Temas de periódicos mexicanos con LDA

Samuel Salas Meza
Tecnologías para la
Información en Ciencias
UNAM Morelia
salas._@hotmail.com

ABSTRACT

Para este proyecto hice un análisis de los temas más importantes en las noticias que se cubrieron en los periódicos "El Universal", "La Jornada" y "Milenio" entre el 31 de enero y el 20 de febrero. Los datos se obtuvieron con web scraping directamente de los periódicos fuente. Luego quité las palabras vacías (stop words) y obtuve las raíces de las palabras. Después obtuve la cantidad óptima de temas y encontré los temas más importantes en todos los artículos. Al final hice unas visualizaciones y presento los resultados.

ACM Reference Format:
Samuel Salas Meza. 2017. Temas de periódicos mexicanos con LDA. In *Proceedings of Análisis LDA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/LDA

1 DESCRIPCIÓN DE LOS DATOS

El análisis se hizo sobre una base de datos de noticias de "El Universal", "La Jornada" y "Milenio" que obtuve entre el 31 de enero y el 20 de Febrero a través de técnicas de web scraping. Cada registro tiene título de la noticia, resumen, autor, periódico, noticia, fecha y lugar. En "noticia" se guarda el texto principal y es el objeto de análisis. En total hay 2,506 registros en la base de datos ¡pero podría ampliarse fácilmente!

Para publicar en un periódico se sigue un formato general. Eso facilitó el proceso de bajar los datos. Lo único que tuve que hacer es encontrar los tags de cada campo y bajarlos paulatinamente para que no me banearan (como me pasó en el proyecto de Minería de Texto). Para evitar tener registros demasiado malos, establecí algunos controle. El primero fue especificar que si no encontraba un objeto con el tag de la noticia, no lo insertara a mi base de datos. También elegí como llave primaria al título, entonces no hay dos títulos iguales. Cuando no se encontraba un título se guardaba como "sin título". Como no puede haber dos registros con la misma llave primaria (y cuento con que al menos una vez no encontró un título), las noticias sin título no se guardaban. Des esta forma, todos los registros tenían al menos noticia y título. Además tener título como la llave primaria me podía proteger contra notas publicadas en dos periódicos (si se daba el caso). En todos los demás campos fui más laxo y permití tener objetos sin lugar, fecha, autor etc. porque en algunos periódicos no siempre especificaban datos como el lugar, el autor era el periódico o cambiaban el orden de algunos tags (que detenía el scraping si no tenía estas libertades).

Análisis LDA, 03/03/2021, Morelia

ACM ISBN 978-1-4503-1234-5/17/07.
https://doi.org/10.1145/LDA

2 DESCRIPCIÓN DE LA TAREA DE APRENDIZAJE NO SUPERVISADO

Para encontrar los temas usé LDA (Latent Dirichlet Allocation). Este algoritmo está construido bajo el supuesto de que los documentos están conformados por uno o más temas. También asume que cada tema tiene algunas palabras fuertemente relacionadas a él. Es decir, que un subconjunto de las palabras relacionadas al tema T ocurre frecuentemente en un documento del tema T. Así, las palabras P relacionadas al tema T incrementan la probabilidad de que el documento D esté relacionado al tema T. En otras palabras, las palabras tienen pesos que nos ayudan a predecir el tema del documento. Hay algunos otros supuestos como que cada documento puede ser descrito como un conjunto de palabras (bag of words) pero los anteriores son los más importantes.

Este algoritmo trata de encontrar N temas en un conjunto de documentos por medio de funciones de probabilidad. Más concretamente, trata de encontrar la probabilidad de que un documento D pertenezca al tema T y de que la palabra P pertenezca a el tema T. De acuerdo a estos resultados actualiza la probabilidad de que la palabra P pertenezca al tema T multiplicando P(T|D)*P(W|T). De estos resultados va eligiendo el mejor. Al principio todas las palabras se asignan de manera aleatoria, pero conforme pasa el tiempo se van acomodando en un tema que tiene más sentido. Por la manera aleatoria en la que se inicializan los temas, es común obtener diferentes resultados dependiendo de la semilla aleatoria.

