

# Modelo de tópicos para los discursos mañaneros del presidente Andrés Manuel López Obrador

---

Armando Salinas Lorenzana

30 de agosto de 2024

## 1. PROBLEMA

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  $A \approx A^{(k)} = U^{(k)}\Sigma^{(k)}V^{(k)'}$ , 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, respectivamente. Al aplicar esta 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  $V_{n \times k}^{(k)}$  y  $U_{m \times k}^{(k)}$  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 transcrip-

ciones que corresponden al presidente, contenido en los archivos “PRESIDENTE ANDRES MANUEL LOPEZ OBRADOR.csv”.

- a) 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) Obtén  $k$  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) 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) 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. Normalízalos de manera adecuada para que sean comparables y gráficalos 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.

#### 1.0.1. SOLUCIÓN

Este ejercicio puede verse resuelto en el archivo `codigo.py`.

- a) En este trabajo realizamos el análisis de los reportes de las conferencias mañaneras del Presidente Andrés Manuel López Obrador, dado el repositorio de estos informes proporcionados. El análisis se realizó desde el año 2019 hasta el año 2023, por lo que el conjunto de documentos de reportes fueron importados en dicho archivo, fueron procesados y posteriormente analizados.

La primera parte después de importar los datos fue procesar el texto. Para esto los datos se importaron en un vector, donde cada vector contiene un reporte en forma de cadena. Entonces se le aplicó un preproceso a todos los reportes, primeramente se le quitaron los *stop words*, es decir palabras que no tienen mucho significado, por sí solas y que no aportan demasiada información, esto ayuda a centrarse en palabras clave importantes que son más informativas para la búsqueda o el objetivo

que deseamos, en este caso, identificar tópicos, por tanto todas estas palabras fueron borradas, sin embargo hay un punto importante a comentar en esta parte, en un principio este proceso de eliminar los *stop words* se realizó con una lista de estas palabras para el lenguaje español importado de *nltk.corpus*, y tras realizar el análisis de los reportes después del preproceso, me percaté de que aún habían muchas palabras que no aportaban mucho al análisis, por lo que me encargué de buscar un nuevo repertorio de *stop words* que me pudiera ser más útiles, por fortuna encontré un repertorio, más extenso y el cual se utilizó para este trabajo dado a los mejores resultados que fueron arrojados tras implementarlo, esta misma lista será subida en conjunto con la tarea.

Posteriormente de quitar los *stop words*, se quitaron los signos de puntuación, comas, puntos, puntos y comas, dos puntos, etc. También todas las palabras fueron convertidas a minúsculas para tener un estándar de las palabras, y se les quitaron las tildes a todas las palabras y por último se le quitarón los números que podrían haber en los textos, debido a que por lo regular resultan ser elementos muy específicos del contexto y no representa información global de los temas, por esta razón fueron considerados eliminarse de dichos reportes. Tras implementar este preproceso a los datos, obtuvimos un reportes como el que se muestra a continuación

```
dias reiterar manifestacion deseo ano ano anos pienso ano optimista mencione  
condiciones inmejorables mejore economia pais crecimiento economia generen  
empleos gente salarios mejoren condiciones vida garantice derecho educacion  
salud consiga paz pais trabajando proposito conferencia matutina ano presentamos  
convocatoria reclutamiento jovenes integracion guardia nacionalinvitamos jovenes  
mexico mayores anos contribuir importante labor seguridad publica nacional  
adiestramiento policial especializado solida capacitacion respeto derechos humanos  
iniciaron cambios ido avanzando empezar asegurar equilibrios macroeconomicos...
```

De esta forma podemos darnos cuenta que se pierde completamente el contexto pero las palabras que terminan quedándose resultan ser muy importante y significativas, palabras que puedan asociarse con tópicos. Por otra parte para el número de palabras a considerar, se propusieron varias opciones de las cuales se percató que un número no muy grande de vocabulario terminaba dando buenos resultados, para este caso se consideraron 500 palabras.

