CONFERENCIAS MATUTINAS AMLO

Este ejercicio es sobre análisis de tópicos. Un tópico es una variable latente que representa o resume conceptos importantes de un texto, como el significado o las ideas principales del mismo. Un tópico se conforma por varias palabras relacionadas semánticamente entre sí de acuerdo a cierto contexto. En el área de procesamiento de lenguaje natural (NLP), forma parte de una tarea general llamada recuperación de información (IR). Para nosotros, desde la perspectiva de machine learning, la consideraremos como una tarea de aprendizaje no supervisado a partir de una representación vectorial particular de los textos.

Considera una representación documento-término como las que vimos en clase. Una forma sencilla de extraer estructuras latentes entre documentos y términos es usando análisis semántico latente (LSA), el cual se basa en factorizaciones apropiadas de esa matriz. Sea $A_{m\times n}$ la matriz TF-IDF de rango r, con m renglones (documentos) y n columnas (términos). Una aproximación de rango k de esta matriz, está dada por la factorización SVD $A \approx A^{(k)} = U^{(k)}\Sigma^{(k)}V^{(k)\prime}$, donde $\Sigma^{(k)}$ es diagonal con los k eigenvalores más grandes de A y $U^{(k)}$, $V^{(k)}$ contienen los correspondientes eigenvectores izquierdos y derechos que definen una base ortonormal para los espacios columna y renglón, respectívamente. Al aplicar ésta factorización en matrices documento-término, podemos extraer las relaciones semánticas y conceptuales entre documentos y términos expresadas en un conjunto de componentes (o tópicos) k, mediante representaciones densas y de baja dimensión, donde $V^{(k)}_{n\times k}$ y $U^{(k)}_{m\times k}$ nos proporcionan una representación de los términos y documentos, respectivamente, en términos de los k tópicos, y $\Sigma^{(k)}$ nos proporciona la importancia de cada tópico. En Python, puedes usar la implementación de `sklearn.decomposition.TruncatedSVD`.

En este ejercicio, realizarás un análisis de tópicos en las transcripciones de las conferencias matutinas de la presidencia de México, los cuales puedes acceder en este repositorio. Para construir tu modelo de tópicos, considera los textos de las conferencias por semana durante los años 2019 a 2023, usando las transcripciones que corresponden al presidente, contenido en los archivos "PRESIDENTE ANDRES MANUEL LOPEZ OBRADOR.csv".

a) **Representación TF-IDF de los textos:** Obtén una representación TF-IDF de los textos. Define el tamaño del vocabulario y realiza el preproceso que consideres necesario en los textos, considerando que para un análisis de tópicos, no es recomendable que el vocabulario sea tan grande, y es mejor conservar palabras cuyo uso dentro del texto, pueda asociarse con tópicos. Documenta y justifica tus parametrizaciones.

b) **Obtención de k tópicos mediante la descomposición SVD:** Elige un k adecuado y justifícalo. Representa cada tópico mediante un word cloud de los términos que forman cada tópico según la importancia expresada en las magnitudes de los renglones de $V^{(k)}$. ¿Puedes asignar un "nombre" representativo de cada tópico?

c) **Uso del modelo de tópicos ajustado:** Usando el modelo de tópicos ajustado en el paso previo, obtén la representación correspondiente de cada una de las conferencias del presidente durante los años del estudio, calculando la matriz documento-tópico mediante el producto $XV^{(k)}$ (o con el método `transform` de `TruncatedSVD`). Asigna cada conferencia a su tópico correspondiente usando como criterio el valor máximo de cada renglón de la matriz. Usa visualizaciones de baja dimensión basadas en PCA, Kernel PCA y t-SNE de la asignación de tópicos que obtuviste. ¿Observas patrones interesantes? Describe brevemente tus hallazgos.

d) **Uso de NMF para mejorar la interpretabilidad:** Un problema que surge al usar SVD es la falta de interpretabilidad, ya que no es claro cómo pueden considerarse los

valores negativos en las matrices U y V. Una forma de resolver este problema es usar una factorización no-negativa de matrices (NMF), que es adecuada para matrices con entradas no negativas, como las TF-IDF. Para una matriz A de rango r con entradas no-negativas, NMF calcula una aproximación de rango $k < r$ mediante la factorización $A \approx A^{(k)} = W^{(k)}H^{(k)}$, donde $W^{(k)}$, $H^{(k)} \geq 0$. En scikit-learn puedes usar la clase `NMF` del módulo `sklearn.decomposition.NMF`. Repite los incisos anteriores usando esta descomposición. ¿Cuál te parece mejor y por qué?

