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 —

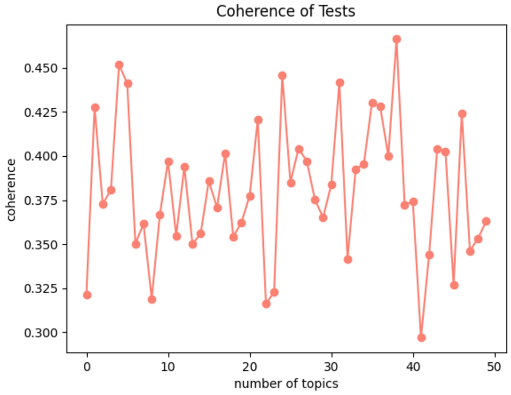
Project's repository

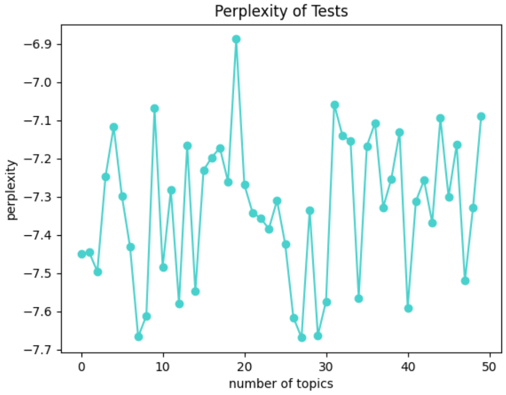
https://github.com/computer-science-crows/study-of-lda-method

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 50 times, and the following coherence and perplexity values were obtained:





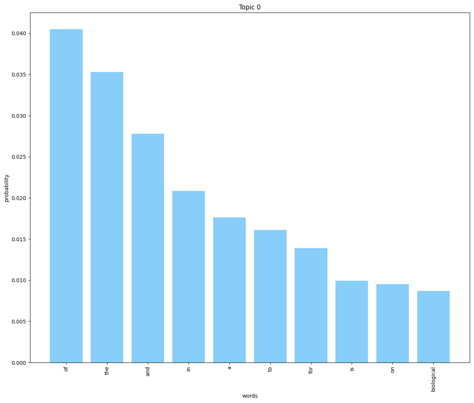
The first thing we can notice is that different coherence and perplexity values are obtained in each run, even though we are working with the same corpus and number of topics.

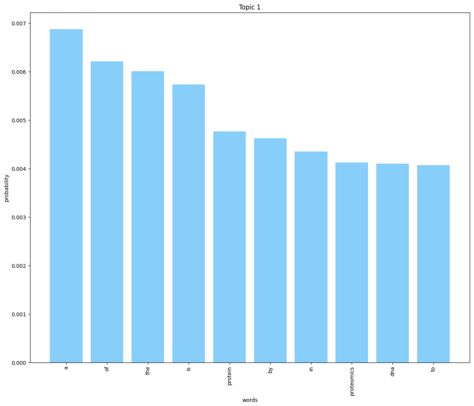
In the LDA (*Latent Dirichlet Allocation*) model, coherence and perplexity can vary when running the model multiple times with the same set of words. This can be attributed to various factors, such as random initialization of the model, parameter selection, quality of the training corpus, and the amount 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.

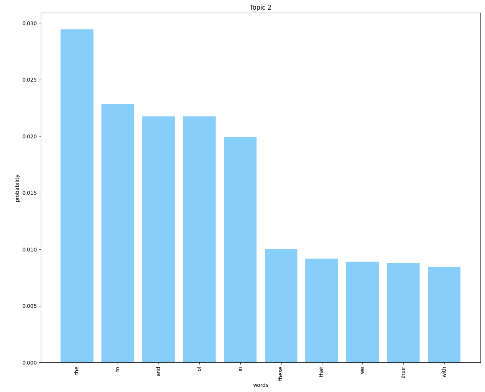
The average values of perplexity and coherence across tests is approximately -7.33 and 0.38.