Con ayuda de la librería `sklearn.feature_extraction.text` se pudo importar una función para generar la matriz de TF-IDF, que es una forma de representar documentos de texto como vectores numéricos, la cual es comúnmente utilizada en el procesamiento del lenguaje natural (NLP) para medir la importancia de una palabra en un documento en relación con una colección de documentos o corpus, en este caso los reportes de las conferencias mañaneras. Este método es particularmente útil en tareas como la recuperación de información y la clasificación de texto. Refleja la frecuencia de una palabra en un documento específico y mide la importancia general de la palabra en todo el corpus.

- b) Ya que se realizó todo el preproceso de los reportes en el inciso a) se procedió a realizar la descomposición SVD, de aquí pudimos extraer los vectores propios y los valores propios, lo cual nos serían de gran ayuda para determinar el número de tópicos adecuados a considerar, dado que obtuvimos la gráfica de la varianza en cada vector propio, el cual podemos observar a continuación.

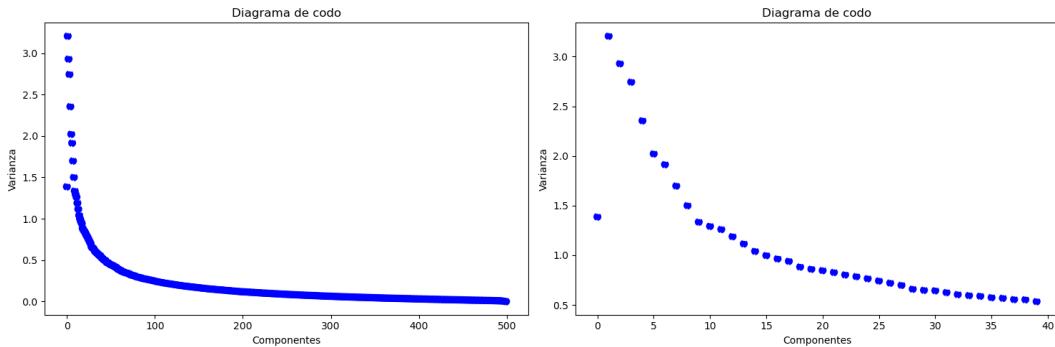


Figura 1.1: Diagrama de codo de la descomposición SVD a la matriz TF-IDF. En la figura izquierda podemos ver el perfil completo del diagrama de codos, como es evidente y de esperarse, la mayor parte de la varianza se encuentra en los primeros componentes, del lado derecho tenemos la misma figura pero con un zoom en la ventana [0,40], aquí podemos apreciar a una menor escala que la mayor parte de la varianza se encuentra entre los primeros 20 componentes, sin embargo, para simplificar aún más el número de componentes, pódemos observar un quiebre alrededor del componente 8, lo cual nos da un indicio de decidir el valor de  $k$  en torno a este número.

Tras observar la figura del diagrama de codo tenemos un candidato para el valor de  $K$ , siendo  $k = 8$ . Para este trabajo se consideró este valor del número de tópicos. Una vez que conocemos el número de componentes, obtuvimos las graficos word cloud, que no es más que una representación visual de texto que muestra la frecuencia de las palabras mediante diferentes tamaños de fuente o colores. En una nube de palabras, las palabras más frecuentes en el texto aparecen en un tamaño de fuente mayor o en un color más destacado, mientras que las palabras menos frecuentes se muestran en tamaños más pequeños o colores menos llamativos. Este tipo de visualización es útil para identificar rápidamente los términos más prominentes o relevantes en un conjunto de datos textuales, lo que permite obtener una visión general del contenido de manera intuitiva y estéticamente agradable. Por lo cual se obtuvieron 8 word clouds. Que se muestran a continuación.



En los word clouds podemos distinguir las palabras más frecuentes de cada tópico, observando las palabras encontradas en cada tópico podemos asociar cada uno con un tema en especial.

- Tópico 1: En este tópico podemos encontrar palabras clave, como México, pueblo, gobierno, país, gente. Este conjunto de palabras son distintas entre sí y no representan nada en común, este conjunto de palabras forma parte del vocabulario que el presidente utiliza día con día en sus conferencias, por lo que este tópico representa esas palabras que siempre se usan, y no representan un tema en específico.
- Tópico 2: En éste tópico vemos palabras como salud, médicos, hospitales, vacunas, por lo que podríamos considerar que se trata de **salud**.
- Tópico 3: En éste tópico podemos ver palabras como tren, energía, mil, petróleo, maya, eléctrica, millones, aeropuerto. Éste tópico contiene palabras muy distintas que no comparten un tema en común, por lo que este tópico se vuelve realmente difícil de interpretar.

