# Capítulo 1: metodología

Antes que nada voy a resumir muy rápidamente todo lo que aconteció durante la investigación. Descargué los PDFs de TESIUNAM, la base de datos de la UNAM que alberga y hace disponibles los trabajos de titulación de toda la universidad. Para hacerlo tuve que escribir unos scripts en Python ([https://github.com/diego-g-fonte/crawler](https://github.com/diego-g-fonte/crawler%5B^s.I.1%5D)),[[1]](#footnote-2) que descargan todos los PDFs de todas las páginas de resultados de cualquier búsqueda en TESIUNAM, además de producir un documento JSON con los metadatos de dichos resultados (nombre de lx autorx, título, año de publicación y número de identificación del PDF asignado por TESIUNAM).

Para cuantificar el estilo usé una estrategia utilizada en los estudios de atribución autoral que está basada en las frecuencias relativas de las palabras. Frecuencia relativa sólo quiere decir que del 100% de las palabras de un texto, X% es la palabra "de", Y% es la palabra "la", etcétera. Por ejemplo, en la tesina titulada "El género policíaco y el gótico dentro de *The Virgin Suicides* de Jeffrey Eugenides", la palabra "de" conforma el 4.16% de las palabras, "la" el 3.29%, etcétera. Generé una tabla con las 150 palabras más comunes en todos los trabajos de titulación y su frecuencia relativa en cada uno. En estilometría, las frecuencias relativas de las palabras más comunes son lo que conforma el estilo de un texto. Como escribe Matthew Jockers en su capítulo "Style":

The working assumption of all such investigation is that writers have distinct and detectabale stylistic habits, or "tics."

. . .

It is exactly these subtle "features" (pronouns, articles, conjunctions, and the like), however, that authorship and stylometry researchers have discovered to be the most telling when it comes to revealing an author's individual style.

. . .

There are many ways in which a feature set may be derived, but the growing consensus is that the analysis of high-frequency words (mostly function, or closed-class, words) or n-grams[[2]](#footnote-3), or both, provides the most consistently reliable results in authorship attribution problems. (Jockers 63-4)

Ya teniendo la tabla con las frecuencias relativas, usé una función ("dist" en R[[3]](#footnote-4)) que crea un espacio multidimensional (en este caso de 150 dimensiones) y coloca un punto por cada trabajo de titulación en ese espacio. Es más fácil pensar en un espacio de dos dimensiones en el que el eje X es la frecuencia relativa de la palabra "de", y el eje Y la de la palabra "la". Dentro de este espacio, el trabajo "El género policíaco ..." se convertiría en un punto con las coordenadas (4.16, 3.29). Otro trabajo de titulación, por ejemplo "Propuesta de lecciones-muestra para la enseñanza de la comprensión de lectura en inglés en el marco de la ENA", tendría las coordenadas (7.11, 4.01). Podemos trazar una línea entre esos dos puntos y medir qué tan cerca o lejos están (en otras palabras, qué tan similares o diferentes son). La computadora puede hacer ese mismo proceso en 150 dimensiones y calcular las distancias entre todos los puntos. Finalmente, generé una gráfica que visualiza el promedio de las distancias entre los trabajos de titulación publicados en un año, de los años 2006 al 2023.

Para cuantificar los temas utilicé una técnica llamada modelado de temas. Usé un programa (MALLET[[4]](#footnote-5)) que, al procesar los textos, produce una tabla con una cantidad X de temas (elegí 50) y la cantidad de ese tema que tiene cada texto. Lo podemos pensar como que cada texto tiene un 100% de "tema", del cual 50% puede ser Jane Austen, 25% narratología, etcétera. Con esa tabla usé el mismo método que con la tabla de frecuencias relativas y generé una tabla similar que visualiza el promedio de las distancias entre los trabajos de titulación, pero ahora tomando en cuenta los temas.

# 1. ¿Cómo conseguí los datos?

# Más arriba dije que los scripts con los que descargué los PDFs generan un documento JSON. A continuación voy a explicar qué es un documento JSON y cómo funcionan, a grandes rasgos, los scripts que escribí.