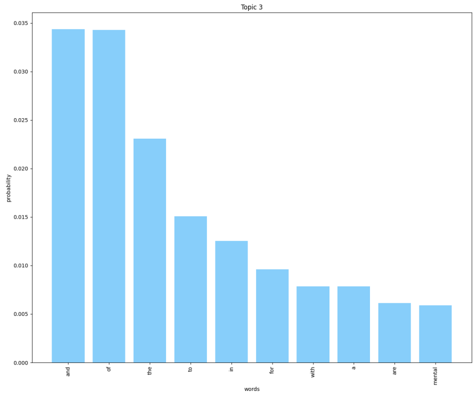
i. Topic descriptions obtained with one specific launch

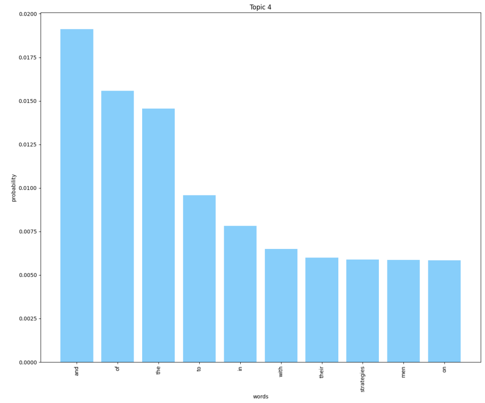
To further analyze the performance of the model let's look at the topic description obtained with one specific launch, in this case, test 5:

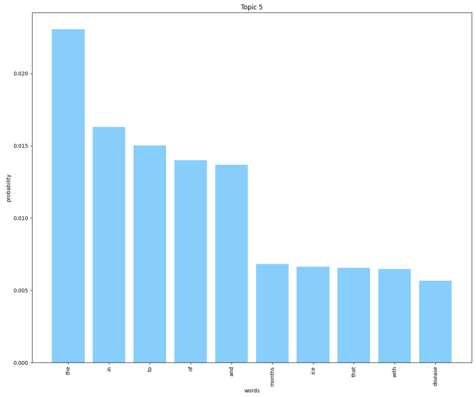


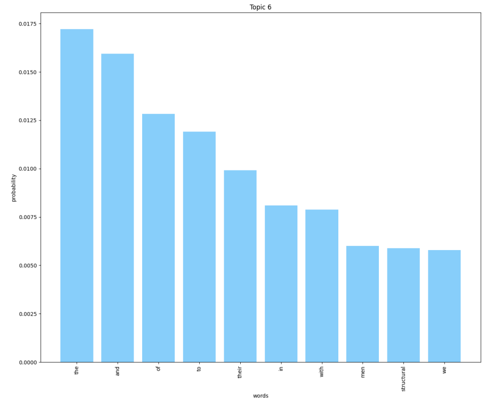


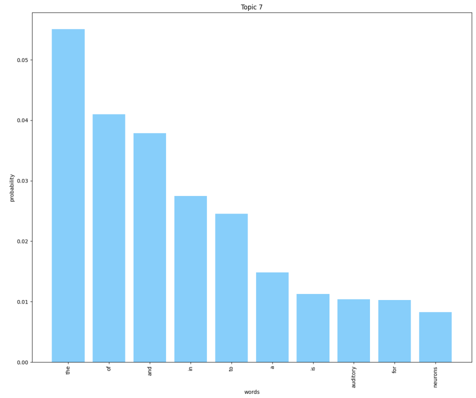












A significant issue was observed. The majority of the words that emerged in the topic descriptions were identified as stopwords.

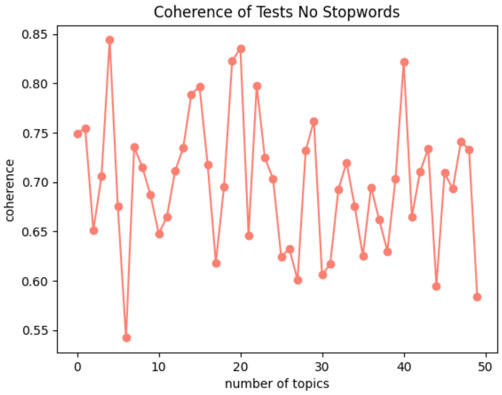
Stopwords are commonly used words in a language that are usually filtered out before or after processing of natural language data. Typically they consist in articles, prepositions, conjunctions, or pronouns. Examples of stopwords in English include "a", "the", "is", "are", and so on. The presence of these stopwords in the topic descriptions is problematic because they appear in almost every document. They are generally uninformative in the context of an LDA model, as they do not contribute to the theme or subject of a topic. The high prevalence of stopwords in the topic descriptions makes it challenging to interpret the topics and understand the underlying themes in the dataset.

