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* —**

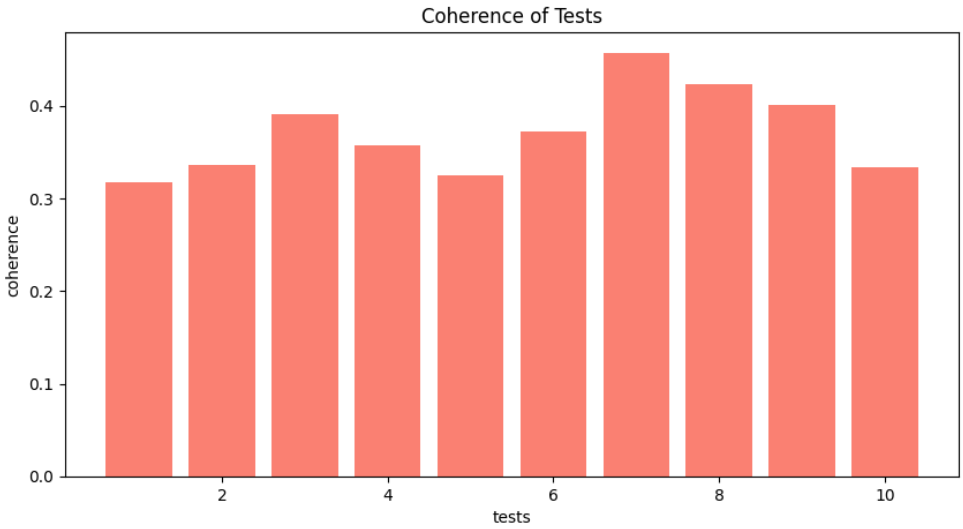
Project's repository

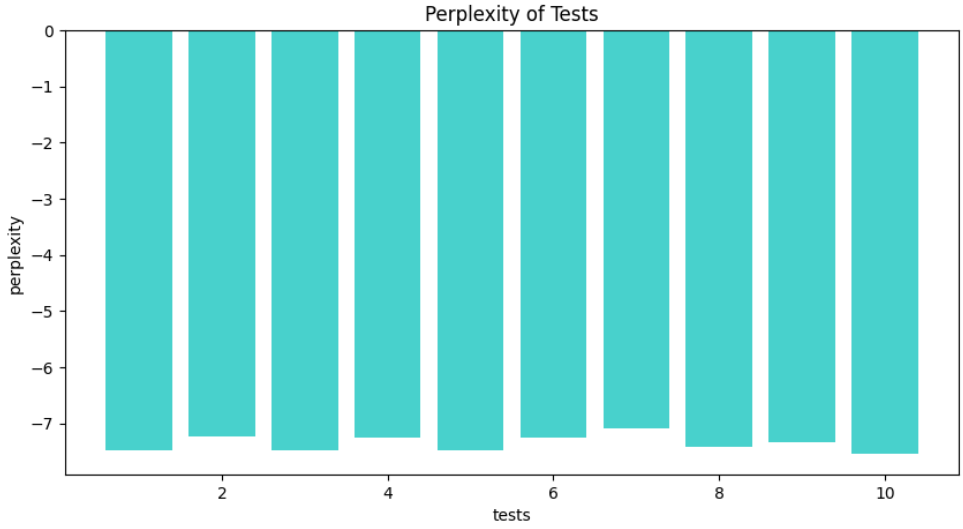
https://github.com/computer-science-crows/study-of-lda-method

I. Initial Analysis

i. Different coherence and perplexity

En el modelo de LDA (Latent Dirichlet Allocation), la coherencia y la perplexity pueden variar cuando se ejecuta el modelo varias veces con el mismo conjunto de palabras. Esto se debe a la naturaleza estocástica del algoritmo y a los diferentes factores que pueden influir en los resultados.





La perplexity es una medida utilizada en los modelos de LDA para evaluar qué tan bueno es el modelo en términos de ajuste a los datos. Cuanto más bajo sea el valor de la perplexity, mejor será el ajuste del modelo. Sin embargo, la perplexity puede variar entre diferentes ejecuciones del modelo debido a la inicialización aleatoria y a la forma en que se generan los temas y se asignan las palabras. Por lo tanto, es posible que obtengas diferentes valores de perplexity cada vez que ejecutes el modelo con el mismo conjunto de palabras.