e) Usando los resultados del método que te parezca más conveniente, (SVD, NMF) construye un indicador semanal para cada uno de los k tópicos durante el periodo de estudio, basado en su frecuencia de aparición. Normalizalos de manera adecuada para que sean comparables y grafícalos como una serie de tiempo. Lo anterior, puede darte un panorama general de la dinámica de los temas que se han tratado en las conferencias matutinas. Realiza un reporte ejecutivo de tus análisis y hallazgos, resaltando las ventajas y desventajas de las metodologías exploradas y da tus conclusiones, incluyendo sugerencias para mejorar el análisis

S O L U C I O N E S

Es necesario comentar antes de iniciar a responder los incisos para lo que fue diseñado este problema, responder acerca de la forma de cargar el repositorio y el preprocesamiento del texto.
Los datos de texto que manejaremos son los datos contenidos en los archivos "PRESIDENTE ANDRES MANUEL LOPEZ OBRADOR.csv", que son recopilaciones de participaciones que ha realizado el presidente de México en cada una de sus mañaneras en nuestro caso nos enfocaremos en los años desde 2019 hasta 2023, y el problema lo abordaremos desde una perspectiva por semanas.
Los datos se encuentran contenidos en el repositorio de Github `https://github.com/NOSTRODATA/conferencias_matutinas_amlo`, de donde se descargaron todos los datos y al final en todas las carpetas contenidas en el se descargaron todos los documentos correspondientes a las participaciones del presidente. Despues se hizo un proceso de división de los textos por semana, es decir, se agruparon los textos de una semana para hacer nuestro análisis de topicos por semanas.
Después de pasar por todo lo anterior para poder apenas tener nuestros datos cargados por semanas en nuestro ambiente de jupyter, iniciamos con un preprocesamiento del texto.
Antes a realizar alg´un análisis con los datos, tenemos que realizar un preproceso y normalización del texto. Para lo anterior se realizo lo siguiente:

1. Convertir las letras en minúsculas
2. Eliminamos signos de puntuación, acentos y otros signos diacríticos.
3. Eliminamos caracteres repetidos, saltos de línea
4. Se quitaron números
5. Eliminamos palabra funcionales (stop words)
6. Aplicamos stemming y lematizaci´on.

Lo anterior se realizo para evitar un sesgo. Se probó con diferentes combinaciones, pero el conjunto de datos que presentaba mejores resultados fue cuando se realizaron las anteriores tareas de limpieza.

Ahora también es importante mencionar el tamaño de vocabulario. Se encontró que entre más pequeño sea el tamaño del vocabulario existe una mayor cantidad de opiniones con distancias cercanas a 1, en cambio cuando el tamaño era demasiado grande las distancias eran muy cercanas a 0. Es decir, lo anterior tiene mucho sentido debido a que la distancia esta muy relacionada con las palabras que se usaran en el vocabulario, para este caso se uso un vocabulario de 1000 palabras

Inciso a) Para esta parte del problema hacemos uso de una libreria de Sklearn llamada TfidfVectorizer, en Python es utilizada para convertir una colección de documentos de texto en una matriz de características TF-IDF (Term Frequency-Inverse Document Frequency). Las parametrizaciones para este indice son las siguientes:

a) Se excluyen términos que aparecen en más del 80 de los documentos, lo que ayuda a descartar palabras comunes que no son útiles para identificar tópicos específicos.

b) Se excluyen términos que aparecen en menos de dos documentos para evitar incluir palabras demasiado raras que no contribuyan significativamente a la identificación de tópicos.

c) Limita el vocabulario a las 1000 palabras más importantes para enfocarse en los términos más relevantes y mantener el modelo manejable y eficiente.

d) Elimina palabras comunes del idioma español que tienen poca relevancia en el análisis.

Lo anterior se hace para obtener una mejor representación de cuales son las la características para el proceso de análisis de texto.

