Alternative Topic Modeling Optative Course

Study of LDA Method

Laura Victoria Riera Pérez
Marié del Valle Reyes

Senior year. Computer Science.
School of Math and Computer Science, University of Havana, Cuba

June 26, 2023

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

***Keywords* —**

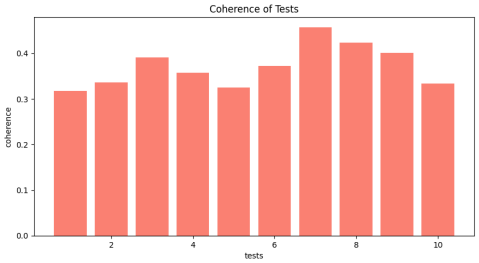
Project's repository

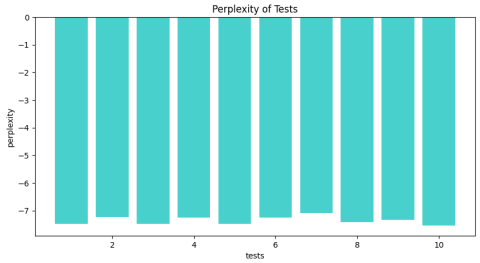
https://github.com/computer-science-crows/study-of-lda-method

I. Initial Analysis

i. Different coherence and perplexity

In the LDA (Latent Dirichlet Allocation) model, coherence and perplexity can vary when running the model multiple times with the same set of words. This is due to the stochastic nature of the algorithm and the various factors that can influence the results.





Perplexity is a measure used in LDA models to evaluate how well the model fits the data. A lower perplexity value indicates a better fit of the model. However, perplexity can vary across different model runs due to random initialization and the way topics are generated and words are assigned. Therefore, you may obtain different perplexity values each time you run the model with the same set of words.

Coherence, on the other hand, is a measure that assesses the coherence of the topics generated by the model. Coherence is based on the relationship between words within each topic and is used to determine how interpretable the topics are. Similar to perplexity, coherence can also vary across different model runs due to randomness and other factors.

The variability in coherence and perplexity can be attributed to various factors, such as random initialization of the model, parameter selection, quality of the training corpus, and the amount of available data. Additionally, different implementations of LDA may have variations in how coherence and perplexity are calculated, which can also contribute to differences in results.

It is important to note that both coherence and perplexity are approximate measures and do not provide a definitive evaluation of the quality of the LDA model. They should be used in conjunction with other evaluation techniques and analysis to gain a more comprehensive understanding of the model results.

In summary, coherence and perplexity can vary when running the LDA model multiple times with the same set of words due to randomness and other factors involved in the algorithm. It is important to consider these variations and use other evaluation techniques to obtain a more complete picture of the model's quality.

ii. Solution

To address the issue of variability in coherence and perplexity in the LDA model when running it multiple times with the same set of words, the following strategies can be considered:

1. Adjust the model parameters: The parameters of the LDA model, such as the number of topics and training iterations, can influence the results. Adjusting these parameters might improve coherence and perplexity. You can experiment with different values and evaluate how they affect the results.

2. Use a fixed random seed: Random initialization of the model can introduce variability in the results. By setting a fixed random seed before each model run, you can ensure consistent initializations and obtain more stable results.

3. Increase the amount of training data: The amount of training data can also impact result stability. If you have a small set of words, variability may be higher. Consider adding more data or expanding the training corpus to achieve more consistent results.

4. Perform averaging across multiple runs: Instead of relying on the results of a single model run, you can perform multiple runs and average the results. This can help reduce variability and obtain a more reliable estimate of coherence and perplexity.