La coherencia, por otro lado, es una medida que evalúa cuánta coherencia tienen los temas generados por el modelo. La coherencia se basa en la relación entre las palabras dentro de cada tema y se utiliza para determinar qué tan interpretables son los temas. Al igual que con la perplexity, la coherencia también puede variar en diferentes ejecuciones del modelo debido a la aleatoriedad y a otros factores.

La variabilidad en la coherencia y la perplexity puede deberse a varios factores, como la inicialización aleatoria del modelo, la selección de parámetros, la calidad del corpus de entrenamiento y

la cantidad de datos disponibles. Además, diferentes implementaciones de LDA pueden tener variaciones en la forma en que se calcula la coherencia y la perplexity, lo que también puede contribuir a las diferencias en los resultados.

Es importante tener en cuenta que tanto la coherencia como la perplexity son medidas aproximadas y no proporcionan una evaluación definitiva de la calidad del modelo de LDA. Se deben utilizar en conjunto con otras técnicas de evaluación y análisis para obtener una comprensión más completa de los resultados del modelo.

En resumen, la coherencia y la perplexity pueden variar cuando se ejecuta el modelo de LDA varias veces con el mismo conjunto de palabras debido a la aleatoriedad y a otros factores involucrados en el algoritmo. Es importante considerar estas variaciones y utilizar otras técnicas de evaluación para obtener una imagen más completa de la calidad del modelo

ii. Solution

Para solucionar el problema de la variabilidad en la coherencia y perplexity en el modelo LDA al ejecutarlo múltiples veces con el mismo conjunto de palabras, se pueden considerar las siguientes estrategias:

1. Ajustar los parámetros del modelo: Los parámetros del modelo LDA, como el número de temas y las iteraciones de entrenamiento, pueden influir en los resultados. Es posible que ajustar estos parámetros pueda mejorar la coherencia y perplexity. Puedes experimentar con diferentes valores y evaluar cómo afectan los resultados.

2. Utilizar una semilla aleatoria fija: La inicialización aleatoria del modelo puede ser una fuente de variabilidad en los resultados. Al establecer una semilla aleatoria fija antes de cada ejecución del modelo, puedes asegurarte de que las inicializaciones sean consistentes y obtener resultados más estables.

3. Aumentar la cantidad de datos de entrenamiento: La cantidad de datos de entrenamiento también puede afectar la estabilidad de los resultados. Si tienes un conjunto de palabras pequeño, es posible que la variabilidad sea mayor. Considera agregar más datos o ampliar el corpus de entrenamiento para obtener resultados más consistentes.