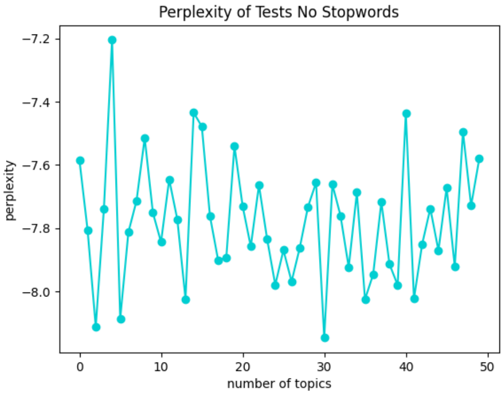
In terms of the model's performance metrics, the coherence value is relatively low, 0.40, which is consistent with the observation that the topics are difficult to interpret due to the high presence of stopwords. The perplexity value was -7.33 which does not provide a clear indication of the model's performance.

II. Removing stopwords

The removal of stopwords in the Latent Dirichlet Allocation (LDA) model indeed plays a crucial role in enhancing the quality of the topics generated. By eliminating these common words, which typically do not contribute to the meaning or theme of a topic, the model can focus on the most significant words for topic identification. This not only improves the interpretability of the topics but also reduces the dimensionality of the word space used, making the model more efficient.

A new Python script was created in which the lines of code responsible for removing stopwords from the given set of words in TokenVieuxM.txt were uncommented. This modification is expected to improve the performance of the model.

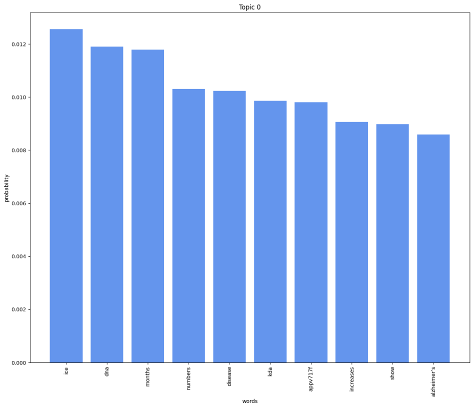


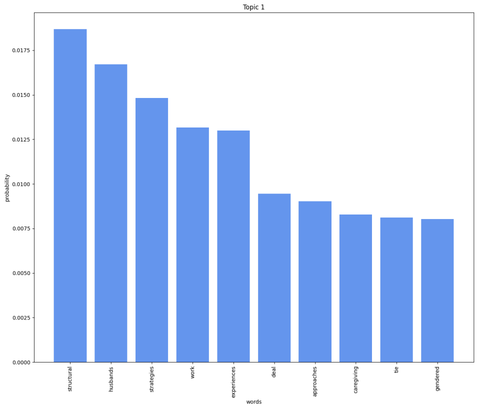


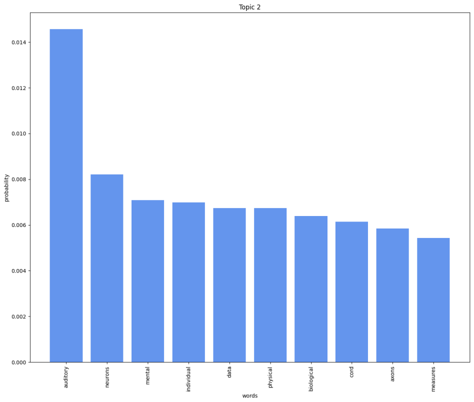
The average values of perplexity and coherence across tests is approximately -7.78 and 0.70.