- Tópico 4: En éste tópico podemos ver que las palabras con más frecuencia fueron electricidad, comisión, energía, pemex, electrica, federal, gas corrupción. Por lo que éste tópico lo podemos relacionar con la **energía**.
- Tópico 5: En éste tópico podemos ver que las palabras más frecuentes fueron seguridad, mil, guardia, violencia, seguridad, homicidios, jóvenes, robo, defensa. Por lo que podemos encontrar cierta relación entre estas palabras y podemos decir que éste tópico se relaciona con la **seguridad**.
- Tópico 6: En éste tópico podemos ver que las palabras más frecuentes son vacuna, vacunas, guardia, adultos, mayores, seguridad. Podemos ver en éste tópico que las dos palabras más frecuentes son vacuna y vacunas, muy seguramente las vacunas contra el covid-19, que fueron un tema constante cuando se comenzaron a aplicar las vacunas en méxico. Sin embargo, al observar las demás palabras no podemos encontrar una relación con las vacunas, por lo que de forma general éste tópico se podría volver difícil de interpretar y ponerle una etiqueta.
- Tópico 7: En éste tópico podemos ver que las palabras más frecuentes son unidos, méxico, política, migrantes, tratado, países, economía, américa. Quizás éste tópico podría tratarse del tema de **migración** aunque no con mucha certeza, por todas las demás palabras que contiene.
- Tópico 8: En éste tópico podemos ver que las palabras más frecuentes son salud, tren, medicamentos, corrupción, aeropuesto, maya, tramo. Nuevamente éste tópico se vuelve difícil de interpretar por la diversidad de las palabras que no tienen una relación en común.

Como pudimos observar, la interpretabilidad de alguno de los tópicos no fue la más clara, debido a que tenían palabras que no tenían una relación en conjunto, por lo que algunos tópicos no fue posible ponerle una etiqueta.

- c) Se realizó posteriormente una representación de baja dimensión en PCA, kernel PCA y t-SNE de la asignación de tópicos que se obtuvieron. Previo a esto se calculó la matriz documento-tópico mediante el producto  $XV^{(k)}$ , y se clasificó cada documento a su correspondiente tópico considerando el índice del mayor elemento en la fila. Esta representación de baja dimensión se muestra a continuación.

