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 29, 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* —**

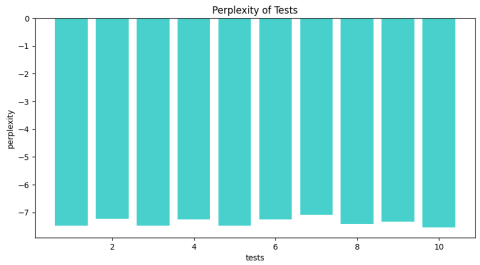
Project's repository

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

I. Initial analysis

The Latent Dirichlet Allocation (LDA) model is a popular machine learning technique used for topic modeling, which allows us to extract abstract topics from a collection of documents. Two metrics that are commonly used to evaluate the quality of a topic model are perplexity and coherence. *Perplexity* is used to evaluate how well the model fits the data. A lower perplexity value indicates a better fit of the model. *Coherence*, on the other hand, is a measure that assesses the coherence of the topics generated by the model. It is based on the relationship between words within each topic and is used to determine how interpretable the topics are.

The program was executed 10 times, and the following coherence and perplexity values were obtained:



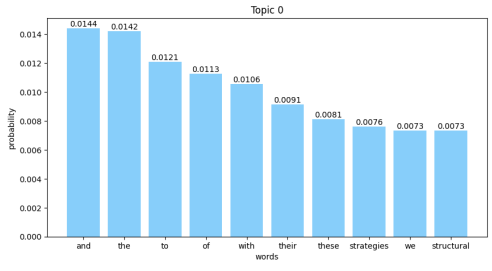
The first thing we can notice is that different coherence and perplexity values are obtained in each run, even though we are working with the same corpus and number of topics.

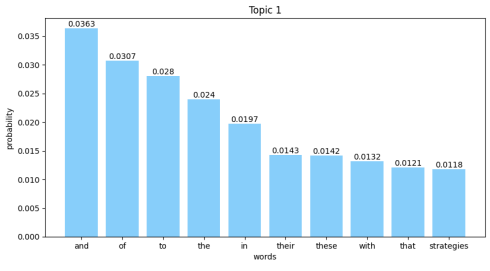
In the LDA (*Latent Dirichlet Allocation*) model, coherence and perplexity can vary when running the model multiple times with the same set of words. This can be attributed to various factors, such as random initialization of the model, parameter selection, quality of the training corpus, and the amount 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. To address the issue of variability in coherence and perplexity in the LDA model of gensim when running it multiple times with the same set of words we can use a fixed random seed ensuring consistent initializations and obtaining more stable results.

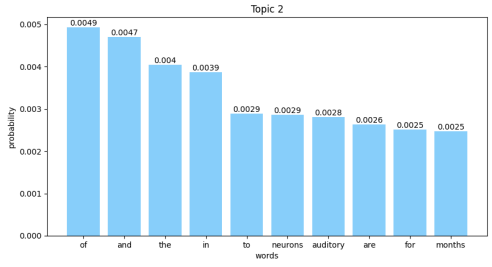
The average values of perplexity and coherence across tests is approximately -7.62 and 0.72.

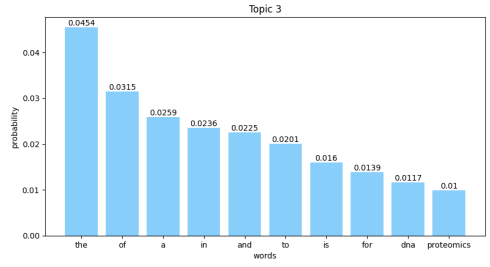
i. Topic descriptions obtained with one specific launch

