Cantidad de Capítulos por cada tomo

Tomo 1: 15

Tomo 2: 7

Tomo 3: 15

Tomo 4: 28

Tomo 5: 28

Tomo 6: 408

Tomo 7: 6

Tomo 8: 15

Tomo 9: 16

Tomo 10: 13

Tomo 11: 18

Tomo 12: 12

***Resultados***

Nota: En todos los tomos utilizaremos bigrams para realizar el topic modelling, para ver mejor los tópicos se recomienda analizar con el landa del gráfico en 0, 0.5 y 1.

***Tomo 1:***

No se obtuvo resultados ya que los bigramas, trigramas y quintigrams lanza el siguiente error:

ValueError: [E088] Text of length 2414997 exceeds maximum of 1000000. The v2.x parser and NER models require roughly 1GB of temporary memory per 100,000 characters in the input. This means long texts may cause memory allocation errors. If you're not using the parser or NER, it's probably safe to increase the `nlp.max\_length` limit. The limit is in number of characters, so you can check whether your inputs are too long by checking `len(text)`.

probé con valores mayores al estándar de min\_count=5, threshold=10. Esos valores fueron:

Min\_Count = 10, 15, 50

Threshold = 10, 50, 150, 500

sin embargo sigue saliendo el mismo error y por falta de tiempo pase a los que se pudieron procesar.

***Tomo 2:***

Modelo LDA:

Se detecto 3 grandes temas cual contrasta los 7, podemos inducir que hay algunos capítulos que hablan del mismo tema.

El valor de coherencia es de 0.31207689360862806 muy bajo para un volumen de datos muy grande, mientras que el valor de Perplexity es de -8.46033110794157 valor que tomamos como aceptable por ser muy bajo. Este modelo logró detectar los 3 temas centrales del documento:

Tema 1: Orden de Marcha del Ejército Argentino y Carta de acción de Güemes

Inferimos que el tema 1 que logró separar la herramienta habla sobres las órdenes de marcha del ejército argentino a cargo de los diferentes oficiales, además del envío de Güemes a Buenos Aires por un supuesto amorío, y los posibles enfrentamientos que se realizará con el ejército enemigo.

Tema 2: Elección de Gobernador

Podemos inferir que este tema habla de la elección de gobernador que se realiza en el puesto de Marqués, el nombramiento de Güemes como gobernador de la intendencia de Salta y además de algunos datos sobre la esposa de Güemes.

Tema 3: Rastreo del ejército enemigo

Este tema aborda sobre los movimientos del enemigo que tiene en la frontera norte, además de anoticiar los enfrentamientos al Gral San Martin

Como se puede observar en la gráfica de los tópicos el tema 1 y 3 están muy relacionados, esto se debe a que ambos temas se habla muy constantemente del ejército enemigo, mientras que el tema 2 está muy alejado por hablarse de la elección del Gobernador de la provincia de Salta y al no mencionar nada sobre movimiento militar el tema se distingue de los restantes.

Modelo LDA con Mallet

Este modelo logra identificar 5 temas, cual mejora los 3 ya obtenidos con el modelo LDA solo, estos temas son:

Tema 1: Pago del sueldo adeudo al Gral Güemes

Tema 2: Movimiento del ejército enemigo

Tema 3: Sanción de Güemes por supuesto amorío

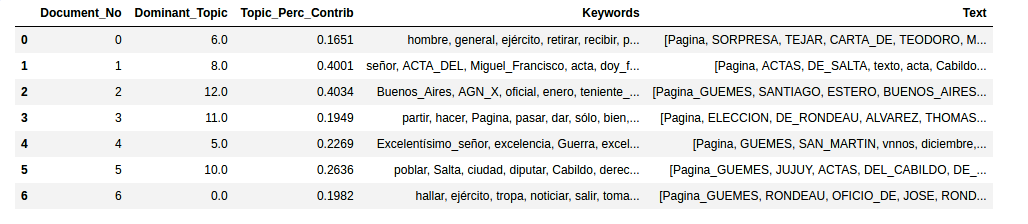
Tema 4: Elección de Gobernador de la provincia de Salta

Tema 5: Documentos de la elección de gobernador (Güemes-Jujuy)

Podemos observar que este modelo logra separar en más temas al tomo, logrando asi de a poco alcanzar con la cantidad de capítulos que hay, pero sin embargo sigue reconociendo que varios capítulos hablan de un tema en particular.