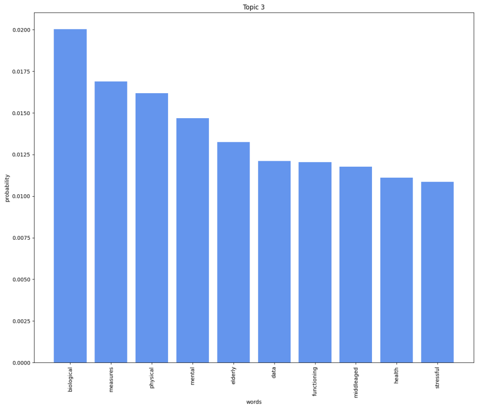
i. Topic descriptions obtained with one specific launch and no stopwords

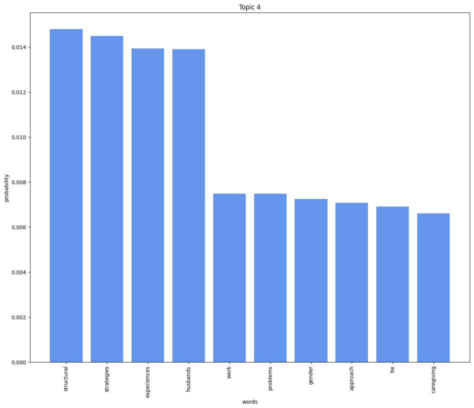
Once again, let's look at the topic description obtained with one specific launch of test 5:

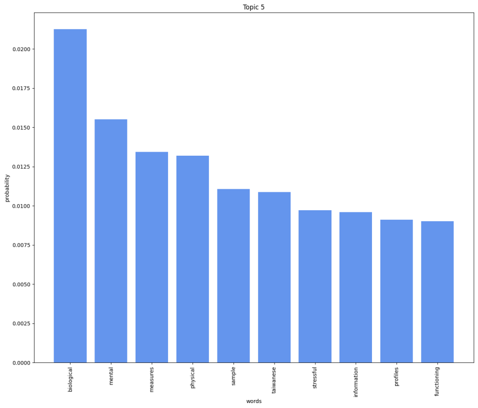

Topic 0
probability
0.000
0.002
0.004
0.006
0.008
0.010
0.012
ice
dna
months
numbers
disease
kda
appv717f
increases
show
alzheimer's
words

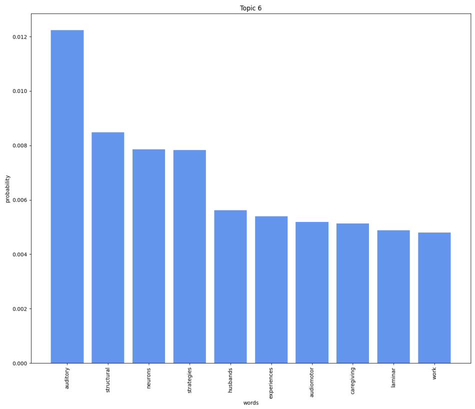

Topic 1
probability
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
structural
husbands
strategies
work
experiences
deal
approaches
caregiving
tie
gendered
words

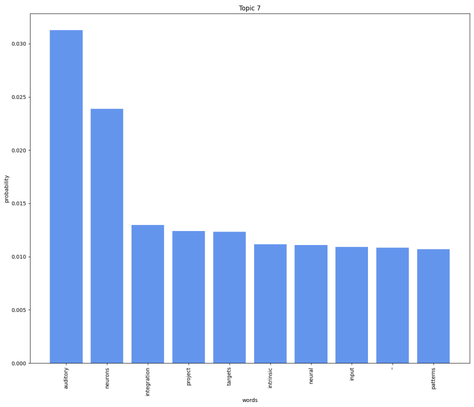

Topic 2
probability
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
auditory
neurons
mental
individual
data
physical
biological
cord
axons
measures
words


Topic 3
probability
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
0.0200
biological
measures
physical
mental
elderly
data
functioning
middleaged
health
stressful
words


Topic 4
probability
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
structural
strategies
experiences
husbands
work
problems
gender
approach
tie
caregiving
words


Topic 5
probability
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
0.0200
biological
mental
measures
physical
sample
taiwanese
stressful
information
profiles
functioning
words