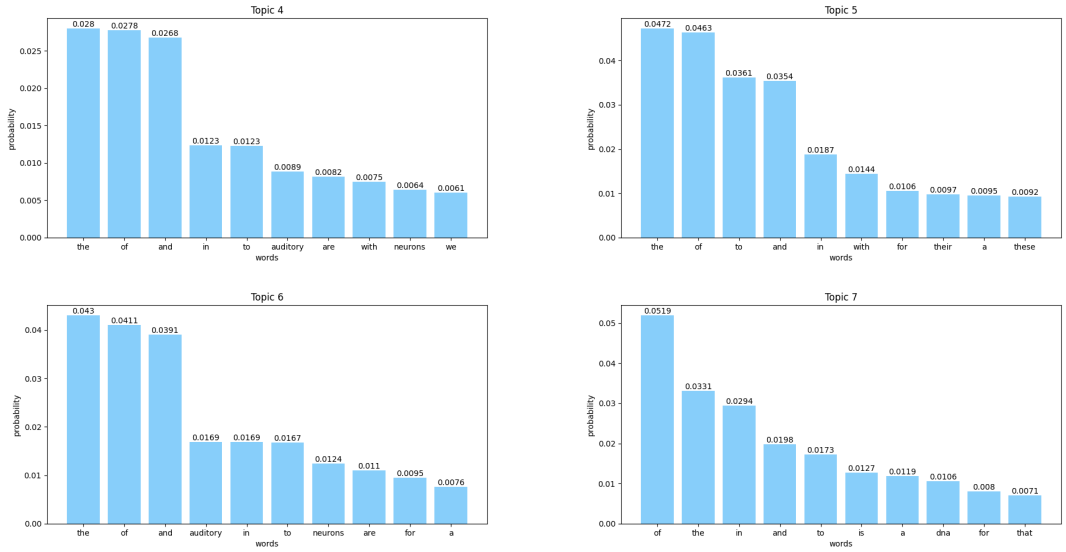
To further analyze the performance of the model let's look at the topic description obtained for one specific launch, in this case, test 8:











La perplexity es -7.255103257328984, y la coherencia 0.3718197713201331.

Poner se tomo el test 8 como ejemplo, y se puede observar que la mayoría de las palabras son stopwords, y analizar los valores de coherencia y perplexity.

II. Stopwords

Stopwords 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 stopwords in English include "a", "the", "is", "are", and so on. Stopwords are used to eliminate words that are so commonly used that they may not contribute much to the analysis or understanding of text data.

Explicación de stopwords

i. 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.

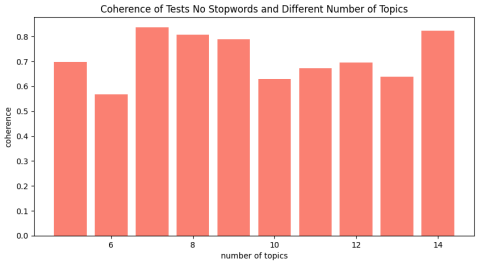
2.a Poner que la mayoría de las palabras son stopwords y buscar por qué esto sucede

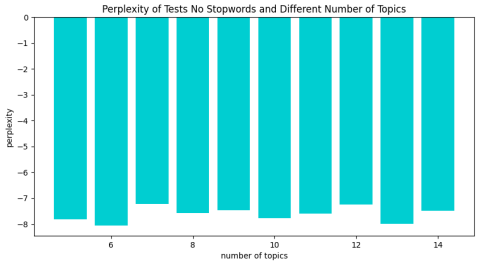
Stopword removal can be performed using pre-defined lists of stopwords specific to each language. These lists contain common words that are considered stopwords and can be easily found online. For example, for the Spanish language, you can find stopwords lists containing words like "el," "y," "de," etc.

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

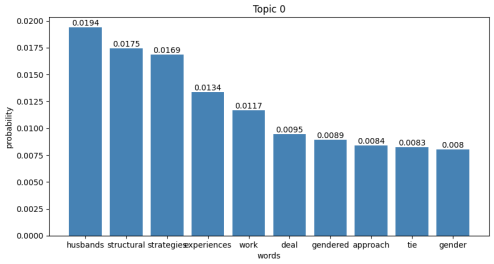
ii. Code Modification

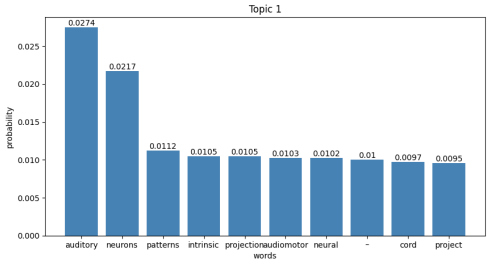
Se creo un nuevo .py en donde se descomentaron las lineas de código que se encargaban de eliminar las stopwords del conjunto de palabras dado en TokenVieuxM.txt.