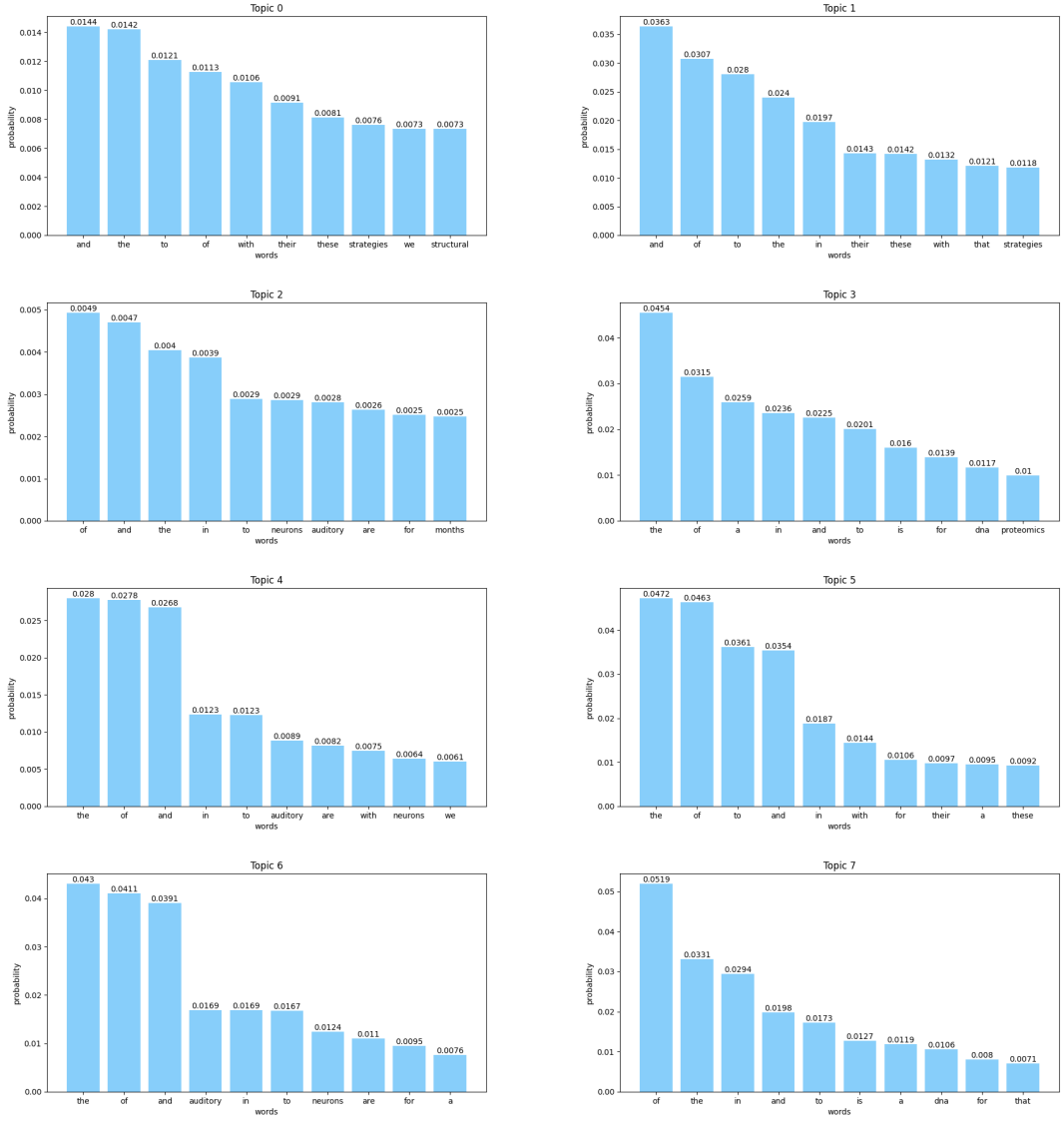
4. Realizar promedios de múltiples ejecuciones: En lugar de confiar en los resultados de una sola ejecución del modelo, puedes realizar múltiples ejecuciones y promediar los resultados. Esto puede ayudar a reducir la variabilidad y obtener una estimación más confiable de la coherencia y perplexity.

5. Utilizar técnicas de validación cruzada: La validación cruzada es una técnica que puede ayudar a evaluar el rendimiento del modelo en diferentes particiones del conjunto de datos. Al realizar validación cruzada, puedes obtener una medida más robusta de la coherencia y perplexity, ya que se evalúa el modelo en diferentes subconjuntos de datos.

Es importante tener en cuenta que, aunque estas estrategias pueden ayudar a reducir la variabilidad en la coherencia y perplexity, es posible que aún exista cierta variación en los resultados. Esto se debe a la naturaleza estocástica del algoritmo y a la complejidad inherente de los datos de texto. Es recomendable evaluar los resultados de múltiples ejecuciones y utilizar otras técnicas de evaluación para obtener una comprensión más completa del modelo LDA.

iii. Test 8

La mayoria de las palabras son stopwords.



La perplexity es -7.255103257328984, y la coherencia 0.3718197713201331.

II. Removing stopwords

La eliminación de las stopwords (palabras vacías) es un paso común en la preparación de datos para el modelado de tópicos con LDA. Las stopwords son palabras muy comunes y frecuentes en un idioma determinado, como "el", "y", "de", "a", etc. Estas palabras no aportan mucho significado o información relevante para la identificación de temas y pueden afectar negativamente la calidad de los resultados de LDA.

Aquí hay algunas razones por las que se deben remover las stopwords al realizar LDA:

1. Reducción del ruido: Al eliminar las stopwords, se reduce el ruido en los datos. Las stopwords son palabras tan comunes que aparecen en casi todos los documentos y no proporcionan información distintiva sobre los temas. Al eliminarlas, se reduce la cantidad de palabras irrelevantes en el análisis y se enfoca en las palabras más significativas para la identificación de tópicos.