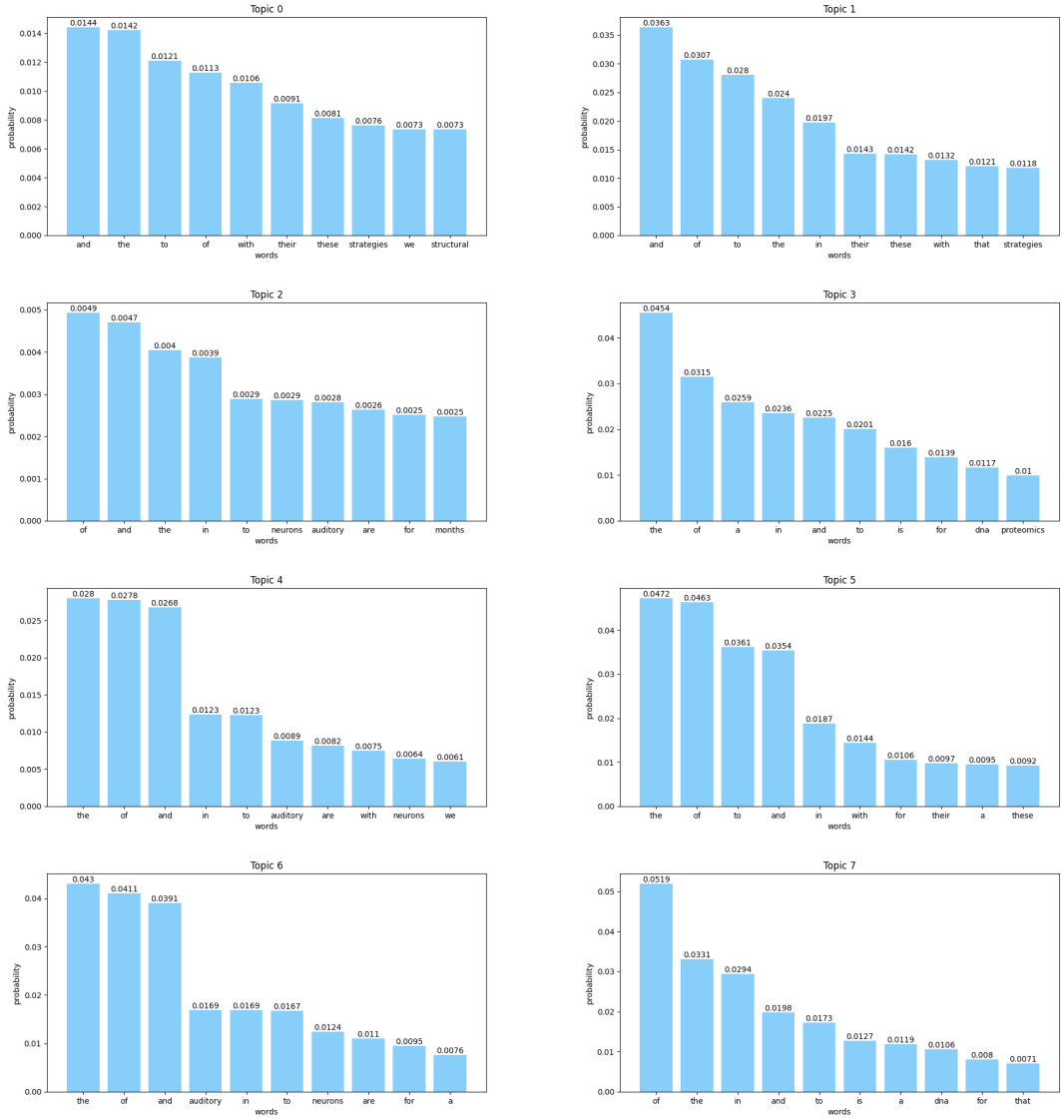
5. Use cross-validation techniques: Cross-validation is a technique that can evaluate the model's performance on different data partitions. By performing cross-validation, you can obtain a more robust measure of coherence and perplexity, as the model is evaluated on different subsets of data.

It's important to note that while these strategies can help reduce variability in coherence and perplexity, there may still be some variation in the results. This is due to the stochastic nature

of the algorithm and the inherent complexity of text data. It is recommended to evaluate the results from multiple runs and use other evaluation techniques to gain a more comprehensive understanding of the LDA model.

iii. Test 8

La mayoria de las palabras son stopwords.



La perplexity es -7.255103257328984, y la coherencia 0.3718197713201331.

II. Removing stopwords

The removal of stopwords is a common step in data preparation for topic modeling with LDA (Latent Dirichlet Allocation). Stopwords are highly common and frequent words in a given

language, such as "the," "and," "of," "to," etc. These words do not contribute much meaning or relevant information for topic identification and can negatively affect the quality of LDA results.

Here are some reasons why stopwords should be removed when performing LDA:

1. Noise reduction: By eliminating stopwords, the noise in the data is reduced. Stopwords are words that are so common that they appear in almost every document and do not provide distinctive information about topics. Removing them reduces the amount of irrelevant words in the analysis and focuses on the most significant words for topic identification.

2. Improved topic interpretability: Removing stopwords enhances the interpretability of topics generated by the LDA model. Stopwords tend to appear in multiple topics and do not help clearly distinguish the themes. By removing them, the most relevant and distinctive keywords of each topic are highlighted, making interpretation and analysis easier.

3. Dimensionality reduction: Removing stopwords reduces the dimensionality of the word space used for topic modeling. This can help improve computational efficiency and reduce memory consumption. By eliminating highly frequent yet uninformative words, a more compact and efficient representation of documents can be achieved.

Stopword removal can be performed using pre-defined lists of stopwords specific to each language. These lists contain common words that are considered stopwords and can be easily found online. For example, for the Spanish language, you can find stopwords lists containing words like "el," "y," "de," etc.