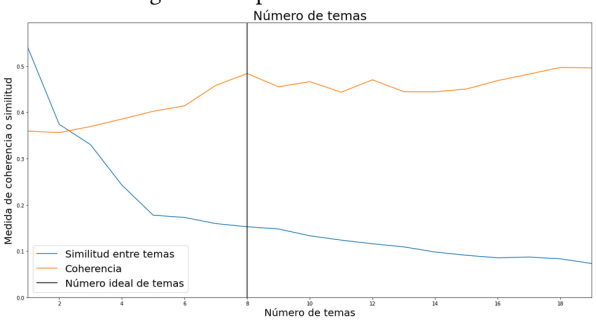
Para elegir el número de temas no encontré información muy contundente, pero sí había varias estrategias que podían usarse. La que más me convenció fue usar dos métricas que en conjunto daban una recomendación del número óptimo de temas. La primera es la similaridad de Jaccard, que básicamente es la intersección de dos conjuntos dividida entre la unión. Esta la usé para encontrar la similitud que existía entre los conjuntos resultantes de usar X y X+1 temas. Se tomaron X y X+1 para poder iniciar el análisis desde 1 tema (que puede servir para corpus "densos") pero tal vez sería más precisio usar X-1 y X. Así se obtiene la "ganancia" al usar X temas. En términos más claros, si encuentro una similaridad muy baja en el índice 8, significa que entre 8 y 9 temas hay mucha diferencia, que es un indicador de una separación "más pura" (particularmente, puede ocurrir cuando el grupo nuevo que se encontró no se parece a los que se habían encontrado o cuando el grupo extra permitió "desambiguar" uno o más grupos resultando en palabras principales son más disímiles). Eso me indicaría que es mejor elegir 9 grupos, pero esa reducción estaría marcada en el índice 8. De cualquier modo, en los resultados el codo está mucho antes y sólo se eligen

más temas en favor de una mejor coherencia. En otras palabras, este cambio no afectaría a los resultados particulares de la práctica.

La otra medida usada, como se mencionó, es la coherencia, que podríamos interpretarla como algo parecido a lo contrario a la perplejidad. Hay muchas maneras de calcularla y aunque revisé las fórmulas, eran demasiadas para entenderlas y elegir la mejor. Entonces decidí esta vez seguir los consejos de los artículos que revisé en vez de fabricar un entendimiento intuitivo. Lo que sí explico es que usé una medida de complejidad llamada cv. Finalmente, para decidir cuántos temas usar tomé el máximo entre:

Coherencia - Similitud - (Número de temas * 0.01)

Agregué ese último término de "número de temas * 0.01" para favorecer tener menor número de temas. Entre más aumenta el número de temas aumenta el número de palabras, entonces, naturalmente, tiende a reducirse el grado de similitud. Es muy difícil que eso no pase, entonces creo que se necesita un control. Eso no venía en ningún lado que leí, pero me pareció que era lo más lógico para balancear el sesgo a más temas. Estos fueron los resultados que me dio esta ecuación al graficarlos para modelos aleatorios con 1-20 temas:



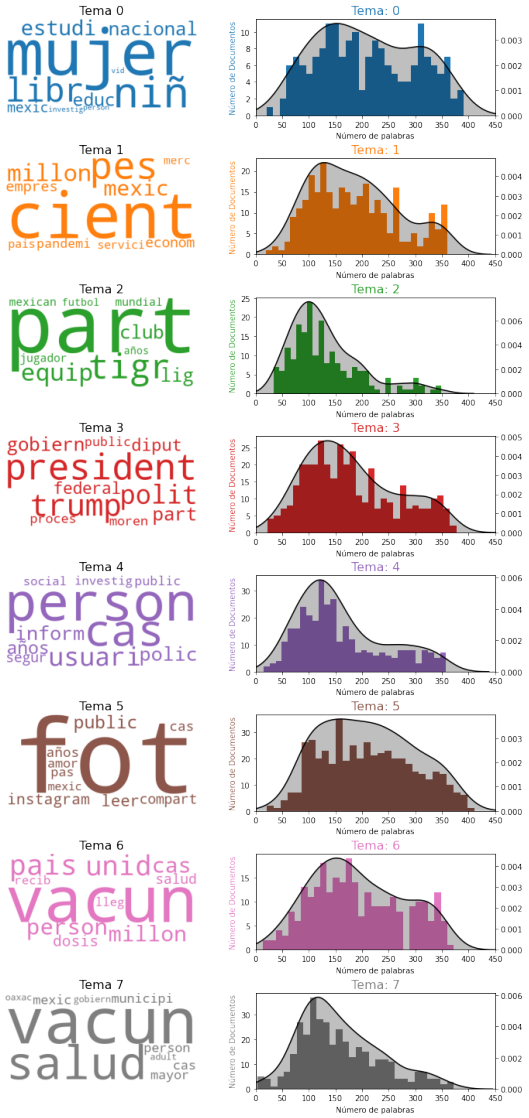
Al final me salieron 8 temas con esa fórmula. El resto de los parámetros no son tan interesantes y por eso los comento únicamente brevemente. Una implementación que me dejaba aprovechar cómputo en paralelo en mi computadora y usé dos núcleos, usé una semilla aleatoria con el número 1, usé 13 pasadas a todo el corpus como recomendación de un artículo de Towards Data Science, y pedí que me regresara una lista de temas con las palabras más probables de cada tema para una visualización.

3 ANÁLISIS Y DISCUSIÓN DE LOS RESULTADOS

El siguiente análisis hace referencia a las gráficas de la derecha. El WordCloud muestra el orden de la relevancia de las palabras con el tamaño, aunque su tamaño no es proporcional al peso. Esas palabras en conjunto (y otras con menor relevancia) forman un tema.

La interpretación de cada tema per sé es algo que no quisiera discutir a detalle porque sería algo muy sesgado a mi propia percepción. Se pueden hacer observaciones generales como que los temas parecen girar en torno a salud ('seguridad'), política y entretenimiento. Más allá de eso, quisiera resaltar algunas peculiaridades de los resultados. Por ejemplo, resulta interesante que los temas 6 y 7 (índices iniciados en 0) tienen 3 palabras importantes en común y una de ellas es la más importante en ambos temas. Lo más probable, de acuerdo al contexto, es que esos dos temas tengan una fuerte relación con la endemia del Covid-19. Lo extraño es que no están agrupados.

Temas y Distribución de palabras (con probabilidad de densidades)

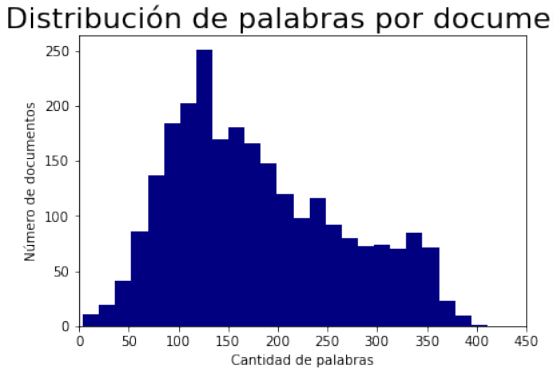


Como nos dice la gráfica de la sección pasada, eso no es ningún accidente. Al contrario, es una decisión que aumenta la coherencia según las métricas de la librería (con la fórmula de coherencia que usé). Algunas posibles explicaciones son que hay tanta información del tema que se alcanza a dividir en subtemas o que las comunalidades son de cierta manera marginales y buena parte del resto de las palabras características son totalmente diferentes.

Para tener palabras más distintivas de esos temas, podríamos seleccionar todos los documentos que sobrepasan cierto umbral de relevancia en los temas relacionados al Coronavirus (esto ya nos lo da LDA). Después podemos remover nuevas palabras vacías definiéndolas como las palabras que ocurren en más del X por ciento de documentos. Por ejemplo si una palabra como vacuna está en

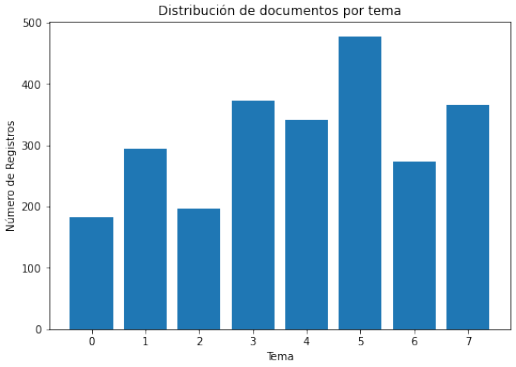
el 80 por ciento de los documentos, aporta poca información y conviene tomarla como palabra vacía. Eso permitiría encontrar un subconjunto de temas más informativo.