2. Mejor interpretación de tópicos: Al eliminar las stopwords, se mejora la interpretabilidad de los tópicos generados por el modelo de LDA. Las stopwords tienden a aparecer en múltiples tópicos y no ayudan a distinguir claramente los temas. Al eliminarlas, se destacan las palabras clave más relevantes y distintivas de cada tópico, lo que facilita su interpretación y análisis.

3. Reducción de la dimensionalidad: Al eliminar las stopwords, se reduce la dimensionalidad del espacio de palabras utilizado para el modelado de tópicos. Esto puede ayudar a mejorar la eficiencia computacional y reducir el consumo de memoria. Al eliminar palabras muy frecuentes pero poco informativas, se puede lograr una representación más compacta y eficiente de los documentos.

La eliminación de las stopwords se puede realizar utilizando listas predefinidas de stopwords específicas de cada idioma. Estas listas contienen palabras comunes que se consideran stopwords y se pueden encontrar fácilmente en línea. Por ejemplo, para el idioma español, puedes encontrar listas de stopwords que contienen palabras como "el", "y", "de", etc.

En resumen, la eliminación de stopwords en el proceso de LDA es importante para reducir el ruido en los datos, mejorar la interpretación de los tópicos y reducir la dimensionalidad del espacio de palabras. Esto ayuda a obtener resultados más precisos y significativos en el modelado de tópicos con LDA

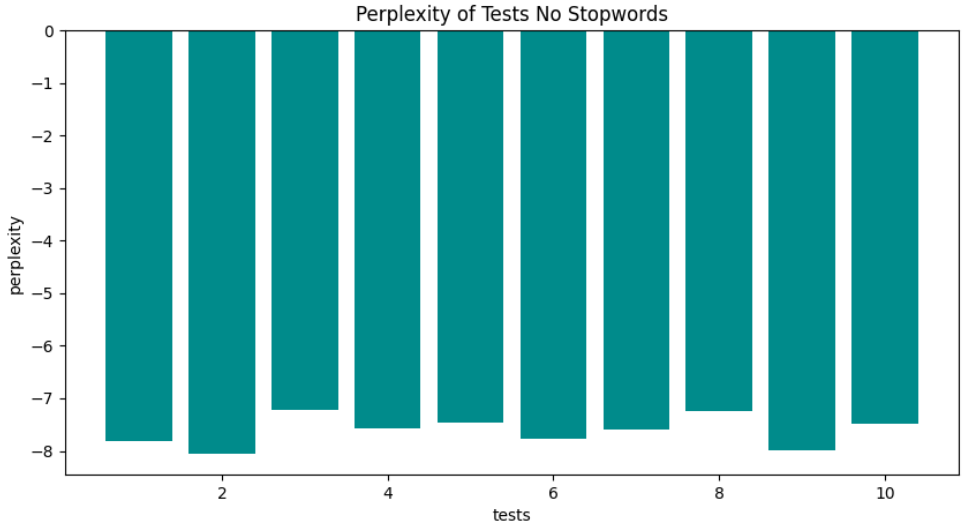
i. Stopwords

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.