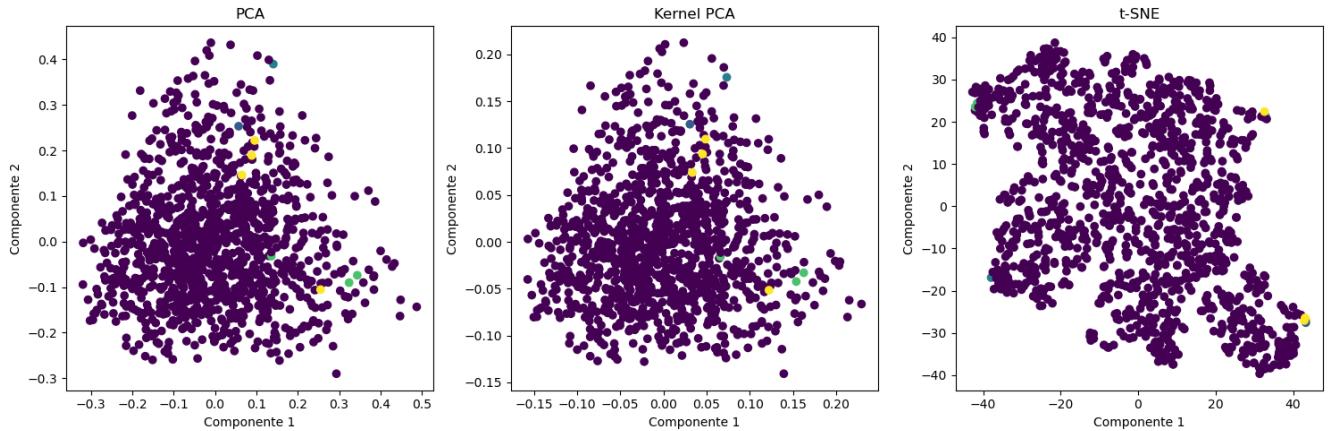
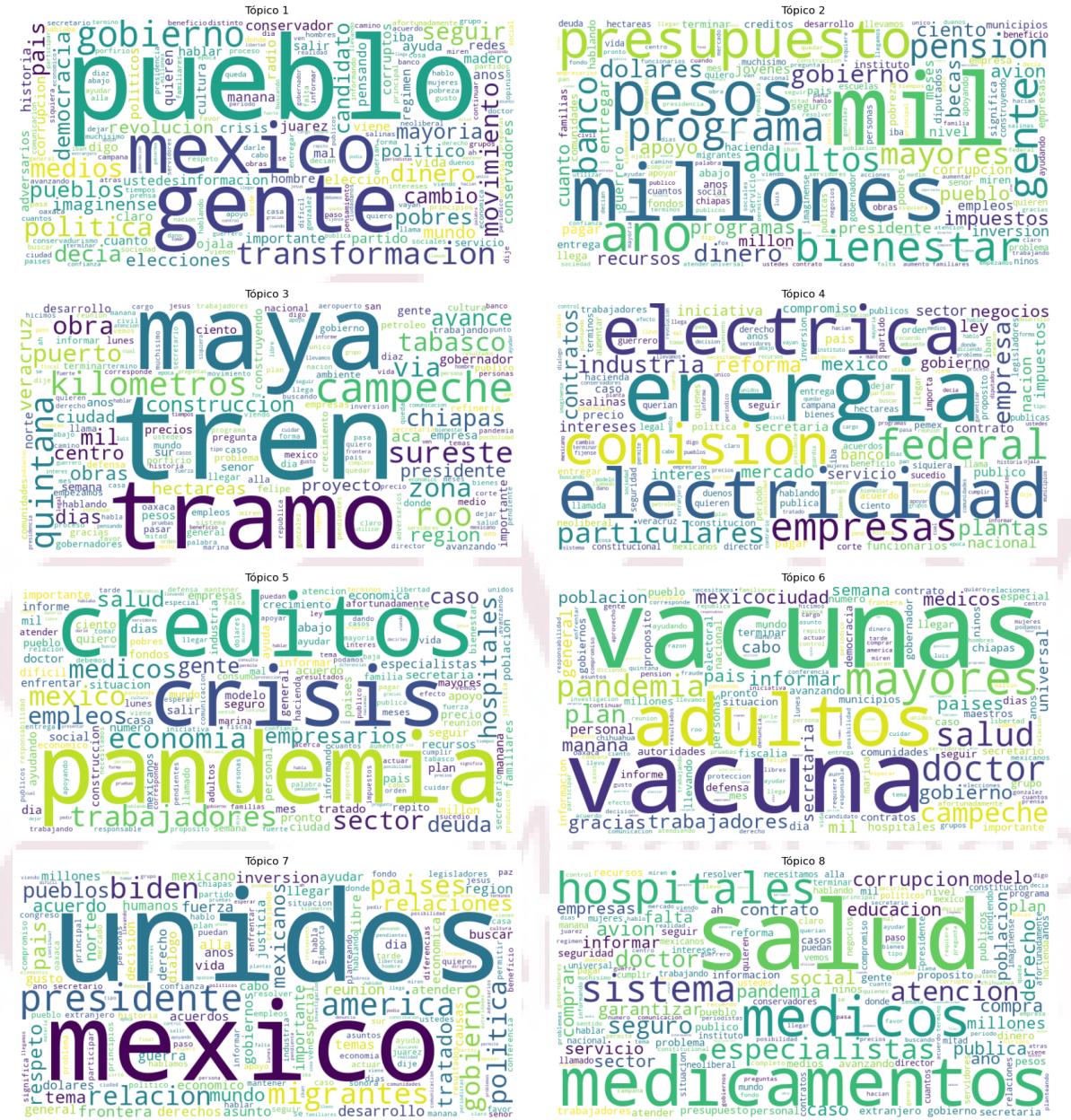


Figura 1.2: Representación en baja dimensión mediante tres métodos distintos. Como podemos observar en las tres representaciones, podemos darnos cuenta que la clasificación que se realizó no fue la mejor puesto que clasificó la mayoría de documentos en un mismo tópico, por esta razón podemos ver la mayoría de puntos un mismo color, y muy pocos elementos de un color distinto, en esta interpretación, no es posible hacer una buena distinción de los tópicos seleccionados, solo puede observarse una estructura global pero no es de gran utilidad.