Las gráficas de conteo de palabras tienen una línea con una estimación de cómo se vería la distribución probabilística poblacional en cada tema. Como podemos ver hay temas que tienden a ser más efímeros como el 2, que tiene una distribución que favorece menos palabras en los documentos. También hay otros temas con distribuciones más uniformes como el 0. Con esta información podemos hacer hipótesis estructurales de los temas. Por ejemplo, podemos decir que el tema 2, el 4 y el 7 son temas que tienen información más “concreta" o “directa". Eso puede ocurrir de muchas maneras. Por ejemplo, en el caso de los deportes puede que el formato esté muy establecido porque se conocen las cosas relevantes en una noticia de deporte como el marcador, el análisis general y algunos momentos importantes. También puede ser que la información que se aporta sea una actualización a una narrativa mayor y por eso sea más corta. La razón en concreto debería ser analizada con otras herramientas, LDA, en este caso, nos muestra la pregunta y un método de clasificación. Tratar de responder esta pregunta con el número de palabras o el promedio (o mediana) de las probabilidades de las palabras de un tema sería un error porque las distribuciones de palabras en los extremos podrían ser muy variables y poco confiables. En contraste, hay temas con distribuciones más equitativas como el 0 o el 5. Mi hipótesis es que estos temas tienden a tener un formato menos definido y a variar más dependiendo de la noticia. Por ejemplo, parece que el tema 5 está ligado a entretenimiento y es un tema muy variado dependiendo de los eventos más esporádicos. Para analizar eso tal vez se podrían dividir los documentos del tema 0 por conteo de palabras (en rangos) y después se podrían usar otras técnicas incluida LDA para saber de qué hablan los documentos extensos de cierto tema y cómo se relaciona eso a los documentos efímeros del mismo tema. Como referencia, esta es la distribución de palabras de todas las noticias.



Algo que vale la pena mencionar es que algunas palabras hacen evidente que la base de datos se recolectó en un periodo de tiempo corto. Por ejemplo, en el tema 2 aparece la palabra “tigr", que parece hacer alusión al equipo Tigres. Pienso eso porque los Tigres tuvieron un partido internacional muy importante durante el tiempo que se recolectaron los datos. Es un evento aislado con mucho impacto en el futbol que al parecer se filtró en el tema de deportes. Si se recolectaran datos durante algunos años, esperaríamos que los Tigres no siempre estuvieran en primera plana en la sección de deportes (si eso existe). Por esa falta de frecuencia en los documentos, esperaríamos que no aparecieran como palabra clave en el tema de deportes.

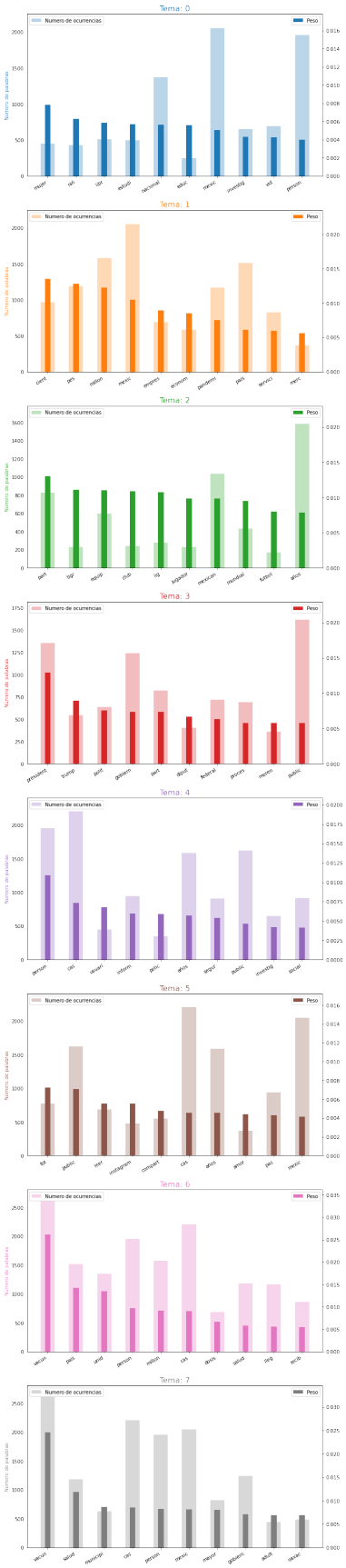
Ahora, tomando una perspectiva más general, podemos analizar cuál es la distribución de la frecuencia de los temas. Para eso se muestra la siguiente gráfica de barras.