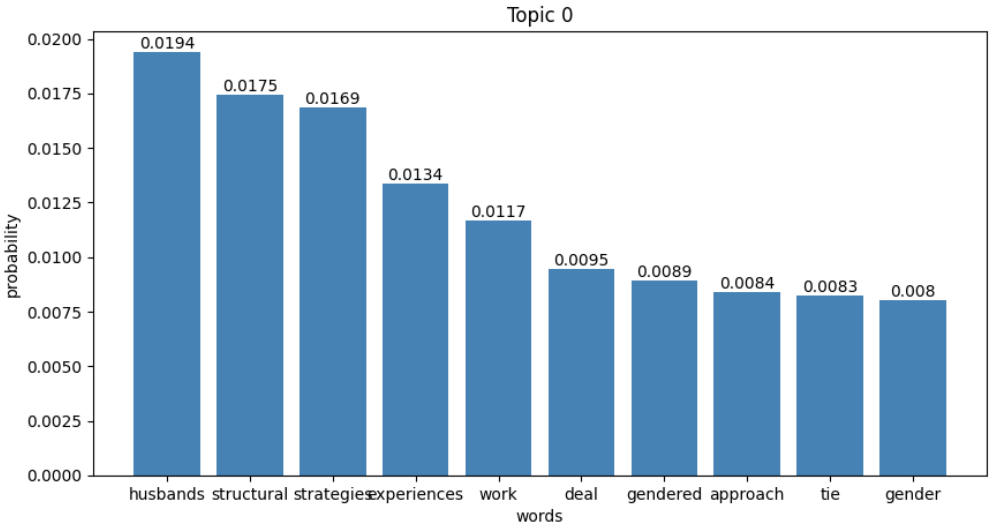
ii. Code Modification

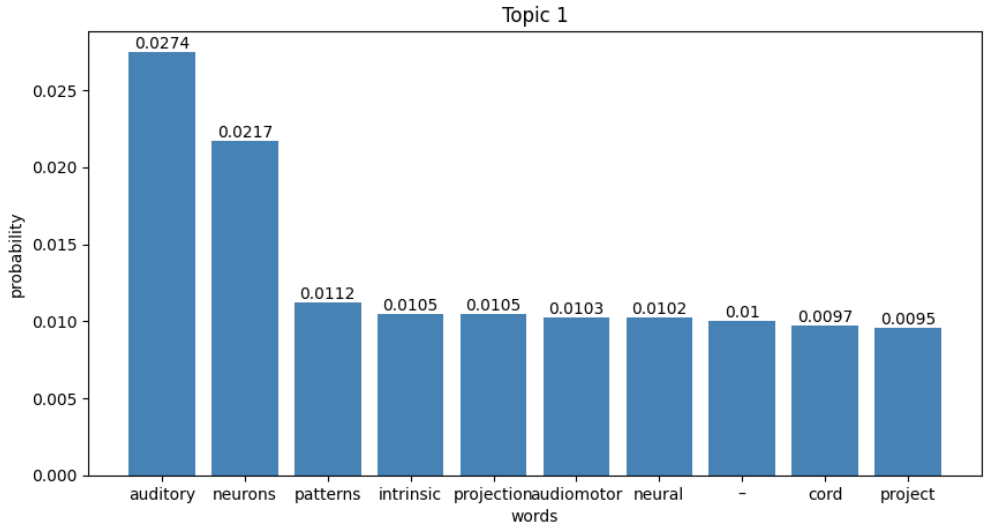
Se creo un nuevo .py en donde se descomentaron las lineas de código que se encargaban de eliminar las stopwords del conjunto de palabras dado en TokenVieuxM.txt.

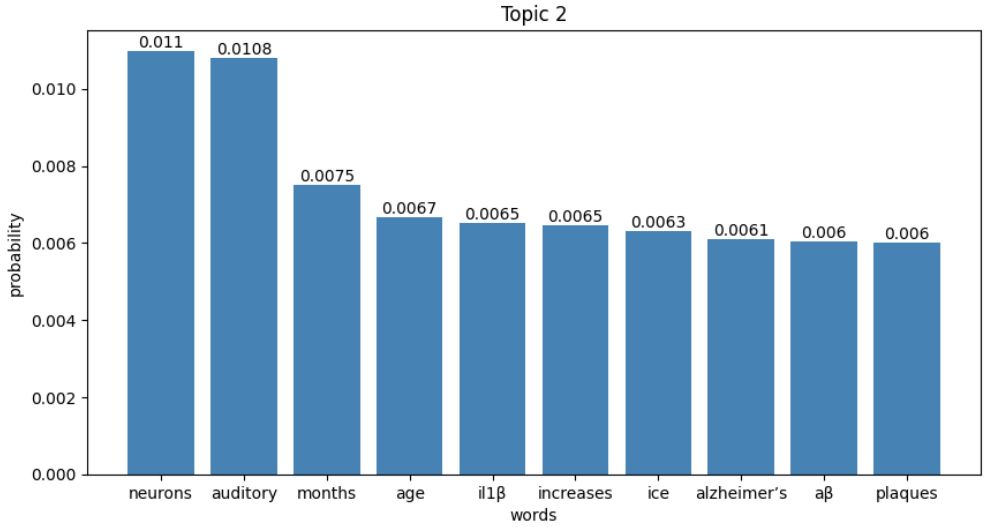

Coherence of Tests No Stopwords
coherence
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
2
4
6
8
10
tests