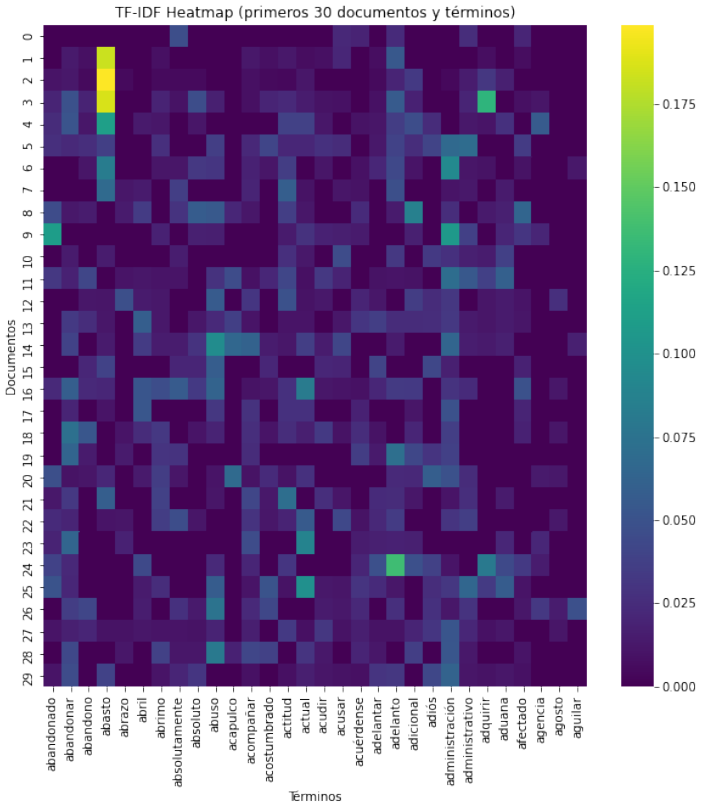
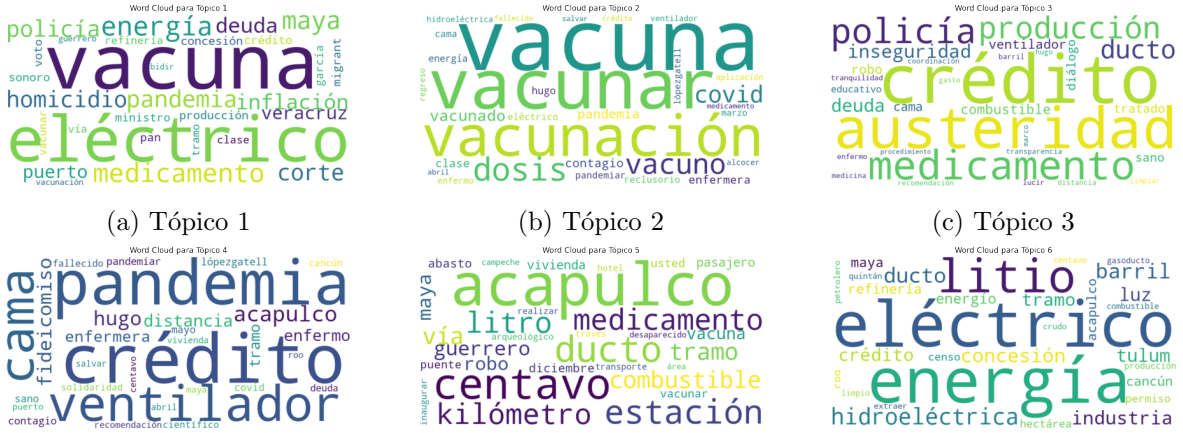


Figura 0.1: TF-IDF Heatmap (primeros 30 documentos y términos)

Al aplicar el proceso debido de esta libreria obtenemos la matriz TF-IDF donde cada fila representa un documento y cada columna representa un término o palabra del vocabulario extraído de todos los documentos. Los valores en la matriz son los pesos TF-IDF que reflejan la importancia de cada término en cada documento, ajustados por la frecuencia del término en todos los documentos.

Como resultado parcial de este inciso podemos hacer un grafico de calor de los primeros 30 documentos y términos. Donde en la imagen 2.1 podemos observar como un diagrama de calor las palabras que tienen mas relevancia dentro de los primeros 30 documentos, un ejemplo claro es la palabra abasto que muestra una clara presencia en los documentos del 2 al 8, donde seria un interesante propuesta de análisis en que temporada se estaba hablando de abasto y a que hacia referencia el presidente al hablar de este tema.

