text. Stop words which are highlighted through color-coding prove to be particularly convenient when it gets to search for something or merely monitor the composition of the text.

In this practical part, we will show how to apply several colors to stop words in the selected sample text along the predefined set of the stop words and their related hexadecimal codes. In this example we will be using Python as the programming language.

Step 1: Define the Stop Words and Colors:

Here the stop words list may differ depending on the application and the language to be used. In NLP, there exist well-curated lists of stop words which are essentially lists comprising of common stop words in a given language. These lists may be also made domain specific or task specific and researchers update these lists periodically to incorporate new language trends and patterns, etc.

```
stop_words = {
    "is": "#FF0000",   # Red
    "a": "#00FF00",    # Green
    "that": "#0000FF", # Blue
    "and": "#FFFF00",  # Yellow
    "are": "#FF00FF",  # Magenta
    "the": "#00FFFF"   # Cyan
    # Add more stop words and their colors here if needed
}
```

Color-Coding Scheme:

For this example, we'll use the following **color-coding scheme:**

- **"is" -> #FF0000 (Red)**
- **"a" -> #00FF00 (Green)**
- **"that" -> #0000FF (Blue)**
- **"and" -> #FFFF00 (Yellow)**
- **"are" -> #FF00FF (Magenta)**
- **"the" -> #00FFFF (Cyan)**
- **"of" -> #FFA500 (Orange)**
- **"for" -> #800080 (Purple)**
- **"to" -> #008000 (Dark Green)**

Step 2: Color-Code the Stop Words in Text

We have used the following text excerpt as an example:

The Source Text:

The use of methods from natural language processing1 has become an indispensable tool in applications of data science pervading nearly every scientific discipline2,3. The main challenge is how to extract meaningful information from large and diverse datasets—most of which are comprised of unstructured texts. One of the most common approaches to represent textual data is the so-called bag-of-words model, in which one ignores the order of words within a given document. To improve the signal-to-noise ratio or decrease the amount of data, this is often accompanied by data filtering as part of the data pre-processing steps4. In practice, such activities can take up to 80% of the research effort5. However, we still lack fundamental insights into how these procedures affect the performance of specific algorithms6. For concreteness, we consider topic modelling7, a paradigmatic unsupervised approach for automatic organization of collections of documents8. One contentious pre-processing step in topic modelling is the removal of semantically uninformative words such as 'the'. The most common approach, which goes back more than 50 years, is to curate a "dictionary of insignificant words"9, commonly referred to as a stopword list10. While some stopword lists can appear to practitioners as standard due to being the default choice in popular applications (such as Mallet11), there is no consensus among experts on which words should be excluded12. Indeed, the use of a 'standard' stopword list is problematic because it ignores the domain-knowledge specificity of stopwords13 and because it is language-specific14. The limitation of static lists has motivated the development of other heuristic approaches based on factors such as the number of occurrences (most and least frequent words), document frequency, and term frequency and inverse document frequency (TFIDF)15, and other, often ill-specified, procedures. The state of uncertainty in the field is illustrated by the fact that the seminal paper on latent Dirichlet allocation "removed a standard list of 50 stop words ... [and] ... words that occurred only once"16, but other works by the same author subsequently removed "standard stop words and those that appear too frequently or too rarely"17 or "all words not in a pruned vocabulary of 4,253 words"18, or chose "1,539 terms that occurred in more than five documents"19 or a "5,000-term vocabulary according to tfidf "20. Even when using the same method, such as TFIDF, different authors use different thresholds; for example, ref. 21 removes words "that have tfidf greater than 0.8". The inconsistency in filtering approaches poses severe challenges to the comparison of results across different studies, rendering it nearly impossible to obtain a coherent picture on the state of the field. This is exacerbated by the fact that the removal of stopwords in topic modelling and text-based unsupervised learning more generally is not well understood22, leading to a sterile debate on the usefulness of such approaches.

Text before using **color-coding** scheme contain **464** words, **106** are stop words distributed as:

is appears 11 times

a appears 10 times

that appears 5 times

and appears 8 times

are appears 1 time

the appears 30 times

of appears 26 times

for appears 3 times

to appears 12 times

Step 3: Output

The output will be a color-coded version of the input text after replacing stopwords with the character "&" in deferent color according the **color-coding scheme** above**:**

Color-coded text:

The use & methods from natural language processing1 has become an indispensable tool in applications & data science pervading nearly every scientific discipline2,3. & main challenge & how & extract meaningful information from large & diverse datasets—most & which & comprised & unstructured texts. One & & most common approaches & represent textual data & & so-called bag-of-words model, in which one ignores & order & words within & given document. & improve & signal-to-noise ratio or decrease & amount & data, this & often accompanied by data filtering as part & & data pre-processing steps4. In practice, such activities can take up & 80% & & research effort5. However, we still lack

fundamental insights into how these procedures affect & performance & specific algorithms6. & concreteness, we consider topic modelling7, & paradigmatic unsupervised approach & automatic organization & collections & documents8. One contentious pre-processing step in topic modelling & & removal & semantically uninformative words such as 'the'. & most common approach, which goes back more than 50 years, & & curate & "dictionary & insignificant words"9, commonly referred & as & stopword list10. While some stopword lists can appear & practitioners as standard due & being & default choice in popular applications (such as Mallet11), there & no consensus among experts on which words should be excluded12. Indeed, & use & & 'standard' stopword list & problematic because it ignores & domain-knowledge specificity & stopwords13 & because it & language-specific14. & limitation & static lists have motivated & development & other heuristic approaches based on factors such as & number & occurrences (most & least frequent words), document frequency, & term frequency & inverse document frequency (TFIDF)15, & other, often ill-specified, procedures. & state & uncertainty in & field & illustrated by & fact & & seminal paper on latent Dirichlet allocation "removed & standard list & 50 stop words ... [and] ... words & occurred only once"16, but other works by & same author subsequently removed "standard stop words & those & appear too frequently or too rarely"17 or "all words not in & pruned vocabulary & 4,253 words"18, or chose "1,539 terms & occurred in more than five documents"19 or & "5,000-term vocabulary according & tfidf "20. Even when using & same method, such as TFIDF, different authors use different thresholds; & example, ref. 21 removes words "& have tfidf greater than 0.8". & inconsistency in filtering approaches poses severe challenges & & comparison & results across different studies, rendering it nearly impossible & obtain & coherent picture on & state & & field. This & exacerbated by & fact & & removal & stopwords in topic modelling & text-based unsupervised learning more generally & not well understood22, leading & & sterile debate on & usefulness & such approaches.

We use a darker background to improve visibility and make the colored text more distinct. By applying this change to the previous color-coded text, the colored text will now appear on a darker background, as below:

The use & methods from natural language processing1 has become an indispensable tool in applications & data science pervading nearly every scientific discipline2,3. & main challenge & how & extract meaningful information from large & diverse datasets—most & which & comprised & unstructured texts. One & & most common approaches & represent textual data & & so-called bag-of-words model, in which one ignores & order & words within & given document. & improve & signal-to-noise ratio or decrease & amount & data, this & often accompanied by data filtering as part & & data pre-processing steps4. In practice, such activities can take up & 80% & & research effort5. However, we still lack fundamental insights into how these procedures affect & performance & specific algorithms6. & concreteness, we consider topic modelling7, & paradigmatic unsupervised approach & automatic organization & collections & documents8. One contentious pre-processing step in topic modelling & & removal & semantically uninformative words such as 'the'. & most common approach, which goes back more than 50 years, & & curate & "dictionary & insignificant words"9, commonly referred & as & stopword list10. While some stopword lists can appear & practitioners as standard due & being & default choice in popular applications (such as Mallet11), there & no consensus among experts on which words should be excluded12. Indeed, & use & & 'standard' stopword list & problematic because it ignores & domain-knowledge specificity & stopwords13 & because it & language-specific14. & limitation & static lists have motivated & development & other heuristic approaches based on factors such as & number & occurrences (most & least frequent words), document frequency, & term frequency & inverse document frequency (TFIDF)15, & other, often ill-specified, procedures. & state & uncertainty in & field & illustrated by & fact & & seminal paper on latent Dirichlet allocation "removed & standard list & 50 stop words ... [and] ... words & occurred only once"16, but other works by & same author subsequently removed "standard stop words & those & appear too frequently or too rarely"17 or "all words not in & pruned vocabulary & 4,253 words"18, or chose "1,539 terms & occurred in more than five documents"19 or & "5,000-term vocabulary according & tfidf "20. Even when using & same method, such as TFIDF, different authors use different thresholds; & example, ref. 21 removes words "& have tfidf greater than 0.8". & inconsistency in filtering approaches poses severe challenges & & comparison & results across different studies, rendering it nearly impossible & obtain & coherent picture on & state & & field. This & exacerbated by & fact & & removal & stopwords in topic modelling & text-based unsupervised learning more generally & not well understood22, leading & & sterile debate on & usefulness & such approaches.

Text after using **color-coding** scheme contain **464** words,**106** are the character &.

7. PROS OF COLOR-CODING STOP WORDS

I. Visual Identification: The usage of these symbols is correlated with the color which enables one to identify easily phrases containing the stop words in the text.