JSON son las siglas de JavaScript Object Notation. JavaScript es un lenguaje de programación, y JSON es una forma de escribir documentos para transportar información fácilmente (que surgió de la forma en la que se usan los objetos en JavaScript[[5]](#footnote-6)). Lo bonito del formato JSON es que es muy fácil de leer, tanto para las personas como para las computadoras, lo que permite transportar información de un lenguaje de programación a otro y además modificar manualmente los datos. Una entrada de los metadatos de un trabajo de titulación se ve así:

{"nombre": "Torres Gutiérrez, Cristian Miguel",

"titulo": ""A speaking eye" : el discurso pictórico en Hero and leander de Christopher Marlowe",

"año": "2017",

"doc\_num": "0755725"}.[[6]](#footnote-7)

Como es evidente, es un formato muy útil para organizar bases de datos.

1.1. ¿Cómo conseguí los metadatos?

Mis scripts para los metadatos (<https://github.com/diego-g-fonte/crawler/tree/main/metadata>) funcionan de la siguiente manera. Primero hay que, desde nuestro navegador web, hacer una búsqueda en TESIUNAM. Yo evidentemente busqué "licenciatura en lengua y literaturas modernas inglesas", pero mi código funciona con cualquier búsqueda. Hay que copiar y pegar el URL de la primera página de resultados en el archivo que llamé "url\_manual". Después de eso, hay que correr los scripts llamados "1.py" y "0.2.py". Estos generan el archivo JSON con los metadatos. Mi código no es perfecto, entonces, después de que los scripts lo hayan escrito, hay que retocar el JSON manualmente un poco. Con eso hecho, se tiene una base de datos con los metadatos de todos los trabajos de titulación de letras inglesas (o de todos los trabajos de titulación que haya arrojado TESIUNAM en la búsqueda inicial). Guardé la base de datos en un documento JSON porque los scripts que descargan los metadatos están en Python, pero el análisis lo hice con el lenguaje de programación R, y necesitaba una manera de transportar la información de un lenguaje a otro.

# 1.2. ¿Cómo conseguí los PDFs?

En total, conseguí 492 PDFs. Si alguien busca en TESIUNAM todos los trabajos de titulación de la carrera de letras inglesas, verá que hay muchos más. Hay dos razones por las que trabajé con 492. La primera es que trabajé con los trabajos de titulación publicados del 2006 al 2023, no con todos los trabajos de titulación existentes de letras inglesas. Descargué los PDFs en el 2024, por lo que los trabajos más recientes de mi investigación son del 2023 (no quiero tener datos de un año que sigue en curso porque estarían incompletos). Que los más recientes sean del 2023 tiene sentido, pero ¿por qué empezar con los del 2006 y no con anteriores? Después de todo, el primer trabajo de la licenciatura fue publicado en 1962. La respuesta es que, antes del 2006, la gran mayoría los trabajos de titulación que alberga TESIUNAM son documentos escaneados, mientras que del 2006 en adelante todos son documentos digitales de origen.

Procesar PDFs producidos directamente desde un procesador de textos es muchísimo más sencillo que procesar PDFs producidos por un escáner, pues los caracteres vienen ya codificados como caracteres, mientras que en los documentos escaneados vienen codificados como pixeles. Para usar PDFs producidos con un escáner se tiene que llevar a cabo uno de varios procesos. El más sencillo y que da resultados de mejor calidad es transcribir manualmente los textos. Este método, claramente, consume mucho tiempo y fuerza de trabajo;por lo tanto, nunca fue una opción que consideré. Otro método es el llamado reconocimiento óptico de caracteres (OCR, por sus siglas en inglés). El OCR utiliza software que reconoce patrones en los pixeles de una foto y los interpreta como caracteres. Este método, a pesar de ser mucho más rápido que el primer método, no da resultados lo suficientemente fidedignos como para confiarles un estudio de esta índole (muchos de los libros en la base de datos de Google Books están codificados con este método, y si alguien ha intentado copiar y pegar texto de esos documentos, sabrá que no da los resultados más confiables). Ninguno de estos dos métodos se acopló a mi metodología, por lo que decidí olvidarme de los PDFs escaneados. Aun así, cuando se realiza la búsqueda en TESIUNAM acotando los resultados a este rango de años, la página nos dice que hay 498 trabajos, mientras que yo digo que conseguí 492.