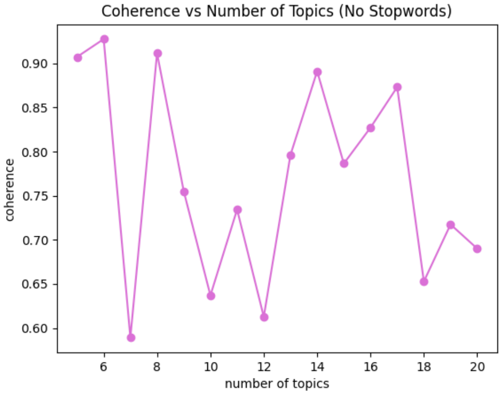


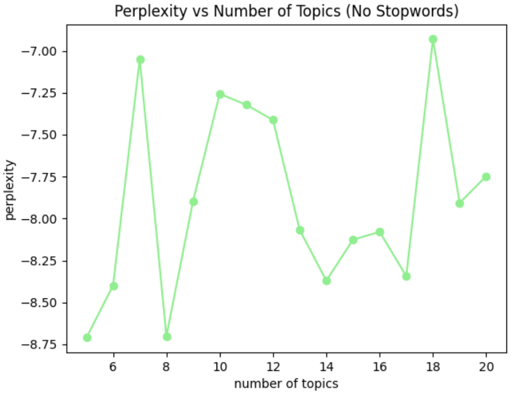


In this case a slightly lower perplexity was obtained, -7.72 while the coherence increased to 0.66. These changes suggest that the removal of stopwords from the dataset improved the performance of the LDA model, both in terms of its predictive ability (as indicated by the decrease in perplexity) and the interpretability of its topics (as indicated by the increase in coherence). This underscores the importance of preprocessing steps in natural language processing tasks.

III. Optimal number of topics for the document collection TokenVieuxM.txt

Choosing the optimal number of topics in an LDA model is a crucial step, but there isn't a definitive rule for this as it largely depends on the specific dataset and the context of the problem.





In the process of determining the optimal number of topics for the LDA model, it is generally accepted that a model with lower perplexity and higher coherence is more desirable. A technique often used is the elbow method, as in K-Means. However, in the absence of a discernible elbow in the plot of these metrics, a different approach is required.

We propose to make this decision by computing the difference between the coherence and perplexity values for each model. This difference will serve as a composite metric that balances both the predictive accuracy (as indicated by perplexity) and interpretability (as indicated by coherence) of the model.

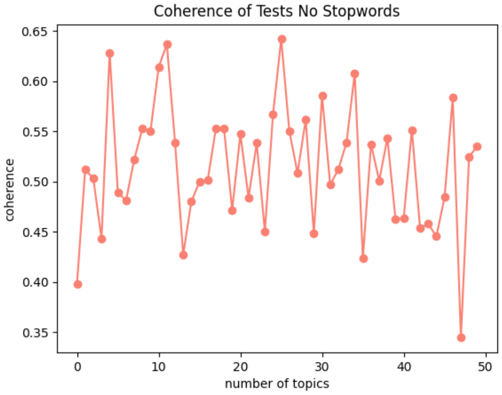
Among the models evaluated, we will select the one with the highest value of this composite metric. However, in cases where multiple models have similar composite scores, we will prioritize the model with the higher coherence value. This is due to our specific need for easily interpretable topics, which is better facilitated by a higher coherence score.

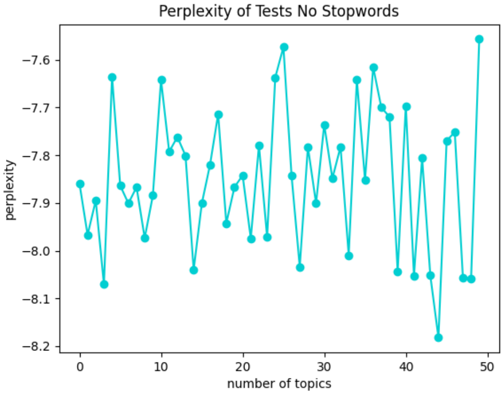
This approach ensures that we select a model that not only fits the data well but also generates topics that are semantically meaningful and interpretable.

According with this, the optimal number of topics is 5.