- d) Este problema de mala clasificación se debe a que al hacer la descomposición SVD, surge el detalle de la interpretabilidad puesto que aparecen valores negativos en las matrices  $U$  y  $V$ . Por esta razón se utiliza la factorización no negativa de matrices que es una técnica de reducción de dimensionalidad y también de clustering que es utilizada para descomponer matrices en las que todos los elementos son no negativos. Esto resulta particularmente útil en el análisis de datos donde las matrices de entrada no contienen elementos negativos, como es el caso en los datos de procesamiento de imágenes y análisis de texto. Por esta razón realizamos la factorización NMF a la matriz TF-IDF.

Una vez aplicada esta factorización a la matriz TF-IDF, obtuvimos una nueva matriz de documentos-topicos, de la cual clasificamos cada uno de los reportes de acuerdo al índice del mayor elemento de la fila del reporte en la nueva matriz de documentos-topicos, de esta forma pudimos realizar nuevamente los word clouds, considerando nuevamente la cantidad de 8 tópicos, de esta forma podemos ver los resultados a continuación.



En los word clouds podemos distinguir las palabras más frecuentes de cada tópico, observando las palabras encontradas en cada tópico podemos asociar cada uno con un tema en especial.

- Tópico 1: En este tópico podemos encontrar palabras clave, como México, pueblo, gobierno, país, gente, gobierno. Nuevamente al igual que los primeros word clouds, éste tópico se hace presente, éste tópico tiene que ver con palabras que el presidente comúnmente menciona cuando simplemente habla de cualquier tema, ya que siempre menciona México, pueblo, gente, gobierno, etc, por lo que podemos etiquetar éste tópico como **México**.
- Tópico 2: En éste tópico vemos palabras como presupuesto, programa, adultos, mayores, pensión, bienestar, dinero, recursos, lo cual tiene que ver con el dinero y el apoyo del gobierno a las personas mayores, adultos o jóvenes, de esta forma podemos encontrar la relación entre las palabras y poder etiquetar éste tópico como el **programas de apoyo del gobierno**.
- Tópico 3: En éste tópico podemos ver palabras como tren, maya, tramo, kilómetros, campeche, electricidad, energías, comisión, federal, etc.

sureste, tabasco, construcción y ya podemos intuir con claridad que se trata del tren maya, debido a las palabras comunes y sobre todo los estados que aparecen donde principalmente se llevó acabo la construcción de las vías para el tren maya, por lo que éste tópico puede tomar la etiqueta de **tren maya**.

- Tópico 4: En éste tópico podemos ver que las palabras que más frecuencia tuvieron fueron electricidad, comisión, energía, electrica, federal, industria. Por lo que éste tópico lo podemos relacionar con la industria de **energía eléctrica**.
- Tópico 5: En éste tópico podemos ver que las palabras más frecuentes fueron pandemia, crisis, economía, empleos, méxico, deuda, hospitales. Todas estas palabras dan indicios que se trata de la pandemia probocada por el covid-19 que vivimos y tocó enfrentarlo mientras el presidente Andrés Manuel estaba en el cargo, de esta forma podemos identificar con claridad que este tópico se relaciona con la **pandemia**.
- Tópico 6: En éste tópico podemos ver que las palabras más frecuentes son vacuna, vacunas, pandemia, adultos, mayores, hospitales, sistema. Podemos ver en éste tópico que las dos palabras más frecuentes son vacuna y vacunas, muy seguramente las vacunas contra el covid-19, que fueron un tema constante cuando se comenzaron a aplicar las vacunas en méxico. Por lo que podemos relacionar este tópico con el tema de las vacunas y todo el proceso que presentó méxico al aplicar las vacunas y distribuirlas en toda la república, de esta forma podemos etiquetar éste tópico con el tema de las **vacunas**.
- Tópico 7: En éste tópico podemos ver que las palabras más frecuentes son unidos, méxico, biden, migrantes, tratado, países, economía, américa, relación. Podemos ver claramente que estas palabras se relacionan estrechamente al fenómeno de la migración por parte de los países lationamericanos que tuvieron que pasar por méxico para llegar a los estados unidos mientras se encontraba gobernando el presidente Biden y permitió la llegada de los inmigrantes, por tanto este tópico lo podemos etiquetar como **migración**.
- Tópico 8: En éste tópico podemos ver que las palabras más frecuentes son salud, médicos, medicamentos, sistema, hospitales, seguro, especialistas, atención y ya podemos ir intuyendo que se trata de salud en general, por lo tanto podemos etiquetar a éste tópico como **salud**.