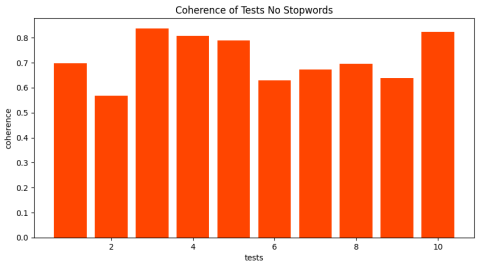
In summary, removing stopwords in the LDA process is important to reduce noise in the data, improve topic interpretation, and reduce the dimensionality of the word space. This helps obtain more accurate and meaningful results in topic modeling with LDA

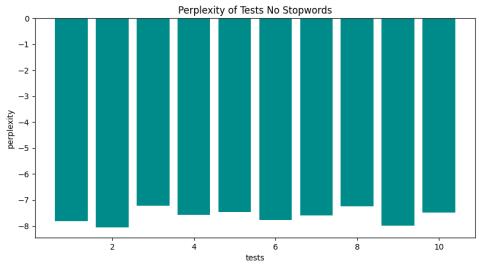
i. Stopwords

Stop words are commonly used words in a language that are often considered insignificant or carry little meaning in the context of natural language processing (NLP) and text mining. These words are typically articles, prepositions, conjunctions, or pronouns. Examples of stop words in English include "a", "the", "is", "are", and so on. Stop words are used to eliminate words that are so commonly used that they may not contribute much to the analysis or understanding of text data.

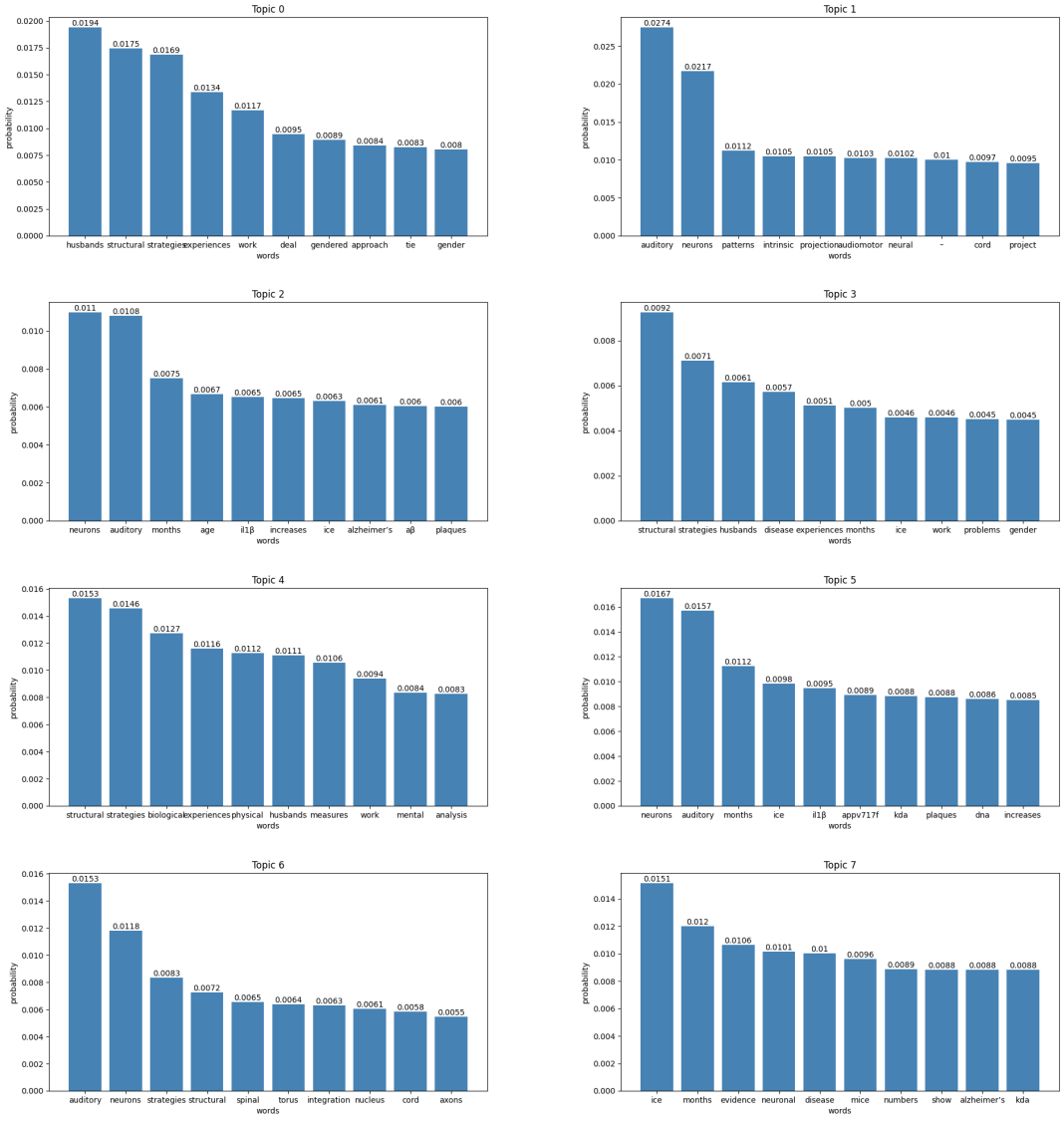
ii. Code Modification

Se creo un nuevo .py en donde se descomentaron las lineas de código que se encargaban de eliminar las stopwords del conjunto de palabras dado en TokenVieuxM.txt.