Inciso b) **Obtén k tópicos mediante la descomposición SVD. Elige un k adecuado y justifícalo.**
Usando como hace referencia el ejercicio en python, la implementación de sklearn decomposition.TruncatedSVD. Aplicamos esta descomposición a nuestra matriz TF IDF, esto con el fin de obtener topicos, que son equivalentes a componentes para que podamos encontrar una representación de temas que habla el presidente. En terminos de número de topicos se hicieron pruebas analizando el comportamiento de estos topicos mediante la proyeccion en wordcloud, y se llegó a que un número de topicos representativo e informativo es 6, se muestran a continuación generados mediante un word cloud.



(a) Tópico 1 (b) Tópico 2 (c) Tópico 3

(d) Tópico 4 (e) Tópico 5 (f) Tópico 6

Figura 0.2: Tópicos generados por la Descomposición SVD

¿Puedes asignar un "nombre" representativo de cada tópico?
Resulta difícil poder asignar un nombre representativo para este caso, debido a que la descomposición no lo hace tan bien la parte explicativa y es por esto mismo de sobre dispersion de ceros que hay en nuestra matriz.
Pero si tuvieramos que hacerlo de esta forma, para los seis tópicos representados en la figura 2.2 podemos asignar los siguientes nombres:

a) **Tópico 1:** No encuentro una relación clara de como poder llamar a este tópico

b) **Tópico 2:** Llamaría a este tópico como **COVID-19**, esto debido a que dentro de sus palabras representativas es vacuna, vacunar, dosis, pandemia, palabras que expresan claramente los principales temas que se hablaron en la cuarentena por COVID-19

c) **Tópico 3: ECONOMIA**, Debido a que algunas de sus palabras más representativas es austeridad, crédito y productividad esto nos da a entender que en este topico se intenta demostrar algo sobre la economía nacional, lo que nos pudiera dar indicios a palabras como deuda, inseguridad y combustible

d) **Tópico 4:**No encuentro una relación clara de como poder llamar a este tópico

e) **Tópico 5: DESASTRES**, demostraría que en este topico se intenta expresar lo sucedido en el estado de Guerrero como lo fue el desastre de Acapulco, el desabasto de medicinas y además de los problemas que hubo con el tramo de metro en CDMX y EDO MEX

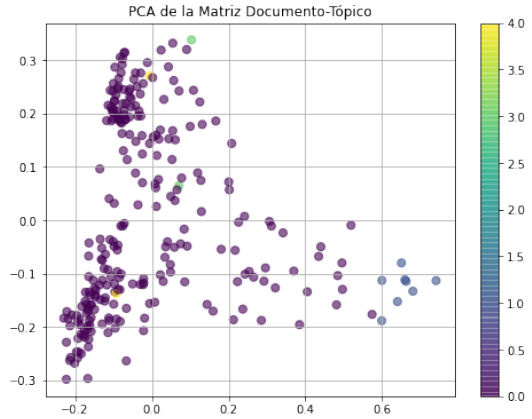
f) **Tópico 6: ENERGIA**, Este tópico es el mas claro ya que demuestra claramente palabras dedicadas a la energia, como litio, electrico, hidroelectrica, industria, barril, en donde posiblemente se desmuestra que se hablaba sobre el abastecimiento energetico del pais

Inciso c) Usaremos el modelo de tópicos ajustado en el paso previo, obtener la representación correspondiente de cada una de las conferencias del presidente durante los años del estudio. Para este caso trabajaremos mediante el metodo de transform de Truncate SVD en Python.

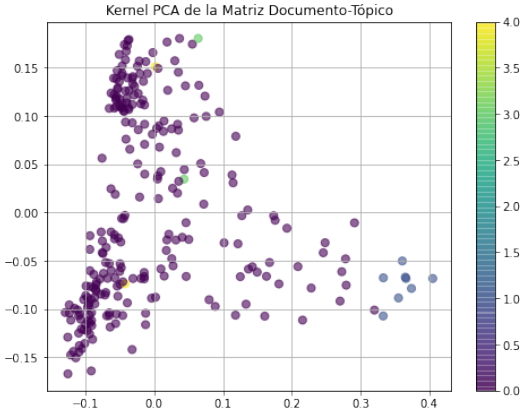
Usamos visualizaciones de baja dimensión basadas en PCA, Kernel PCA y t-SNE de la asignación de tópicos que obtuvimos en el inciso pasado.