Es impresionante cómo después de aplicar la factorización a la matriz TF-IDF y repitiendo los incisos anteriores pudimos obtener mejores resultados, lo cual se vieron reflejados en los word clouds al ver grupos de palabras bastante relacionadas entre sí, pudiendo identificar con mayor claridad el tema que representaban en conjunto, por tanto éste método de NMF resultó mejor que el SVD.

Así podemos concluir que los primeros 8 tópicos de los temas que más se hablaron fueron

1. **México**
2. **Programas de apoyo del gobierno.**
3. **Tren maya.**
4. **Energía eléctrica.**
5. **Pandemia.**
6. **Vacunas.**
7. **Migración.**
8. **Salud.**

Por otra parte, también aplicamos la reducción de dimensiones con los tres distintos métodos y obtuvimos los siguientes resultados.

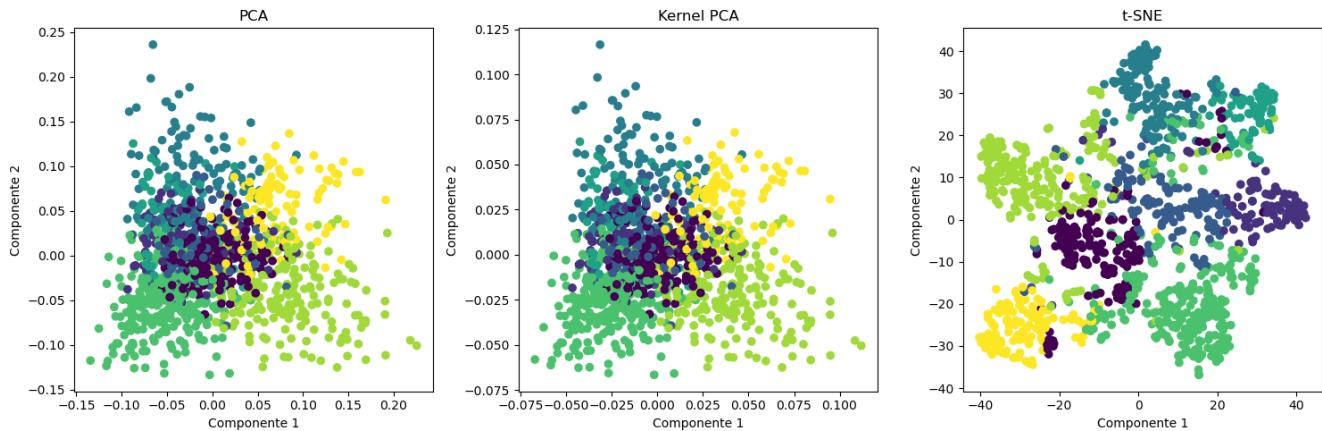


Figura 1.3: Representación en baja dimensión mediante tres métodos distintos con la factorización aplicada. En esta ocasión podemos observar claramente que existen grupos definidos en la representación de baja dimensión, podemos asociar cada color de grupo con su correspondiente tópico de la siguiente manera, Violeta oscuro: México, Violeta claro: Programas de apoyo del gobierno, Azul verdoso: Tren maya, Verde azulado: Energía eléctrica, Verde: Pandemia, Verde amarillento: Vacunas, Amarillo verdoso: Migración, Amarillo: Salud.

Si prestamos atención en los grupos formados, podemos ver que en PCA y kernel PCA, el tópico de México se encuentra en el centro y sus alrededores los demás cluster. Y para t-SNE podemos observar que los grupos se encuentran más separados entre sí y que se superponen menos. Sin embargo, lo más importante es que podemos llegar a apreciar los subgrupos formados y se diferencian con relativa claridad.