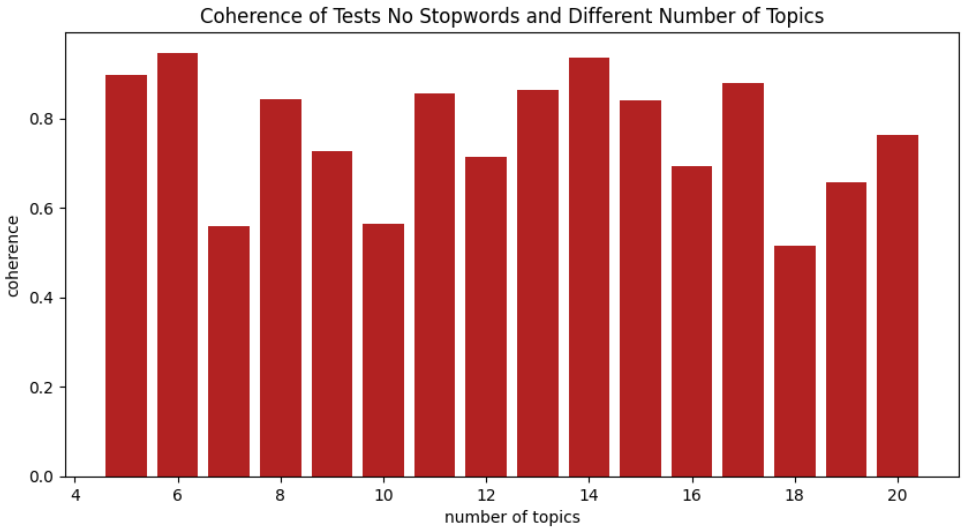
ii.1 Test 8 No Stopwords

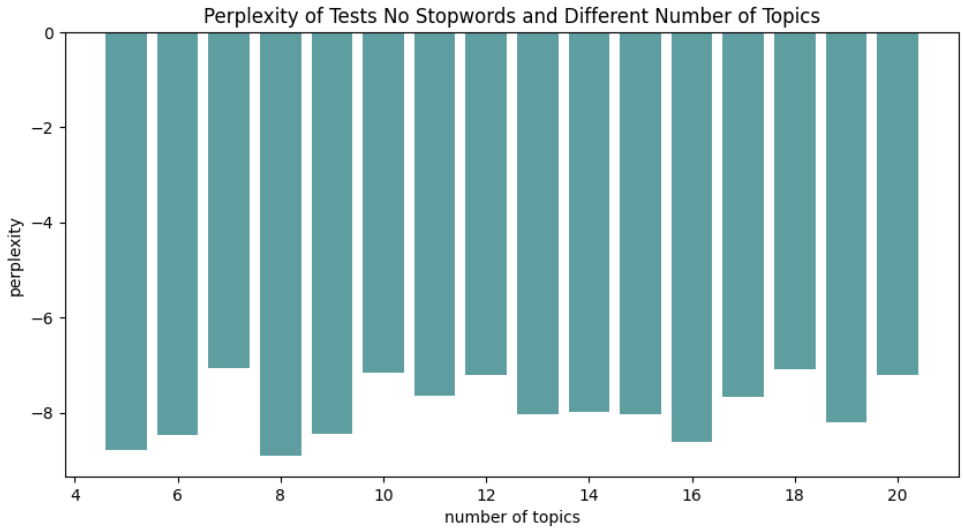


La perplexity es -7.764361092424768 y la coherencia 0.6284575950301475.

hay que comparar los resultados con el test sin remover stopwords y ver la relación entre coherencia y perplexity

iii. Changing number of topics





El número óptimo de tópicos de acurdo a los valores de coherencias debería ser 6.

falta plotear los test con el otro dataset

III. Most typical documents for topics

Para identificar el documento más típico en cada tópico utilizando LDA de Gensim en Python, puedes utilizar el método get_document_topics proporcionado por la clase LdaModel. Este método devuelve una lista de tuplas que contienen el ID del tópico y la probabilidad de ese tópico para cada documento.

References

[1] Cormen, Thomas H. y otros. *Introduction to Algorithms*. The MIT Press. 4ta Edición. Cambridge, Massachusetts. 2022.