Primero asignaremos cada conferencia a sú tópico correspondiente usando el criterio del valor máximo de cada renglón de la matriz, estas visualizaciones las realizaremos con el paquete de SkLearn: PCA, Kernel PCA y TSNE correspondientemente.

Los resultados son los siguientes:



(a) Representación correspondiente de cada una de las conferencias del presidente en PCA



(b) Representación correspondiente de cada una de las conferencias del presidente en Kernel PCA

Figura 0.3: Representaciones

En esta parte podemos observar que PCA y Kernel PCA no hacen ninguna representación representativa de las conferencias, es posiblemente un mismo problema de los valores ceros que encontramos desde aplicar SVD.

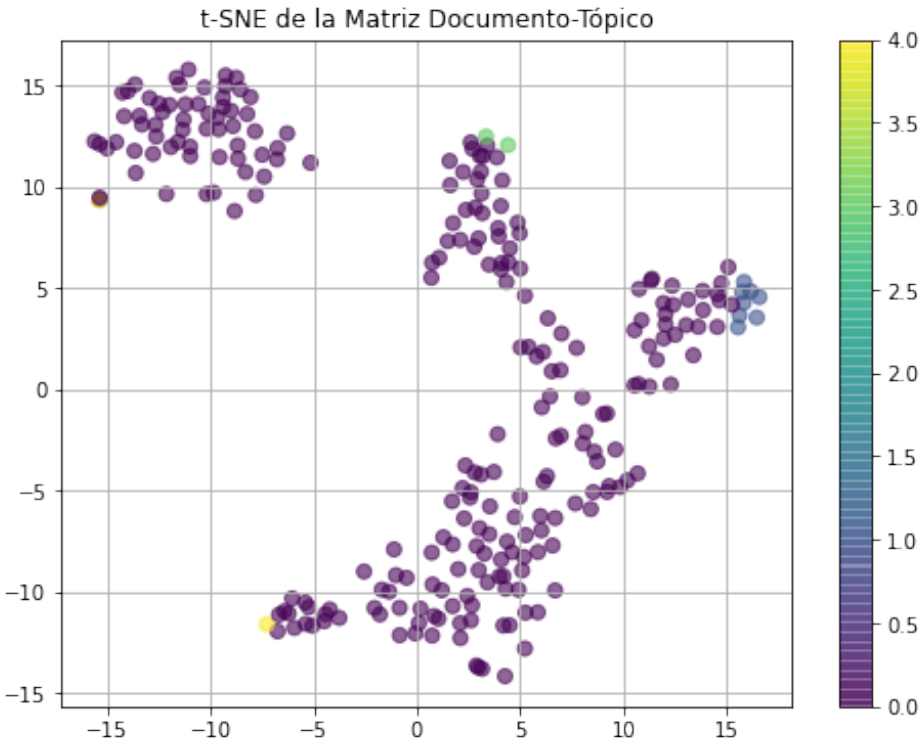


Figura 0.4: Representación correspondiente de cada una de las conferencias del presidente en TSNE

Notemos que T-SNE, intenta rehacer grupos, y es un buen indicio ya que podemos ver que este metodo si pudiera funcionar si aplicamos alguna otra metodología cómo pueden considerarse los valores negativos en las matrices U y V. Si hay grupos de puntos que están claramente separados de otros, esto puede indicar que las conferencias que pertenecen a esos grupos tienen temas muy distintos de las demás

Seguimos teniendo e problema de ceros pero considero que se puede usar este método para el análisis de los siguientes incisos.

Inciso d) En este inciso vamos a abordar el problema que mencionamos en el inciso anterior, la falta de interpretabilidad, ya que no es claro cómo pueden considerarse los valores negativos en las matrices U y V. Una forma de resolver éste problema es usar una factorización no-negativa de matrices (NMF), que es adecuada para matrices con entradas no negativas, como las TF-IDF.