Como se puede ver hay una distribución bastante balanceada. La mayoría de los temas oscilan entre los 300 documentos y hay algunas anomalías alrededor de 200 y 500 temas. Creo que esas anomalías se pueden someter a algunos cuestionamientos. Particularmente me surge la duda de si alguno de estos temas se está usando para poner los documentos que no se sabe dónde más colocar. En particular, creo que el tema 2 está bien construido porque tiene sentido que no haya tantas noticias de deportes (especialmente por la pandemia). Particularmente creo que sería productivo analizar el tema 5 porque es el que más documentos tiene y además tiene una distribución uniforme, que en este caso es una anomalía más. Además las palabras que está usando parecen indicar anuncios del estilo “comparte en instagram". Algo de información extra respecto a esta pregunta se puede obtener con la gráfica que se observa en la siguiente página. La palabra “instagram", por ejemplo, está en menos de 500 artículos, similar a la palabra “compart". La palabra foto está en alrededor de 750 artículos. Esto no es suficiente para concluir que todos los artículos que tienen la palabra instagram se acomodan en el tema 5, pero es otro indicador para revisar los artículos de este tema.

Algo extraño es que no hay una indicación clara de que la distribución de número de palabras de un tema tenga un impacto en el peso relativo que se le da a cada palabra. Si hay menos palabras, podría pensarse que cada palabra tiene más peso. O también podría pensarse que en los temas dispersos, hay tantas palabras que las palabras clave sobresalen más y deberían tener mayor peso. Sea cual sea el caso, no parece haber un patrón que indique una u otra dirección (al menos en las 10 palabras más significativas). Por ejemplo, el tema 4 y el 5 son temas con distribuciones de conteo de palabras diferentes, pero no existe una diferencia significativa en la distribución de los pesos en las palabras más relevantes. Para hacer una conclusión más acertada respecto a esto, se necesita hacer un análisis más profundo.

También se puede observar que el peso de las palabras no sigue la misma distribución que su ocurrencia. Hay temas cuya palabra

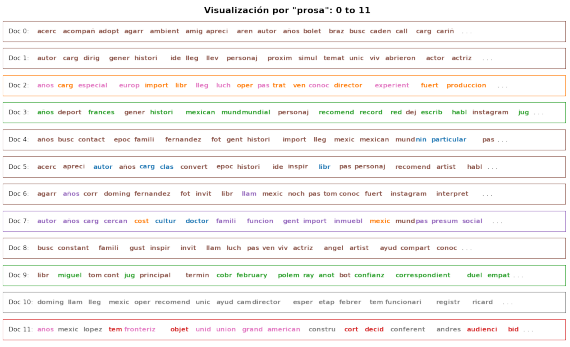


más representativa también tiene muchas ocurrencias y también hay algunos casos donde la palabra más representativa tiene pocas ocurrencias. La información extrapolada de los temas es poca como para hacer un análisis confiable de este tipo, pero al menos en este corpus, podemos notar que el número de ocurrencias no está fuertemente ligado con el tema cuando se trata de las palabras más importantes para un tema. Por ejemplo, la palabra futbol aparece alrededor de 200 veces en el tema 2. Para ponerlo en perspectiva, ese es el número de documentos clasificados en el número 2. La distribución de palabras por documento no necesariamente es equitativa, pero parece que está en los suficientes documentos para ser una palabra con mucho peso para el tema 2. Sin embargo, no es la palabra con mayor peso en el análisis, lo cual es sorprendente ya que es poco probable encontrar esa palabra en algún otro tema.