II. Quick Assessment: Through scanning the colored texts, one is able to recognize the locations and occurrences of stop words without necessarily reading through the full text.

III. Focus on Content Words: This separation by colors serves well at directing attention to the content words, and, thereby, get a better insight on the message contained in the text.

IV. Educational Tool: Color coding can prove beneficial as a teaching aid to help explain about the stop words and how and why it is utilized in the NLP activities.

V. Selective Analysis: Scholars can compare the colored versions of the text in order to conduct detailed examinations of how stop words affect particular NLP tasks that are undertaken by the scholars.

VI. Customization: It must be pointed that the color-coding scheme can be easily adapted to the concrete requirements for the given research as well as for the further visualization.

8. LIMITATIONS

a. Color Perception: The effectiveness of color-coding therefore depends on the ability of the users to discern and distinguish different colors.

b. Limited Representation: Color coding is a way of presenting concept visually and is not powerful as numbers concerned with stop word frequency.
c. Color Choices: Special attention in the selection of colors should be paid taking into account that some users can have difficulties seeing colors.

9. CONCLUSION

Influence of stop words include the relation of the activity toward completing the task, domain of text and objectives of analysis. However, there are ready-made stop word list available, but these are not very much specific to all the requirements. Some of the stop words may have to be domain specific or the task specific stop word list hence making it necessary to get it right.

Designating the stop words in different color is rather useful and can be considered as one of the most effective innovative solutions. Authors also state that positions of stop words can be quickly determined, for example, by depicting them in a non-standard color. The same color codes also assist in the identification of stop words and their effect on the selected NLP tasks and a selective analysis can be made.

Nevertheless, some issues need to be considered and mentioned, for example, variation of color discernibility among users and possible absence of numerical information obtained from color coding. Some layout advantages of this approach need critical color choice matching and accessibility for all the users.

Thus, stop words remain one of the important components of NLP and their processing remains crucial in order to reach the performance goals in text analysis. The approach of the color-coded stop word lets the reader expand on the stop word usage and distribution adding to the interpretation of the results.

In light of the future progression of NLP techniques and the discovery of new textual data, it is valuable to delimit stop words and their assignments as well as investigate for their function in state-of-art language models and for solutions to multi-lingual text simplifications. Through leveraging both the basic and advanced ideas of stop word removal novel approaches as color coding, it is possible to equip researchers and practitioners with new tools and improve the results of natural language processing in the broad range of domains.

10. RESOURCES

[1] **Bird, S., Klein, E., & Loper, E.** (2009). "Natural Language Processing with Python". O'Reilly Media.

[2] **Blanchard A.** (2007). "Understanding and customizing stop word lists for enhanced patent mapping". World Pat Inf.; 29: 308–316. https://doi.org/10.1016/j.wpi.2007.02.002.

[3] **Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.** (2019). BERT: "Pre-training of Deep Bidirectional Transformers for Language Understanding".

[4] **Garg, N., Schiebinger, D., Jurafsky, D., & Zou, J.** (2019). "Word Embeddings quantify 100 Years of Gender and Ethnic Stereotypes".

[5] **Jurafsky, D., & Martin, J. H.** (2019). "*Speech and Language Processing"* (3rd ed.).

[6] **Kirti, A., Kumar, P., & Choudhury, M.** (2020). "Comparative Analysis of Stop Word Removal Techniques for Text Classification in Bengali".

[7] **Lee, J., & Kang, J.** (2020). "Investigating the Impact of Stop Words on the Performance of Neural Text Generation".

[8] **Luo, X., Zhang, R., Li, L., & Wu, X.** (2020). "Revisiting Stop Word Removal for Generic Text Summarization".

[9] **Manning, C. D., Raghavan, P., & Schütze, H.** (2008). "*Introduction to Information Retrieval"*.

[10] **Pečar, J., Kukar, M., Lavrač, N., & Juvan, P.** (2021)." Stop Words for Text Classification in Slovenian".

[11] **Serhad Sarica, Jianxi Luo**, June 2020, "Stop words in Technical Language Processing".

[12] **Senem Kumova Metin, Bahar Karaoğlan**, (June 2017). "STOP WORD DETECTION AS A BINARY CLASSIFICATION PROBLEM".

[13] **Yadollahi, M., Azzopardi, L., & Crestani, F.** (2021). "Improving Query Translation with the Stop Word Lists".

[14] **Wu, S., Zhang, Y., Yang, F., Li, Z., Wang, Y., & Yu, Y.** (2020). "TextRank: Bringing Order into Texts".