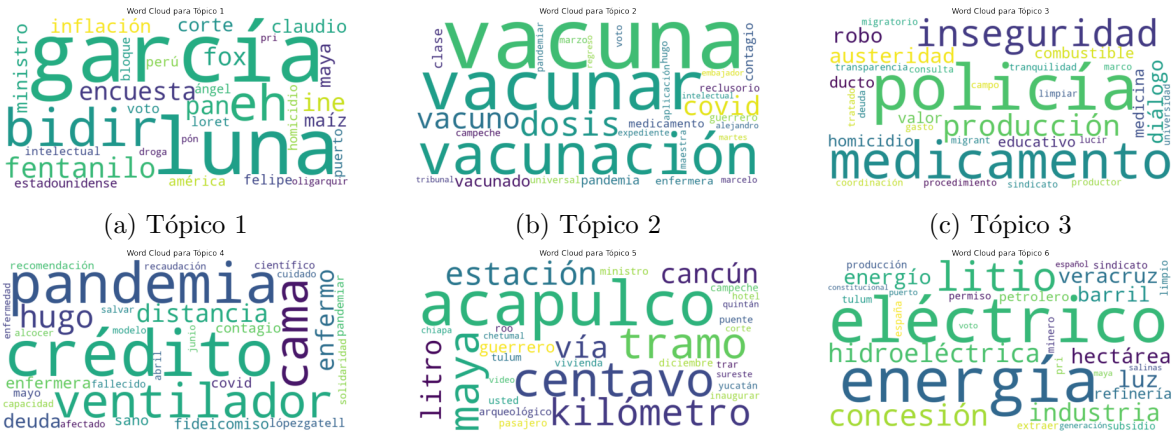
Como se menciona en el problema en python, scikit-learn se usa la clase NMF para apoyarnos de esta herramienta.

Primero vamos a obtener los topicos como lo hicimos en el inciso B, pero ahora mediante la descomposición SVD NMF.

Ahora observamos un WordCloud para analizar los topicos creados pero ahora con la herramienta NMF

¿Puedes asignar un "nombre" representativo de cada tópico?

Ahora observamos que con esta descomposición logramos tener tópicos mas representativos esto debido a que estamos mitigando el problema que mencionamos de como considerar los valores negativos en las matrices.



(a) Tópico 1 (b) Tópico 2 (c) Tópico 3

(d) Tópico 4 (e) Tópico 5 (f) Tópico 6

Figura 0.5: Tópicos generados por la Descomposición SVD-NMF

Se asignará un "nombre", que considero representativo de cada tópico para el hecho de caso de cada estudio.

a) **Tópico 1: POLITICA**, ahora a diferencia de la descomposición anterior vemos que el tópico uno si tiene mayor coherencia al momento de asignar palabras representativas. En este caso asigna "Garciaz "Luna"que pueden hacer referencia al Ex secretario de seguridad pública de México, o también palabras como ÏNE", .Encuesta", "Fox", "PAN", haciendo alución a la politica de México.

b) **Tópico 2:** Nuevamente llamaría a este tópico como **COVID-19**, esto debido a que dentro de sus palabras representativas es vacuna, vacunar, dosis, pandemia, palabras que expresan claramente los principales temas que se hablaron en la cuarentena por COVID-19

c) **Tópico 3: ECONOMIA**, Debido a que algunas de sus palabras más representativas es austeridad, crédito y productividad esto nos da a entender que en este topico se intenta demostrar algo sobre la economía nacional, lo que nos pudiera dar indicios a palabras como deuda, inseguridad y combustible

d) **Tópico 4:**Lamentablemente sigo sin encontrar una relación clara de como poder llamar a este tópico, pero se logran ver palabras de economía y salud.

e) **Tópico 5: TURISMO**, demostraría que en este topico se intenta expresar algo más claro como son los avances y problemas que enfrentan zonas de turismo de México, por ejemplo el desastre que ocurrio en acapulco, Guerero, tambien se habla de la palabra "Maya"que pudiera hacer referencia a la construcción del tren Maya.

f) **Tópico 6: ENERGIA**, Este tópico sigue siendo el mas claro ya que demuestra claramente palabras dedicadas a la energia, como litio, electrico, hidroelectrica, industria, barril, en donde posiblemente se desmuestra que se hablaba sobre el abastecimiento energetico del pais

Respondiendo a la pregunta del inciso en general, es claro que la representación NMF es mejor ya que nos ayuda con el problema de la representación de los valores negativos de nuestra matriz TF IDF.
Lo anterior podemos ver claro un representación por el método de T-SNE, como sigue:

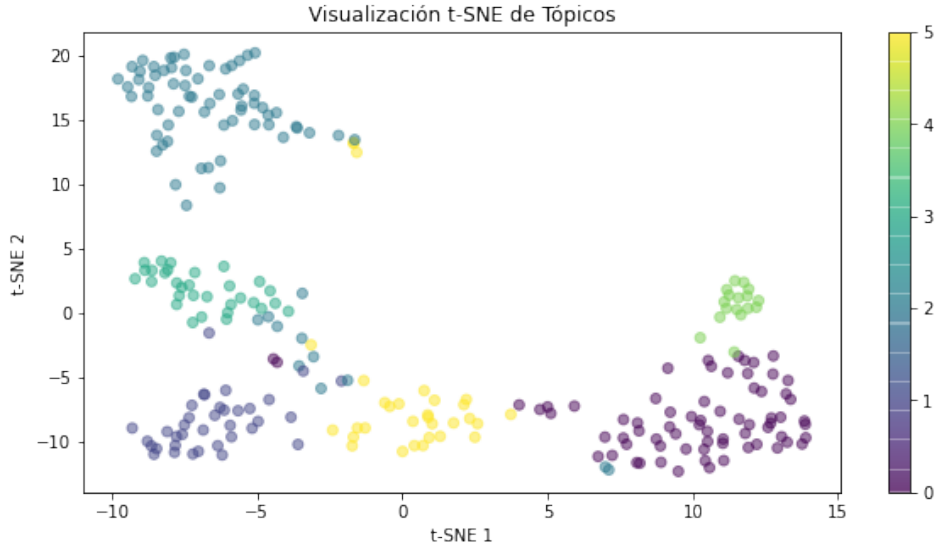


Figura 0.6: Representación correspondiente de cada una de las conferencias del presidente en TSNE

Vemos en esta imagen 2.6 que logra hacer mejores agrupaciones de los topicos, y asi poder dividir cada mañanera en un mejor correspondiente tópico

Inciso e) Usaremos los resultados del método SVD-NMF para construir un indicador semanal para cada uno de los k tópicos durante el periodo de estudio, basado en su frecuencia de aparición. Se van a Normalizar como lo indica el problema para tener bases de comparación y se graficaron en una serie de tiempo como se muestra a continuación.

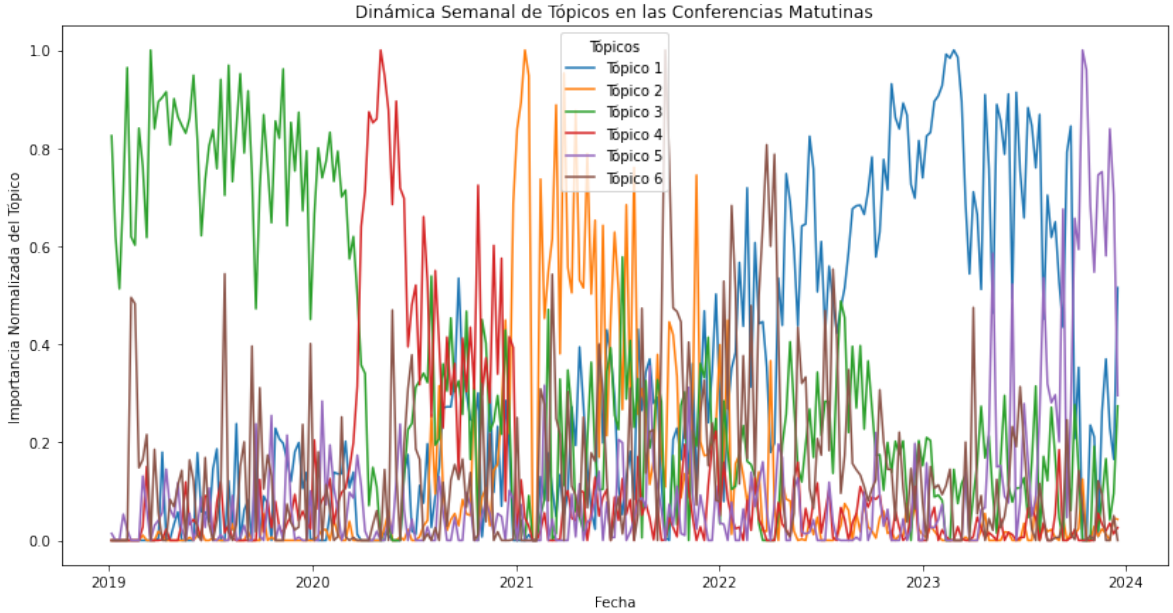


Figura 0.7: Representación correspondiente de cada una de las conferencias del presidente en TSNE