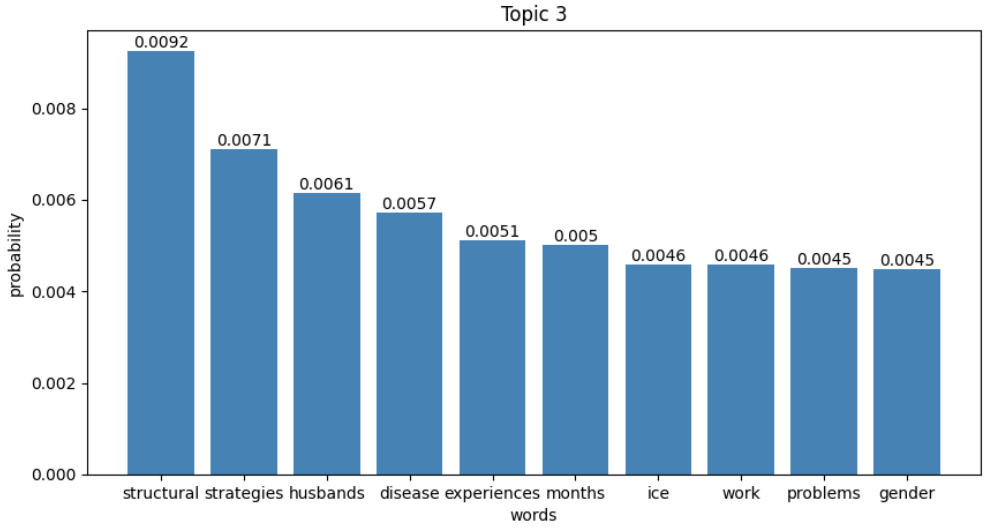

Perplexity of Tests No Stopwords
perplexity
0
−1
−2
−3
−4
−5
−6
−7
−8
2
4
6
8
10
tests