Su valor de coherencia del modelo es 0.43325061622802996 subiendo así un 0.1211737226194019 del valor de coherencia del modelo anterior.

Observando el gráfico de los valores de coherencia con respecto a la cantidad de tópicos podemos observar que la mejor cantidad de tópico es 7 con un valor de coherencia de 0.4432, que es igual a la cantidad de capítulos que hay en el tomo 2.



Modelo TF-IDF

Usando este modelo para contrastar con los anteriores podemos observar una mejora, pues su valor de coherencia es de 0.5575834243743936, mejorando así un 0.1243328081463636 con respecto al modelo LDA con Mallet y 0.2455065307657655 con respecto al modelo de LDA, además su valor de Perplexity es de -12.986508495848403.

**Tomo 3:**

Modelo LDA:

En este modelo se detecto 5 temas constractando así a gran medida de los 15 capítulos que hay en el tomo. Su Coherencia es bajo pues tiene un valor de 0.28034722953751257, y su Perplexity es de -8.625640626168872, valor muy parecido con el gd2.

Los 5 grandes temas centrales que detectó la herramienta son:

Tema 1: Batallas, Auxilio e Independencia.

Inducimos que habla sobre la batalla de Venta y Media, además de las cartas con relación de Güemes French y Rondeau, además de la jura de Güemes de la independencia y algunos documentos sobre la monarquía Inca.

Tema 2: Sorpresa, Derrota y Ayuda.

Este tema relata sobre la batalla de sipe sipe y su derrota, además de la sorpresa enemiga sufrida al teniente Ignacio Regueral cerca de Tupiza, y la disposición de Güemes a socorrer al ejército auxiliar derrotado del alto Perú.

Tema 3: Retirada, Entrevista y Persecución a Camargo.

Se desarrolla la retirada del ejército patriota desde el punto de vista de los realista, las actas redactadas en el cabildo de Jujuy, la persecución y muerte del oficial realista Camargo en Las Lomas cerca de la municipalidad de Cotagaita, además de la entrevista realizada entre Güemes y Pueyrredón y la relación entre Güemes y el soberano congreso del Tucuman.

Tema 4: Ataque a Salta y Pacto de San José.

Este tema desarrolla el ataque (sin llegar a desarrollar acciones bélicas) de las tropas del ejército auxiliar del alto Perú comandadas por Rondeau a Salta, culminando en el pacto de San José de Cerrillos impidiendo así una guerra civil y posibilitando la gesta libertadora del Gral San Martin.

Tema 5: Estado de la caja principal de Salta.

En la gráfica observamos que los temas 1 y 3 están muy relacionados (siendo el primer tema con más peso), esto se debe a que se desarrollan las acciones bélicas en la misma zona, además que los principales actores son el ejército del virreinato y el ejército auxiliar del Perú y lo que desencadenó estas acciones belicas. El tema 2, se puede divisar que está muy alejado de estos dos temas anteriores, pues se habla de la retirada, caza y entrevista que ocurrieron después de la pérdida de la batalla de Sipe Sipe. El tema 4 también está muy cercano ya que se trata de un ataque entre camaradas y el actor principal es el ejército del alto Perú. Sin embargo el tema que más se diferencia de todo el resto es el estado de la caja (tema 5) ya que no tiene nada que ver con acciones bélicas.

Modelo LDA con Mallet

Este modelo también logra detectar 5 temas centrales de cuales podemos inducir:

Tema 1: Disposición de Güemes y Derrota Sipe Sipe.

Tema 2: Güemes jura la Independencia.