Nuevamente podemos recalcar que la factorización NMF ayudó a dar mejores resultados e interpretabilidad al procesamiento de los datos ya que también nos permitió ver los grupos y los patrones separados con claridad en comparación a cuando usamos SVD, que no pudimos interpretar nada y no observamos algún tipo de patrón.

Sin embargo puede surgir una pregunta, al igual que me surgió a mí mismo, dado que consideramos únicamente 8 tópicos y ya vimos con que podemos relacionarlo, pero ¿qué podemos decir de los demás tópicos? ¿qué temas abordarán? como vimos en el diagrama de codos de la varianza, podemos considerar más tópicos, aunque de manera resumida podrían ser 8 como lo consideramos, pero la verdad es que pueden haber más aunque quizás resulten ser temas menos importantes que los temas principales como salud, economía, apoyos, etc. Con esta duda ampliamos la lista a los primeros 20 temas más abordados por el presidente Andrés Manuel en las conferencias mañaneras basandonos en más componentes

1. **México**
2. **Programas de apoyo del gobierno.**
3. **Tren maya.**
4. **Energía eléctrica.**
5. **Pandemia.**

6. Vacunas.
7. Migración.
8. Salud.
9. Corrupción.
10. Educación.
11. Ex presidentes y personajes corruptos.
12. Seguridad.
13. Industria petrolera (PEMEX).
14. Agua.
15. Aeropuerto Felipe Ángeles.
16. Inflación.
17. Democracia.
18. Combustibles (Gas y gasolina).
19. Estados de México.
20. Derechos humanos.

De estos tópicos podemos comentar algunas cosas interesantes de las conferencias mañaneras. Se habló en gran parte de los programas de apoyo del gobierno, y se enfatizó mucho el tema del tren maya, por encima de temas más importantes como el de educación y salud. Podemos apreciar el impacto de la pandemia reflejada en los discursos mañaneros tomando el lugar número 5 y 6 con vacunas. Es interesante que en la posición 4 se ubique la energía eléctrica. También podemos decir que México presentó un efecto enorme de inmigrantes y que también tuvieron parte en el diálogo matutino. Es curioso que exista un tópico completamente dedicado a ex presidentes y personajes corruptos, ya que también tuvieron gran parte en la conversación de las conferencias. El tema de seguridad se encuentra hasta el puesto número 12, curiosamente, a pesar de haber un incremento de violencia en el país desde el inicio del gobierno del presidente, también podemos mencionar que el tema del agua tiene un puesto en los primeros 20 temas, esto quizás puede deberse a los problemas de sequías a lo que se ha enfrentado el país. Todos estos temas de gran interés representan de manera general los discursos por parte del presidente alrededor de los años de diálogo con la prensa.

- e) A continuación mostraremos la serie de tiempo de cada uno de los ocho tópicos seleccionados para tener un visión más general en el tiempo acerca de los temas.