IV. Changing the dataset

Let us now analyze the behavior of the model for the document collection TokenVieuxN.txt. The program was also executed 50 times, and the following coherence and perplexity values were obtained:

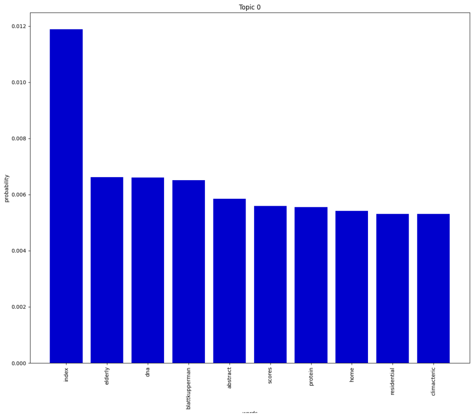


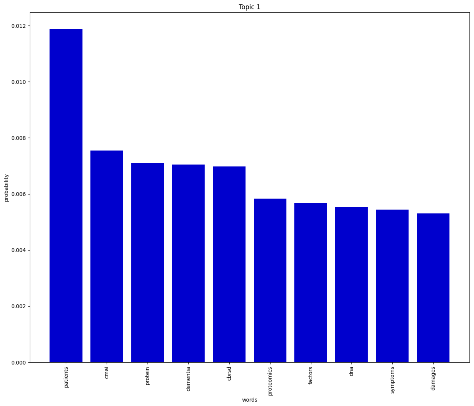


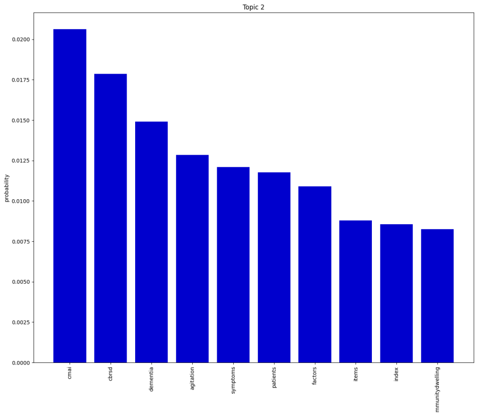
The average values of perplexity and coherence across tests is approximately -7.84 and 0.51.

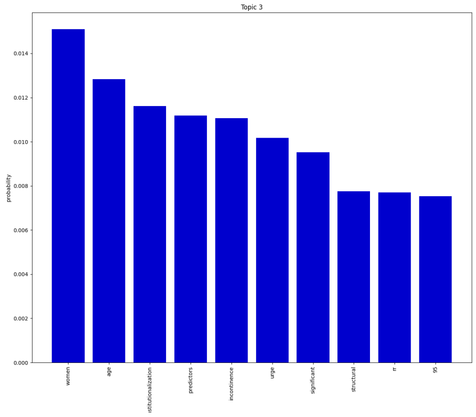
i. Topic descriptions obtained with one specific launch and no stopwords

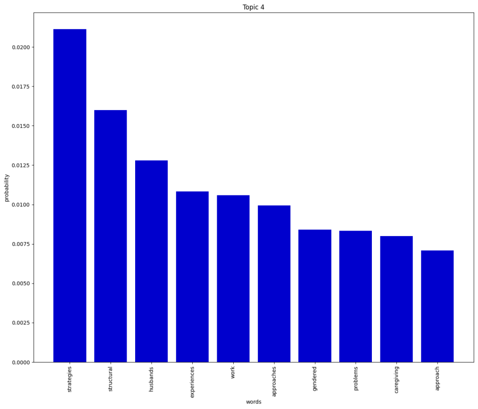
Let's examine the topic description derived from a specific execution of test 5 on this dataset:

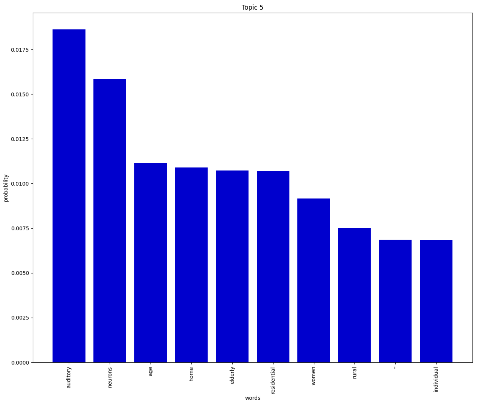


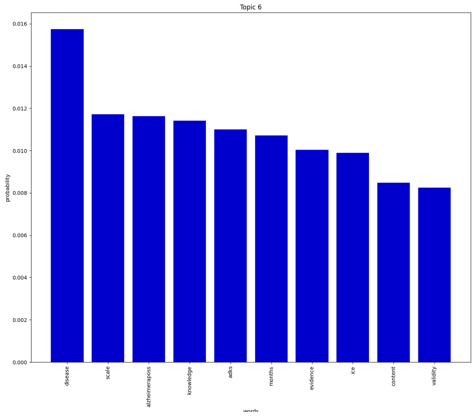


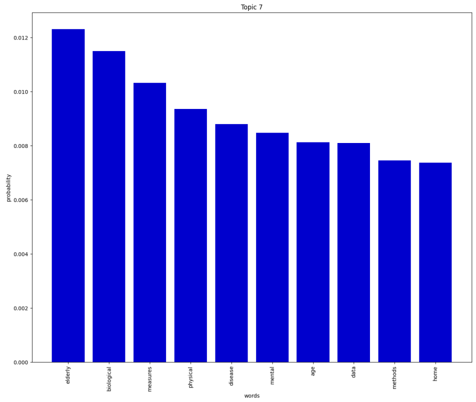








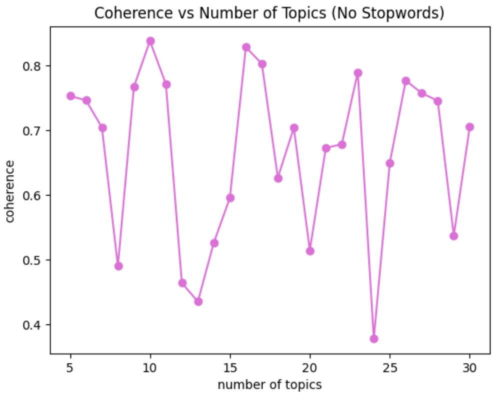


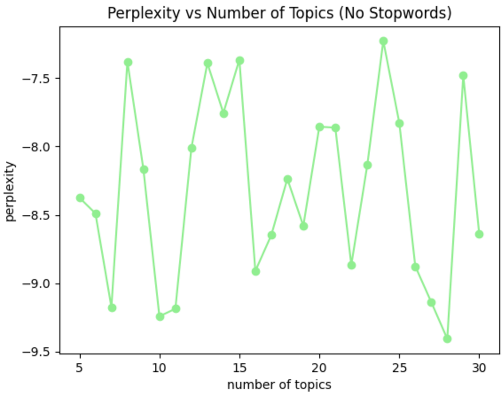


We observe a coherence value of 0.50 and a perplexity value of -7.70. When juxtaposed with the performance of the model without stopwords on the previous dataset, it's possible to notice that the perplexity slightly increased and the coherence decreased. This comparative analysis suggests that the Latent Dirichlet Allocation (LDA) model, configured with 10 topics, provides a better fit for the data derived from the 'TokenVieuxM' dataset.

ii. Optimal number of topics for the document collection TokenVieuxN.txt

As we can observe in the following plots associated with the current dataset, we find an analogous situation to the one encountered with the previous dataset: there is no discernible elbow. We will employ the same approach as previously outlined and use the composite metric to find the optimal number of topics for the LDA model.





The optimal number of topics is 23.

V. 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.

Ver como utilizar el método

Document	Topic ID	Topic Probability
1	4	0.98846
2	3	0.99099
3	1	0.98928
4	9	0.99262
5	8	0.99302

Figure 1: *Dataset 1*

Document	**Topic ID**	**Topic Probability**
1	9	0.98846
2	5	0.99099
3	9	0.98928
4	7	0.99262
5	6	0.99302
6	4	0.99451
7	4	0.34179
8	7	0.98448
9	3	0.98815
10	1	0.98915

Figure 2: *Dataset 2*

VI. 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.

Actualizar referencias