Tema 3: Ataque a Salta

Tema 4: Retirada del ejército auxiliar del Perú (punto vista realista)

Tema 5: Estado de la caja de la provincia de Salta

Sin embargo este modelo no tiene una mejora significativa, porque sigue dividiendo en 5 temas al tomo. Su valor de coherencia es 0.3672328087094547, 0.08688557917194212 más que el LDA.

Si observamos el gráfico de coherencia y tópicos podemos observar que el mejor modelo es de 18 tópicos pues tiene un valor de coherencia de 0.4128. Esto quiere decir que algún capítulo puede hablar más de un tema, pues contamos con solo 15 capítulos, para ello entonces analizaremos capitulo por capitulo para saber cuánto temas se habla en cada uno. Para ello armamos el siguiente cuadro:

|  | LDA | | LDA Mallet | | Mejor Modelo | | Mejora | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Capítulo - Página | Tópico | Coherencia | Tópico | Coherencia | Tópico | Coherencia | LDA | LDA Mallet |
| 22 - 116 | 1 | 0,27321 | 5 | 0,2375 | 3 | 0,2692 | -0.00401 | 0.0317 |
| 30 - 69 | 1 | 0,2218 | 5 | 0,2023 | 3 | 0,2428 | 0.021 | 0.0405 |
| 23 - 59 | 1 | 0,2636 | 5 | 0,21278 | 2 | 0,2945 | 0.0309 | 0.08172 |
| 35 - 37 | 1 | 0,2122 | 5 | 0,197 | 7 | 0,2453 | -0.0143 | 0.0483 |
| 36 - 34 | 1 | 0,2536 | 5 | 0,26721 | 6 | 0,2393 | 0.016771 | -0.02791 |
| 24 - 36 | 1 | 0,253229 | 5 | 0,32351126 | 3 | 0.27 | -0.02469 | -0.053511 |
| 34 - 31 | 1 | 0,27519 | 5 | 0,21239 | 3 | 0,2505 | -0.0039 | 0.03811 |
| 31 - 24 | 1 | 0,2094 | 5 | 0,208595 | 6 | 0,2055 | 0.124267 | -0.003095 |
| 27 - 19 | 1 | 0,174933 | 5 | 0,293781 | 2 | 0,2992 | -0.0039 | 0.005419 |
| 28 - 18 | 1 | 0,292195 | 5 | 0,2494271 | 2 | 0,2596 | -0.032595 | 0.0101729 |
| 25 - 11 | 1 | 0,293796 | 5 | 0,264633 | 3 | 0,2941 | 0.000304 | 0.029467 |
| 33 - 10 | 1 | 0,2336116 | 5 | 0,306627 | 2 | 0,3132 | 0.0795884 | 0.006573 |
| 32 - 8 | 1 | 0,4060186 | 5 | 0,329004 | 2 | 0,3889 | -0.0171186 | 0.59896 |
| 26 - 4 | 1 | 0,292684 | 5 | 0,2686618 | 6 | 0,3011 | 0.008416 | 0.0032438 |
| 29 - 3 | 1 | 0,234684 | 5 | 0,27126499 | 6 | 0,3 | 0.065316 | 0.02873501 |

El cuadro ordenamos por la cantidad de página que contiene cada capítulo. Como se puede notar LDA siempre lo clasificó con un solo tema central entonces no podemos tomarlo como medida, lo que sí podemos tomar como medida es su valor de coherencia, luego ver si mejora este modelo aplicando LDA Millet. Este caso se dieron en los capítulos 23, 27 y 33, pues parte de un modelo LDA con coherencia pequeña y a medida hay más tópicos podemos notar el aumento significativo del mismo, además el incremento de los tema es coherente pues ocurren casos como el capítulo 35 que si bien aumenta significativamente el valor de coherencia no es posible hablar que aumente 7 temas ya que sería ilógico pensar que se agreguen 6 temas más a los 15 ya existente y por lo tanto el mejor modelo sería de 21 temas y no de 18.