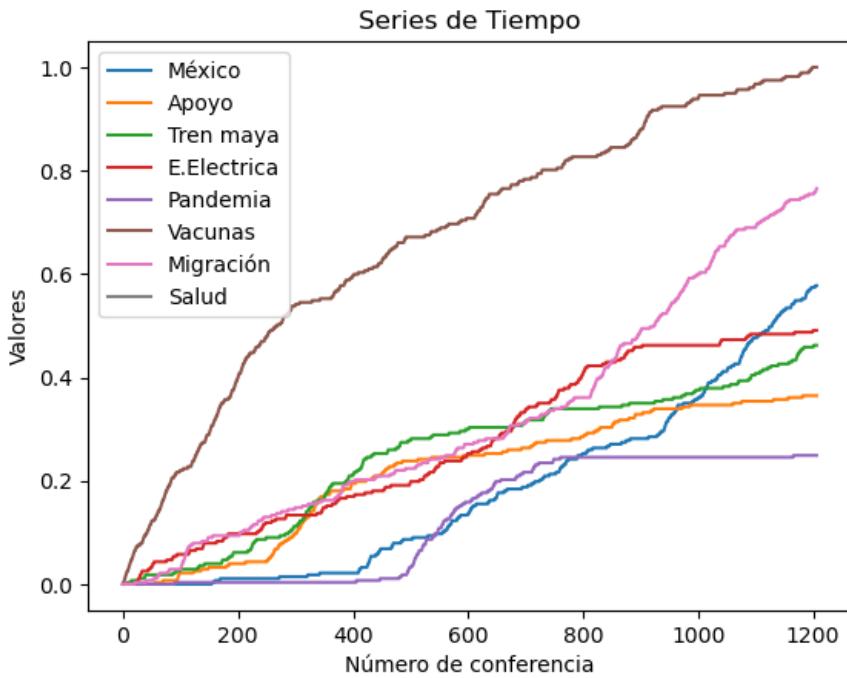


Figura 1.4: Series de tiempo de los tópicos. Para calcular estas series de tiempo se observó la clasificación de cada uno de los reportes de acuerdo a la categoría que le correspondía, de esta forma todas las series de tiempo iniciaban con un valor de cero pero a medida que cada reporte se le asignaba un tópico, se le incrementaba un valor de la unidad a su correspondiente cero, cuando habían series que se les incrementaba un 1, a las series de tiempo de los demás tópicos permanecían con el mismo valor para ese instante, al finalizar, las series de tiempo fueron normalizadas para que estos fueran comparables entre sí.

Como observamos en las series de tiempo, existen series de tiempo que inician en un número de conferencia ya avanzado como es el caso de la pandemia que sabemos que inició dos años después del gobierno del presidente Andrés Manuel, y que después de cierto tiempo se mantiene casi constante debido a que después de pasar todo el tiempo de pandemia y que las vacunas se aplicaran a la población y regresaramos a la normalidad, este tema se tocó muy poco. Sin embargo en ese tiempo de crecimiento, tuvo un crecimiento enorme. Podemos observar que el tema de migración fue un tema constante en todo estos años, siendo su crecimiento la más apegada a la lineal. En cuanto a los temas de tren maya, energía eléctrica, salud y apoyos, se mantuvieron en un mismo valor constante, sin embargo, para el tema de vacunas, podemos observar que tuvo un enorme crecimiento en todo este tiempo.

Tras todo lo comentado en los incisos anteriores podemos decir que este estudio arroja importantes resultados de interés, y nos permitió observar los temas más relevantes de lo que fueron estos casi 6 años de gobierno del presidente. Primeramente en los resultados obtuvimos los primeros 20 temas más tocados en los discursos, ya observamos que de los temas más importantes están los programas de apoyo del gobierno, el tren maya, la energía eléctrica, y la pandemia, etc... podemos destacar también como que la educación no fue un tema tan importante a tocar en comparación de otros temas como el tren maya o la migración, y es quizás por los temas del momento que fueron surgiendo a lo largo del tiempo. Sin embargo también hay un tópico de ex presidentes que resultaron ser más comentados que la seguridad y otros temas más importantes.

En cuanto a la metodología, pudimos observar que tuvimos un problema con la interpretabilidad de la información al utilizar SVD, y no nos permitió ver con claridad en los word clouds los tópicos a los cuales correspondían, y tampoco pudimos ver claramente patrones en la representación de baja dimensión, por otra parte, al usar la factorización NMF pudimos solucionar estos problemas al obtener mayor interpretabilidad, esto nos permitió poder asignar una etiqueta a cada tópico y ver patrones en las representaciones de baja dimensión, de esta forma pudimos notar que esta última resultó ser mejor para este trabajo.

También podemos destacar la importancia de la estructura de los datos para el entrenamiento de un modelo y el análisis de los datos, hacer un buen preprocess de los reportes ayudó a simplificar mucho la información y poder observar los temas de mayor relevancia, aunque este es un tema más general, no podemos dejar pasar que en este trabajo pudimos aprender la importancia de la estructura de los datos. Esto en general permite mejores resultados enfocados al estudio. Aunque habían días que no había un reporte de conferencia mañanera, quizás el presidente pudo haber dejado una persona encargada en su representación para haber tenido un reporte todos los días mientras estuvo ausente. Aunque la basta cantidad de reportes fue suficiente para tener una visión muy general.