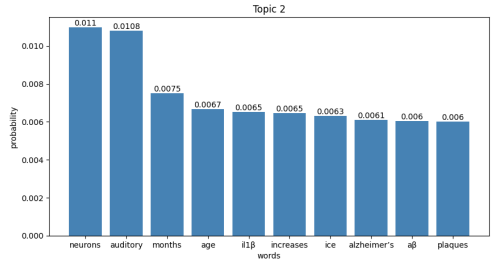


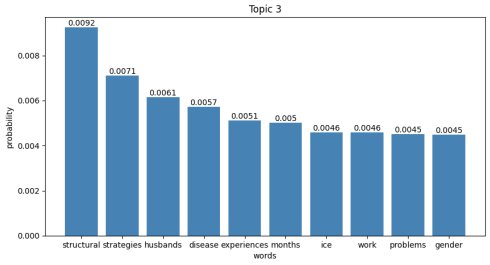


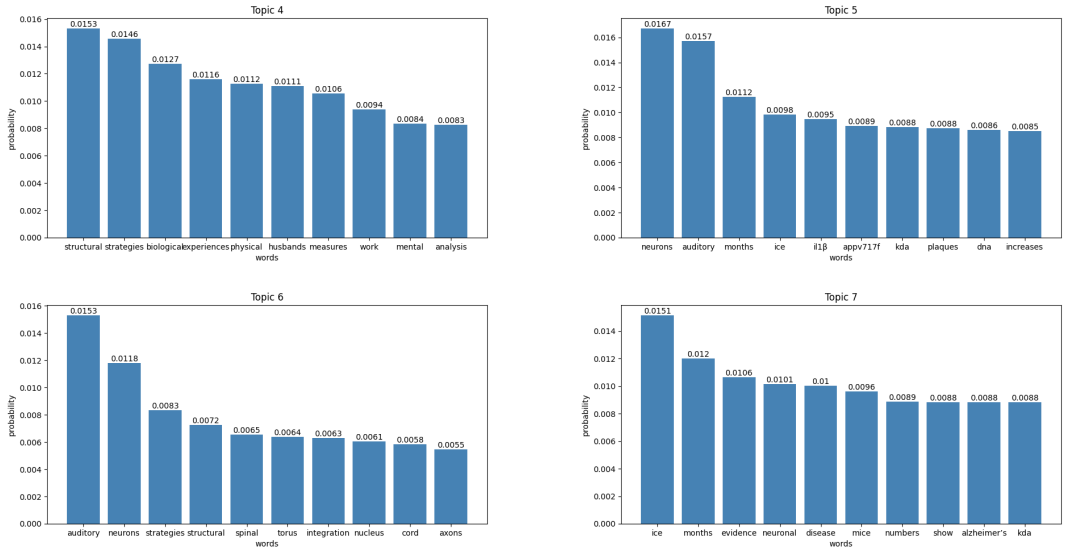
ii.1 Test 8 No Stopwords





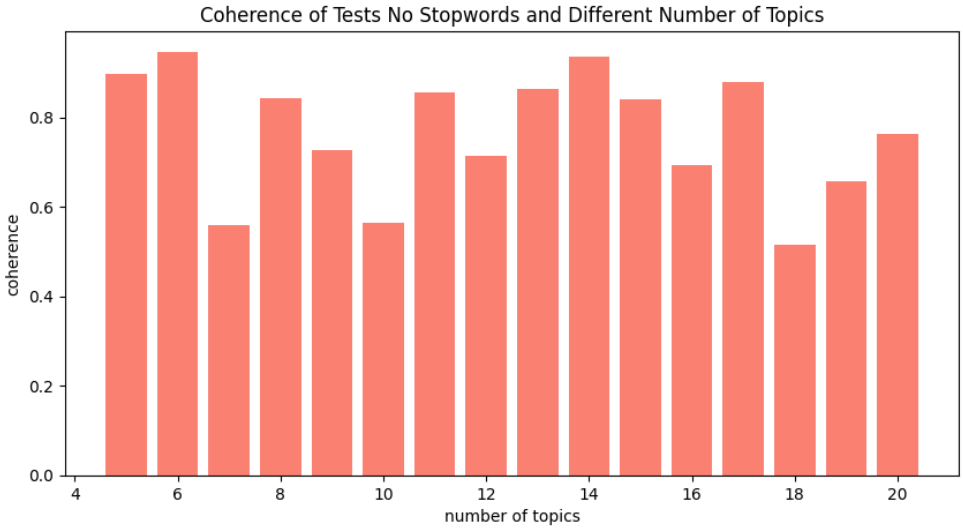






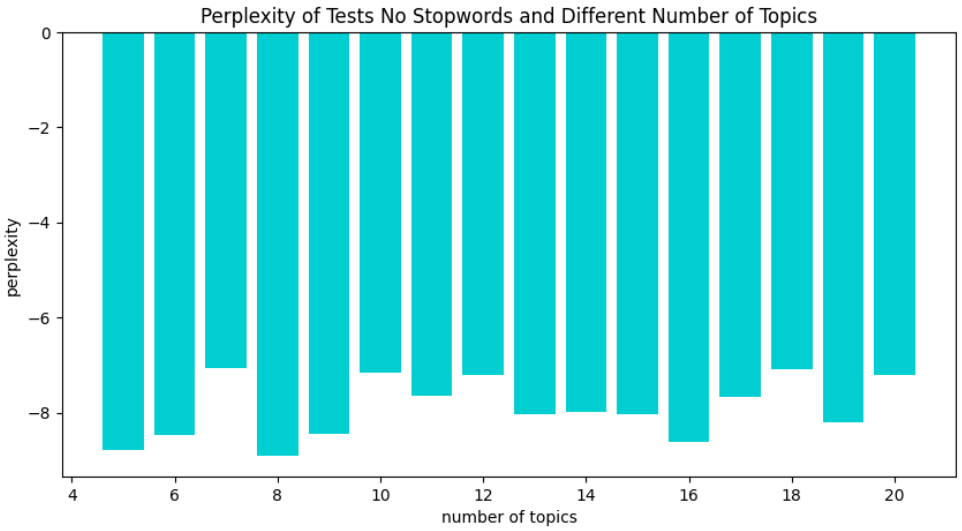