El tema 2 es un caso interesante porque la palabra con más ocurrencias tiene alrededor de 1,600 ocurrencias. Entre los máximos de ocurrencias, este tema tiene el mínimo (o segundo mínimo) entre las 10 palabras más significativas. Además, es el tema con la menor media de la frecuencia entre las 10 palabras con mayor peso. Yo interpreto que eso quiere decir que el lenguaje en ese tema es más especializado. A lo que me refiero es que ese tipo de palabras se usan, en el contexto de noticias, generalmente sólo en temas de deportes. Esta es información importante porque reglas de este tipo facilitan mucho más la catalogación de documentos. En un caso ideal, si supiéramos que algunas palabras sólo se presentan en un tema en particular, sólo tendríamos que buscar esas palabras para identificar el tema. Extendiendo esa idea, nuestro objetivo ideal es sería los temas con palabras características que forman reglas de asociación (tal vez probabilísticas). Aunque estas reglas serían fáciles de romper si se tiene la intención, se esperaría que en el contexto adecuado fueran muy útiles y elucidantes.

Estas gráficas muestran evidencia más contundente de que hay que hacer un análisis de temas de Covid-19 quitando 'palabras vacías especiales al tema'. Se observa que el peso que se otorga al tema 6 y 7 para 'vacuna' es alto en ambos casos, pero que también su frecuencia lo es. Esta información apunta a que la palabra es importante para ambos temas, pero es tan frecuente que no sirve para catalogar documentos como una 'palabra característica'.

Por último mostraré cómo se ve el proceso de LDA para los documentos



El rectángulo indica el color del tema. El color de las palabras indica el tema más relacionado con esa palabra. Esta visualización tiene el propósito de hacer evidente que no todas las palabras tienen que estar fuertemente relacionadas al tema asignado al documento

para que el documento tenga la asignación que tiene. Eso era esperado, lo que esta herramienta permite es analizar los componentes de un texto de manera sencilla y encontrar anomalías. Por ejemplo, es extraño que la palabra 'deport' tenga una asignación al tema 5. Yo interpreto ese tema como más ligado al entretenimiento, y no sería descabellado hacer esa asignación si no hubiera un tema con palabras más fuertemente ligadas a deporte como las del tema 2. Para llevar la narrativa de esta sección a una narrativa circular, creo que esta puede ser evidencia de que las computadoras procesan las cosas de diferente manera que los humanos. Si este tipo de ejemplos, donde la intuición humana falla, se repiten mucho debemos tener mucho cuidado haciendo interpretaciones en los temas per sé. También puede ser evidencia de que pudo faltar hacer más iteraciones para hallar una asignación más coherente.

4 CONCLUSIONES

Los temas que se encontraron en las noticias en las fechas rescatadas son en general de salud (seguridad), política, y entretenimiento. Las particularidades de cómo se encontraron esos temas y qué características estructurales tiene cada uno de ellos nos hablan de información adicional acerca de los temas y de LDA en general. Por ejemplo, sabemos que hay temas que tienden a ser más cortos que otros. También existen temas que tienen palabras en común y que se podrían conocer mejor si se eliminaran palabras que brindan poca información en el tema (que son las que tienen demasiada frecuencia en más de un tema). LDA tiene una búsqueda de palabras que no necesariamente es intuitiva y hay que tener cuidado al interpretar los temas. Probablemente la mejor opción es usar los temas para responder una pregunta o como alimento para un algoritmo que nos pueda ayudar a responder una pregunta.

Aun así quedaron algunos experimentos por correr. El más claro, y en el estado actual es el de separar los temas 6 y 7 para ver qué tipo de cuestiones se resaltan cuando se habla de Coronavirus. Me pareció interesante que LDA encontró más coherente usar dos temas relacionados a Covid-19. También sería conveniente tomar más datos para que los temas resultantes fueran un poco más generales. El análisis toma más en cuenta las noticias relevantes durante sólo 3 semanas, así que hay una contradicción entre usar un algoritmo que encuentra las generalidades en información muy "específica". Correr ese experimento queda pendiente para un experimento posterior. También sería interesante encontrar de manera estocástica si realmente hay una relación entre los lenguajes nicho o especializados a partir de la frecuencia con la que se usan las palabras más importantes. Para eso necesitaríamos correr varios experimentos con múltiples tipos de corpus y encontrar la relación entre la presencia de palabras y la asignación del documento a un tema.