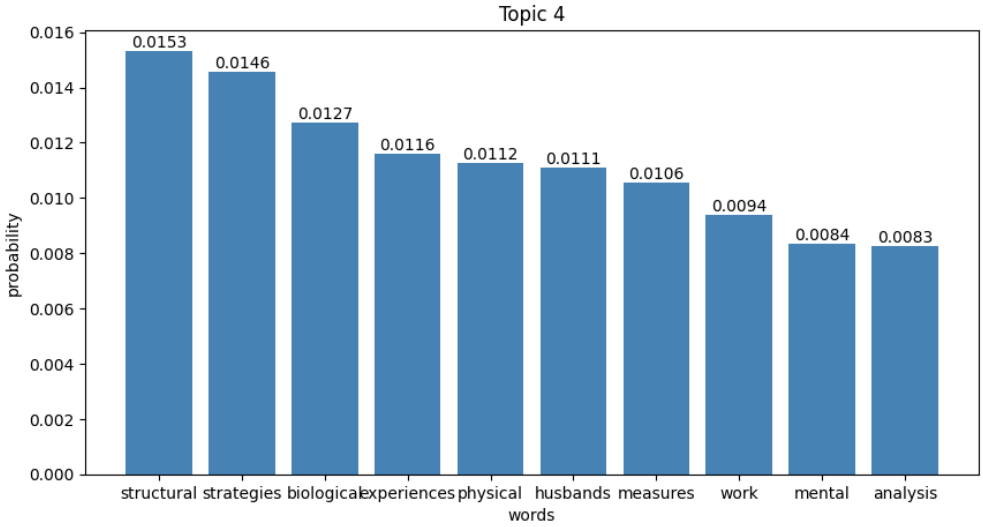
ii.1 Test 8 No Stopwords

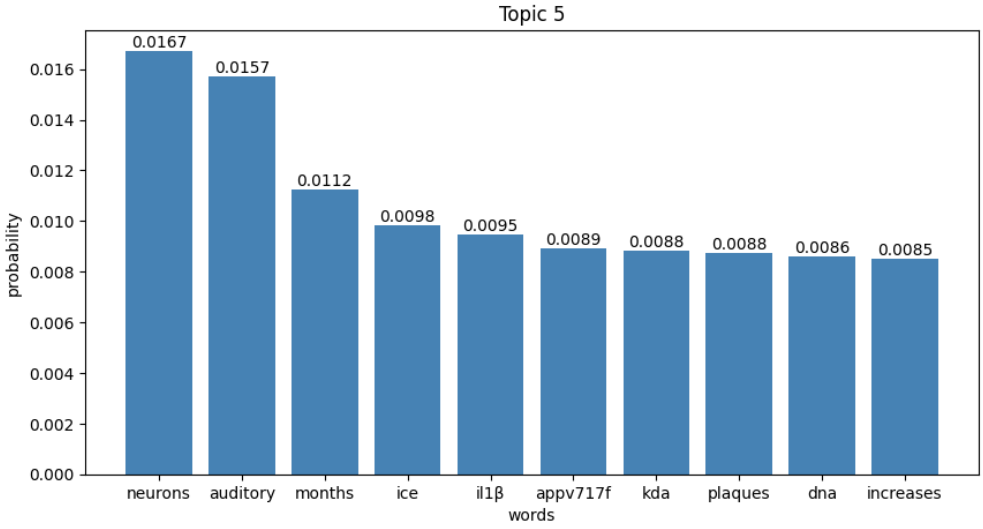


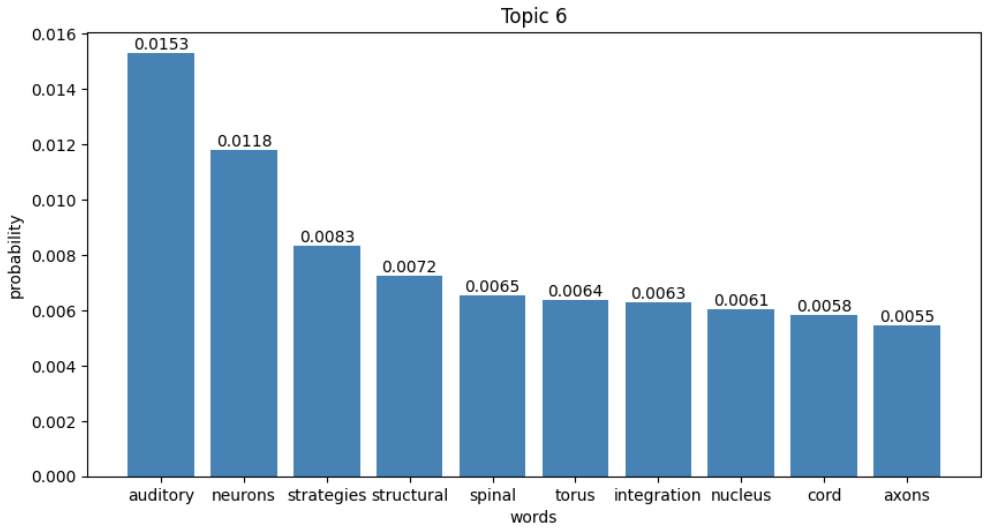


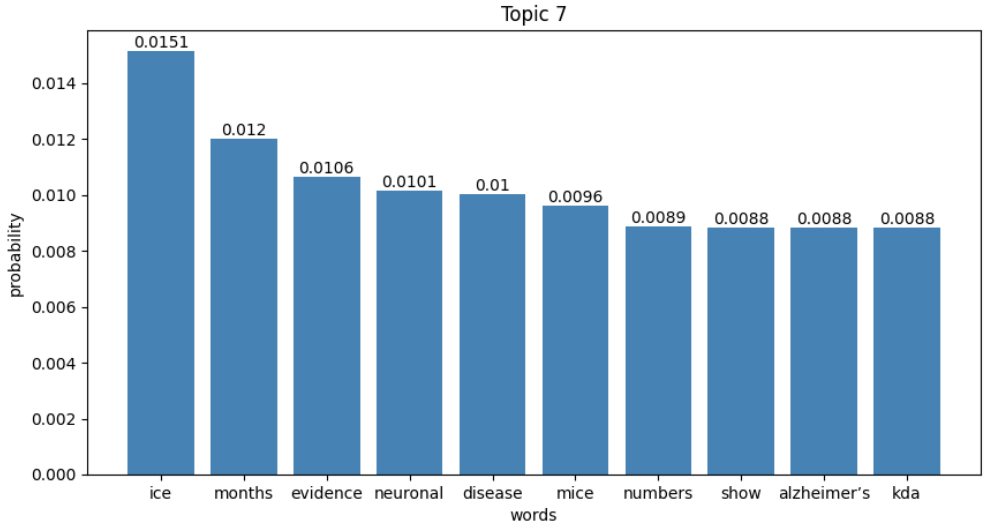

Topic 2
probability
0.000
0.002
0.004
0.006
0.008
0.010
0.011
0.0108
0.0075
0.0067
0.0065
0.0065
0.0063
0.0061
0.006
0.006
neurons
auditory
months
age
il1β
increases
ice
alzheimer's
aβ
plaques
words