La perplexity es -7.764361092424768 y la coherencia 0.6284575950301475.

iii. Changing number of topics



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

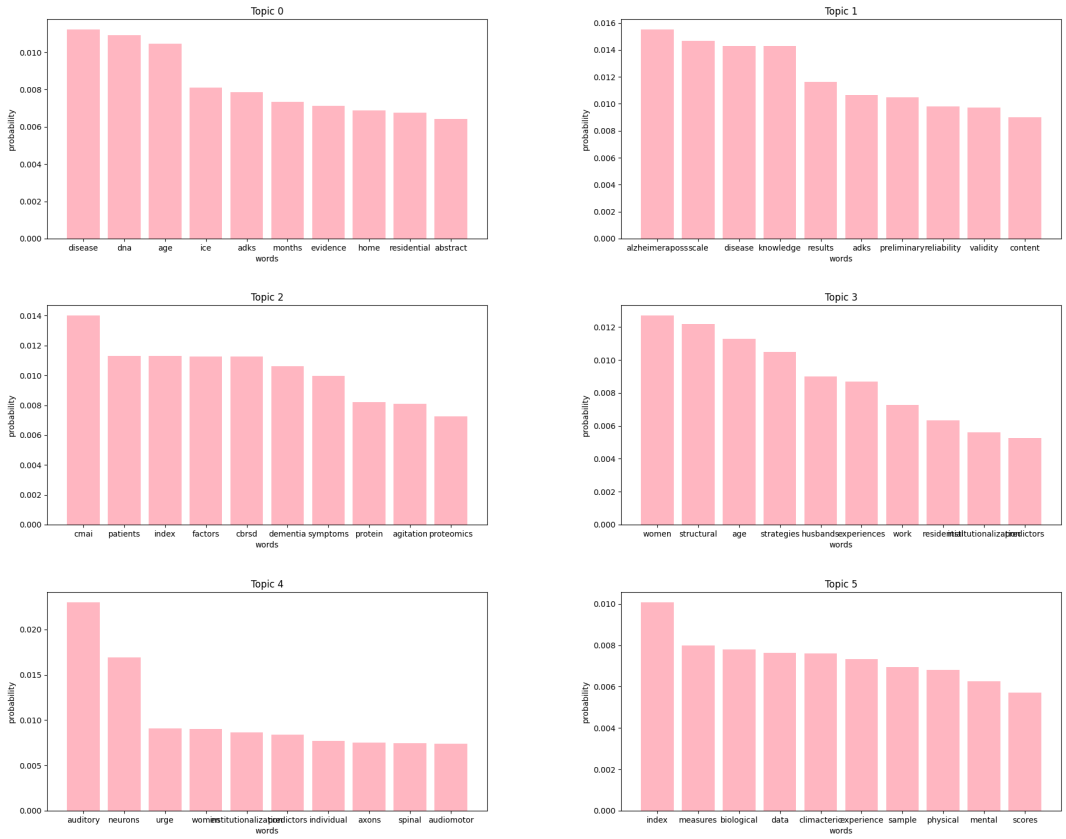
2.d Escribir cual sería el número óptimo de tópicos y por qué