Después de hacer esta práctica me quedaron algunas dudas que me permitirían entender más los resultados de LDA. La primera es ¿cómo se mide la novedad de una noticia? Si obtenemos una noticia con temas que nunca se habían visitado antes ¿qué pasaría La respuesta más aparente es que esas noticias tendrían palabras que el modelo no las ha visto antes. No sé si es posible que un texto sea rico en lenguaje pero pobre en temas, y resulte en una asignación errónea de nuevas noticias. Tal vez algunas otras ideas podrían ser buscar si las palabras que tiene el texto no poseen fuerte inclinación a un tema. Es decir, que encontró una asignación al tema T a través de palabras precariamente relacionadas a él. Otra puede ser buscar si la predicción del texto nos da una distribución de muchos temas como producto de no saber dónde colocar el tema. Si podemos responder esta pregunta, podemos medir qué tan novedosa es la noticia que se quiere publicar dadas las noticias que se han publicado! Otra pregunta que me quedó es si la distribución de número de documentos se asemejaba a la distribución de aportación de temas en el Corpus. Es decir, si la suma de las probabilidades de aportación de un tema a través de todos los documentos es similar al número de documentos de ese tema. Esta pregunta la respondí en la visualización de Streamlit y resulta ser que, para este caso, sí. Además muestro un seleccionador aleatorio de noticias por tema.

REFERENCIAS

Amos D. (Ago 17, 2020): A Practical Introduction to Web Scraping in Python. Obtenido el 13 de febrero del 2020 en:
https://realpython.com/python-web-scraping-practical-introduction/

Bhatt B. (Jun 22, 2018): Intuition behind Latent Dirichlet Allocation (LDA) for Topic Modeling. Obtenido el 14 de Febrero en:
https://www.youtube.com/watch?v=Cpt97BpI-t4t=1s

Breuss M. (NA): Beautiful Soup: Build a Web Scraper With Python. Obtenido el 10 de febrero en:
https://realpython.com/beautiful-soup-web-scraper-python/

Kapadia S. (Ago 19, 2019): Evaluate Topic Models: Latent Dirichlet Allocation (LDA). Obtenido el 16 de febrero en:
https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

Kekre S. (NA): Create Interactive Dashboards with Streamlit and Python. Obtenido el 1 de marzo en:
https://www.coursera.org/learn/interactive-dashboards-streamlit-python

Kho J. (Sep 26, 2018): How to Web Scrape with Python in 4 Minutes. Obtenido el 10 de febrero en:
https://towardsdatascience.com/how-to-web-scrape-with-python-in-4-minutes-bc49186a8460

Kulshrestha R. (Jul 19, 2019): A Beginner's Guide to Latent Dirichlet Allocation(LDA). Obtenido el 15 de febrero en:
https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2

Salah M. (Jul 4, 2020): What is the formula for cv coherence?. Obtenido el 16 de febrero en:
https://stats.stackexchange.com/questions/406216/what-is-the-formula-for-c-v-coherence

Serrano L. (Mar 18, 2020): Latent Dirichlet Allocation (Part 1 of 2). Obtenido el 13 de febrero en:
https://www.youtube.com/watch?v=T05t-SqKArY

user12446118 (Oct 15, 2020): What is the best way to obtain the optimal number of topics for a LDA-Model using Gensim?. Obtenido el 16 de febrero en:
https://stackoverflow.com/questions/32313062/what-is-the-best-way-to-obtain-the-optimal-number-of-topics-for-a-lda-model-usin

Zhao A. (Ene 5, 2019): Natural Language Processing (Part 5): Topic Modeling with Latent Dirichlet Allocation in Python. Obtenido el 14 de febrero en:
https://www.youtube.com/watch?v=NYkbqzTlW3wt=634s