Topic 3
probability
0.000
0.002
0.004
0.006
0.008
0.0092
0.0071
0.0061
0.0057
0.0051
0.005
0.0046
0.0046
0.0045
0.0045
structural
strategies
husbands
disease
experiences
months
ice
work
problems
gender
words


Topic 4
probability
words
0.0153
0.0146
0.0127
0.0116
0.0112
0.0111
0.0106
0.0094
0.0084
0.0083
structural
strategies
biological
experiences
physical
husbands
measures
work
mental
analysis


Topic 5
probability
words
0.0167
0.0157
0.0112
0.0098
0.0095
0.0089
0.0088
0.0088
0.0086
0.0085
neurons
auditory
months
ice
il1β
appv717f
kda
plaques
dna
increases

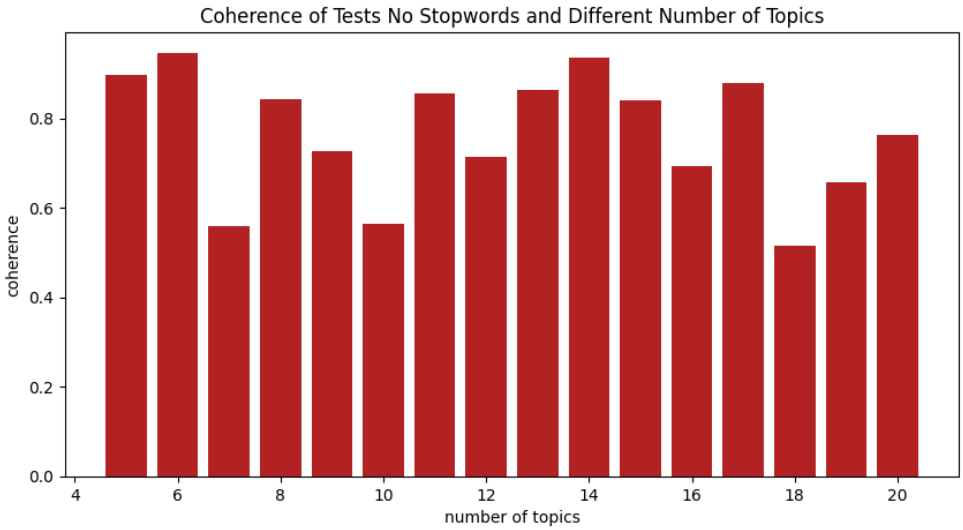


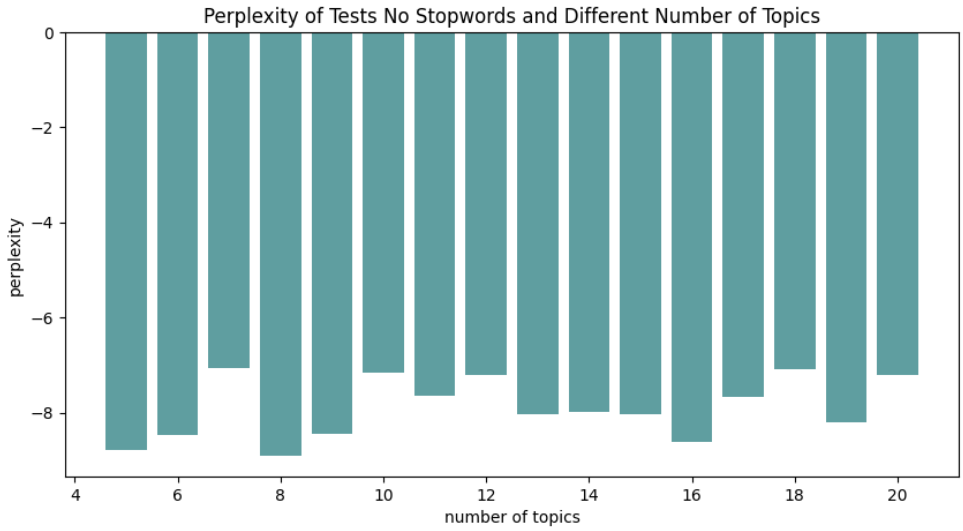


La perplexity es -7.764361092424768 y la coherencia 0.6284575950301475.

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.