Si tomamos en cuenta los capítulos 23, 27 y 33 añaden al modelo 1 tema más cada uno, con esto ya tenemos los 18 temas que hay en el mejor modelo, y su nivel de coherencia aumenta significativamente. En el siguiente cuadro, se ponen los perplexity del modelo LDA y la coherencia del modelo TF-IFD.

| Capítulo - Página | LDA  Perplexity | TF-IDF  Perplexity | TF-IDF  Coherencia |
| --- | --- | --- | --- |
| 22 - 116 | -8.0963 | NaN | 0.811509 |
| 30 - 69 | -8.19429 | NaN | 0.765289 |
| 23 - 59 | -8.1015 | NaN | 0.71273 |
| 35 - 37 | -8.0035 | NaN | 0.758864 |
| 36 - 34 | -8.14989 | NaN | 0.6974325 |
| 24 - 36 | -7.7149049 | NaN | 0.6994225 |
| 34 - 31 | -7.811606 | NaN | 0.7641423 |
| 31 - 24 | -7.636234 | NaN | 0.703706 |
| 27 - 19 | -7.9498034 | NaN | 0.7039213 |
| 28 - 18 | -7.44701627 | NaN | 0.6770703 |
| 25 - 11 | -7.10492856 | NaN | 0.4722702 |
| 33 - 10 | -7.4781673 | NaN | 0.6212791 |
| 32 - 8 | -7.3914518 | NaN | 0.5585912 |
| 26 - 4 | -6.8780568 | NaN | 0.4546569 |
| 29 - 3 | -6.48035866 | NaN | 0.3370856 |

Nota: Un caso muy peculiar es el capitulo 24 pues tiene una mejora casi lineal a medida que se van ingresando nuevos temas.

Modelo TF-IDF:

Tomando este mejor modelo entonces podemos contrastar con el modelo TD-IDF, esté presenta una mejor grandísima pues su valor de coherencia es de 0.5230675568013615, presenta una mejora de 0.24272032726384896 con respecto al LDA y 0.15583474809190684 respecto al LDA Millet ademas su perplexity es de -12.880845282972682. El valor más chico visto hasta ahora.

***Tomo 4:***

Modelo LDA:

En este modelo se detectaron 5 tópicos, un valor lejos de los 28 capítulos que existen en el tomo. Su concordancia es 0.409457110066567336 y Perplexity es -8.31625008482893. Los 4 temas que detectó la herramienta son:

Tema 1: Es el tópico con mayor cantidad de documentos, parece hablar que desde Salta enviaron tropas a marchar a Jujuy tras ocupación de realistas

Tema 2: Este tema trata del encuentro en el Cabildo de Salta y nombramiento interino de Juan Manuel Quiroz.

Tema 3: Su contenido se relaciona cartas al Ejército Auxiliar y la bandera enemiga

Tema 4: Trata sobre la situación financiera y contable de Salta

Los temas 3 y 4 están un poco cerca en la gráfica de tópicos, aun así tratan diferentes temas

Modelo LDA con Mallet:

Se obtiene 5 tópicos al igual que el modelo LDA. Los temas del modelo son:

Tema 1: Avance de realistas en Salta y Jujuy

Tema 2: Belgrano manda a combatir a de La Serna..

Tema 3: Retirada de enemigo.

Tema 4: Medidas de Güemes.

Tema 5: Actas del Cabildo y estado de caja Salta.

Su valor de coherencia es: 0.410436583033346

Se obtiene el valor óptimo de coherencia igual a 0.4353 con 21 tópicos, por lo que hay capítulos que hablan del mismo tema. Con 28 tópicos, que es la cantidad de capítulos que hay en el tomo, el valor de coherencia 0.4125 es más bajo al óptimo pero superior al obtenido de 5 tópicos

Modelo TF-IDF

Con este modelo se obtiene un valor de coherencia 0.5364352818193469, el cual es mejor a los modelos LDA y LDA Mallet. Su valor de Perplexity es -12.376424004852517 mejora el modelo LDA.

**Tomo 5:**

Modelo LDA:

Este modelo distingue 4 tópicos, pero los tópicos 1 y 2 comparten un conjunto de palabras en común al hablar del enemigo.