A primera vista no vemos algo claro dentro del analisis en la serie de tiempo, pero si ampliamos nuestra vista podemos observar datos interesantes, por ejemplo en el topico

No. 3, vemos que en el 2019 fue un año donde este tópico tuvo más presencia ya que fue un año lleno de inseguridad y de problemas economicos como lo puede ser el combustible y robos, entonces al ser un año malo para México podemos ser claros con que este tópico si tuvo mayor presencia en 2019.

El tópico 2 es un caso interesante de este análisis ya que llamamos a este tópico COVID-19, debido a que se mencionan palabras que involucran temas de la cuarenta, y se muestra su clara presencia a mediados de 2020 que fue el año donde se declaro la cuarentena a nivel mundial, y tuvo mucha más presencia en 2021 primeros meses de 2022 ya que en los primeros meses México registró 60 552 casos nuevos de COVID-19, es la cifra más alta de toda la pandemia en el país, que fue el periodo donde se estudiaba la cuarentena y el tema de mayor importancia eran las vacunas.

Topico 1, para el estudio en la serie de tiempo es la barra azul, y este tópico llamado POLITICA tuvo un gran incremento a partir de 2022, esto posiblemente por la cercania a las elecciones de los estados como: Aguascalientes, Durango, Hidalgo, Oaxaca y Tamaulipas, en 2023 se siguio con esta tendencia pero más linea, esto sugiere posiblemente que sucede por la cercania a las elecciones cercanas del 2024.

El tópico 5, que pertenece al TURISMO, tiene mas sentido que sea mencionado en el segundo semestre del 2023 y 2024, esto debido a dos temas para el gobierno muy importante, como lo es Acapulco, ya que recordemos que en está fecha fue el desastre natural que ocurrió en Guerrero, y además otro tema importante es el del tren Maya, ya que por su inaguración tuvo un gran auge y se demuestra su tendencia en la serie de tiempo.

El tópico 6, muestra un aumento en tendencia en 2022 esto puedde ser provocado por varias cosas como que la Cámara de Diputados rechaza la reforma constitucional en materia energética del presidente López Obrador. También en ese año el presidente López Obrador inaugura la Refinería Olmeca con sólo el 65 % de avance en su construcción

Conclusiones:

1. SVD (Singular Value Decomposition) e una técnica matemática que descompone una matriz en tres matrices más simples (U, Σ, V^T). Puede aplicarse a cualquier matriz, independientemente de si sus entradas son negativas o no.

2. Los valores en las matrices U y V pueden ser negativos, lo que a veces dificulta la interpretación, especialmente en el contexto de PLN, donde se espera que las características (como palabras) tengan una contribución no negativa.

3. SVD es particularmente útil para identificar las relaciones latentes en los datos y es ampliamente utilizado en sistemas de recomendación y en la compresión de imágenes.

4. NMF (Non-negative Matrix Factorization) es una técnica de descomposición que también factoriza una matriz en dos, W y H, pero con la restricción de que todas las matrices involucradas tengan solo valores no negativos.

5. La no negatividad hace que los componentes sean más fácilmente interpretables ya que no hay que considerar la dirección negativa de una característica, lo cual es intuitivamente más fácil de entender y se asemeja a "partes.º "porciones"de los datos originales.

6. La Representatividad de los Tópicos: Los tópicos identificados parecen capturar de manera efectiva los temas predominantes en las conferencias del presidente.

7. A partir de las visualizaciones de baja dimensión, se observó agrupaciones de conferencias por tópicos, lo que puede indicar períodos durante los cuales el presidente se enfocó en temas específicos.

8. La interpretación de los tópicos se basa en el contenido y las palabras clave que constituyen cada tópico. Esto puede proporcionar una visión cualitativa de las preocupaciones y enfoques del presidente a lo largo del tiempo.

9. El modelo de TruncatedSVD. Esto transforma los documentos de alta dimensionalidad (conferencias) en un espacio de características reducido (tópicos), facilitando su análisis y comprensión.