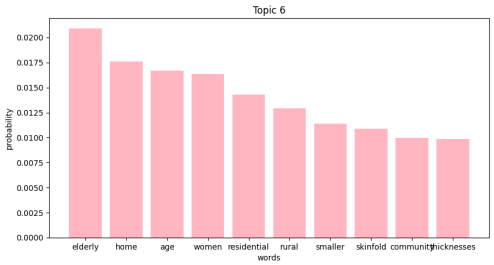


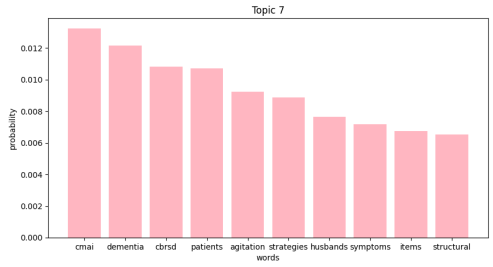
El número óptimo de tópicos de acurdo a los valores de coherencias debería ser 6.

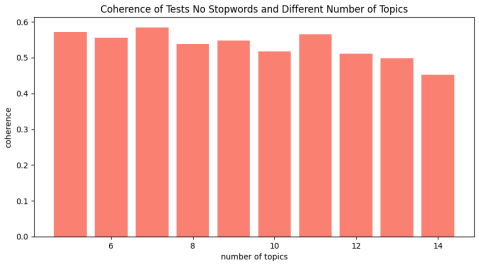
III. Dataset 2

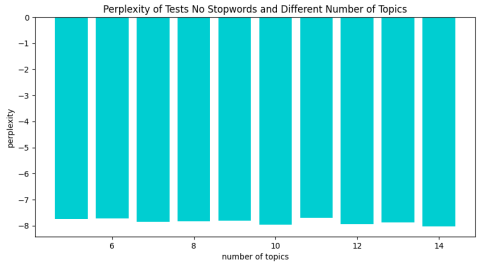
2.e Comparar los resultados de este dataset con el ii.1 y llegar a conclusiones.





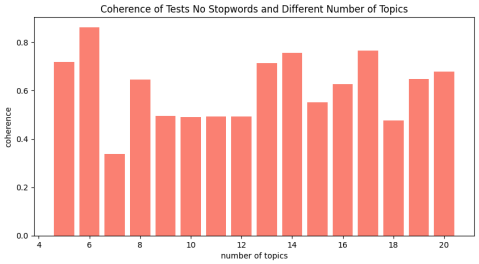


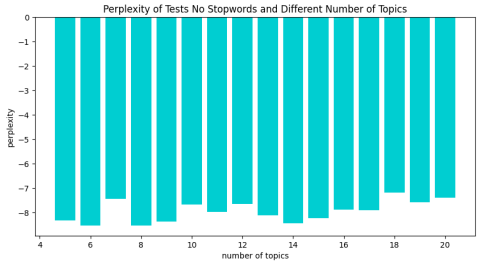




i. Different number of topics

2.f Analizar cuál es el número de tópicos óptimo y justificar





IV. Most typical documents for topics

Ver como utilizar el método

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.

V. Code modifications, tests and plots

The Python program used for this analysis was divided into three main functions: ‘read()‘, ‘parse()‘, and ‘lda()‘.

The ‘read()‘ function is responsible for reading the indicated dataset. The ‘parse()‘ function processes the read dataset, removing unnecessary symbols and preparing the data for the LDA model. The ‘lda()‘ function initializes an LDA model from the Gensim module, to which the parsed dataset and the number of topics to be generated are passed. This function also calculates the coherence and perplexity of the model.

References

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