En este modelo se detectaron 5 temas:

Tema1: Belgrano auxilia a Güemes ante avance enemigo

Tema 2: Estado de la caja principal de Salta y el pedido de fondos para auxiliar la Mision de Zenta

Tema 3: Parece hablar de actas del Cabildo

Tema 4: Enemigos se rinde y son prisioneros, Belgrano protesta maltrato hacia prisioneros.

Las medidas que se obtuvieron son Perplexity: -8.433269846654406 y el valor de coherencia: 0.3612975135493436

Modelo LDA Mallet:

Valor de coherencia obtenido: 0.44068516687286585

Tema 1: Nombramiento de escribano Molina

Tema 2: Taboada viaja a campo enemigo por orden de Güemes

Tema 3: Situación frente al enemigo

Tema 4: Oficio de Güemes y sus oficiales

Tema 5: Actas del cabildo de Salta

El modelo óptimo considera 17 tópicos con un valor de coherencia 0.5016 superior al del modelo LDA. Si se consideran 28 temas su valor de coherencia es 0.4606 lo que da indicio que efectivamente algunos capítulos del tomo tratan el mismo tema

Modelo TF-IDF

Se obtienen las medidas Perplexity: -12.347042946357522 y valor de coherencia: 0.3798100084076902

***Tomo 6:***

Este tomo es muy interesante pues no posee capítulos, sino que es una colección de 408 cartas ordenadas de manera cronológicas. Entonces podemos decir que el Topic Model nos arrajara los temas que se llegan a tocar en el conjunto de estas cartas.

Modelo LDA:

El LDA logró detectar 5 temas, cuyo valor de coherencia es de 0.3126114892797737 y su perplexity es de -8.545542533183822, siendo muy similar a los casos anteriores.

Los temas son:

Tema 1: Carta sobre movimiento enemigo y marcha del ejército patriota.

Tema 2: Cartas entre Güemes y Belgrano.

Tema 3: Cartas de Güemes enviadas a varios próceres.

Tema 4: Cartas Güemes y la Asamblea

Tema 5: Cartas Güemes y Olañeta

Modelo LDA con Mallet:

El modelo LDA Mallet logró detectar también 5 temas, con un valor de coherencia de 0.3283850671423107 es decir que hay una mejora de 0.01577357786253697 que es muy pequeña. Cuyo temas son:

Tema 1: Cartas entre Güemes con el Gobierno Tucuman

Tema 2: Cartas sobre marcha del ejército patriota

Tema 3: Cartas órdenes a Güemes

Tema 4: Cartas entre Güemes y Belgrano

Tema 5: Cartas sobre sobre movimiento enemigos.

Podemos intuir entonces que ya el modelo de LDA es bastante bueno para empezar, y además que este modelo no incluye una mejora muy importante. Si observamos el gráfico de coherencia y tópicos podemos notar que el mejor modelo es de 28 temas con una coherencia de 0.3648 pero hay un modelo de 3 temas con una coherencia de 0.3641, es decir, que difiere con un valor de 0.0007 es muy pequeña esta diferencia, por lo cual si bien no tiene la mayor coherencia podemos decir que con tres tópicos bastaría para englobar esta colección de carta, pero hay peligro al tener solo 3 temas podemos perder precisión. Otro valor razonable parece ser 8 temas pues su coherencia es de 0.3417, es decir, 0.0231 menor que el óptimo, este a mi parecer nos permitirá tener una cantidad adecuada de tópicos con relación al volumen de datos, sin perder mucha precisión y englobar documentos en temas que no están muy relacionados, cosa que pasó en los anteriores gd en el modelo LDA.

Modelo TF-IDF:

Este modelo logra tener una mejoría , pues su valor de coherencia es de 0.4127538557693985, 0.10014236648962477 mayor al modelo de LDA y 0.07105385576939849 al seleccionado (8 tópicos) pero su valor de perplexity es de -10.908294913237654 no tan bajo como los anteriores.