La segunda razón por la que trabajé con 492 documentos es que hay ciertos trabajos cuyas fichas aparecen en la página de TESIUNAM, pero que no están disponibles. Es posible que lx autorx no haya querido publicar su trabajo en TESIUNAM, como es el caso del trabajo "Metanarrativas y el emtramado [sic] en *Enduring Love* de Ian Mcewan", o que, por alguna razón que TESIUNAM no hace pública, simple y sencillamente no hay botón para descargar el documento, como en el trabajo "The song and the water: el mar, la narración y la vida en *The Waves*, de Virginia Woolf".Conseguir los PDFs fue probablemente la etapa de la investigación que más trabajo me costó. No voy a entrar en detalle pero basta con decir que TESIUNAM está hecha de manera tal que los botones de las páginas de resultados que dicen "texto completo" no llevan directamente al texto completo, sino que hacen una serie de redirecciones que mi código tiene que seguir —como los procesos burocráticos en los que hay que ir a una oficina para que te digan que vayas a otra para que te digan que vayas a otra—. No es nada muy complejo, solamente fue tedioso. Si alguien quiere ver los detalles del código, está en <https://github.com/diego-g-fonte/crawler>.

1.3. ¿Cómo pasé los PDFs a texto procesable?

Pasé los trabajos de titulación de PDF a texto simple (es decir, texto sin formato: sin tipografía especificada, negritas, subrayados, itálicas, tamaños de letra variables, justificación, interlineado, etcétera) con los scripts en Python que están en <https://github.com/diego-g-fonte/pdfs_a_texto/>. Por la naturaleza del formato PDF no pude preservar mucha información que hubiera sido valiosa; por ejemplo, la separación de los párrafos, la capitulación de los trabajos, qué texto es citado y qué texto es propio de lx autorx, etcétera. Esto es porque, en el formato PDF, el texto no está almacenado en un cuadro de texto como en los procesadores de texto, sino que cada caracter es un elemento independiente, con información de su posición, su tamaño, su color, etcétera. Como dice el manual del paquete "pdfminer.six", "Most PDF files look like they contain well-structured text. But the reality is that a PDF file does not contain anything that resembles paragraphs, sentences or even words. When it comes to text, a PDF file is only aware of the characters and their placement" (pdfminer.six). Esto resultó en que un texto que originalmente se veía así: [["legal\_tesiunam.png"]], se vea así:

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso DERECHOS RESERVADOS © PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México). El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

O que uno que se veía así: [["pagina\_tesina\_pdf.png"]], se vea así: [["pagina\_tesina\_vscode.png"]].

Antes de empezar a recolectar los datos, cuando apenas estaba diseñando este proyecto, tenía la esperanza de poder cuantificar todas estas cosas que perdí: la longitud promedio de los párrafos por año, el porcentaje promedio de texto citado *vs* texto propio, las tipografías más populares, lxs autorxs más citadxs, los formatos de citación más populares, la popularidad de cada signo de puntuación, entre tantas otras.[[7]](#footnote-8) La realidad me obligó a olvidar estas ambiciones y a estudiar largos hilos de texto sin puntuación[[8]](#footnote-9) y sin formato.

# 1.4. La base de datos final

Como dije, ordené los metadatos en un documento JSON. Después de transformar los PDFs a texto simple, junté los textos con sus metadatos para al fin tener un documento maestro en donde estuvieran todos los datos que necesitaba para mi investigación. El código que hace esto está en h[ttps://github.com/diego-g-fonte/json\_processing](https://github.com/diego-g-fonte/json_processing). El documento final está disponible en https://github.com/diego-g-fonte/analisis2/blob/main/todo.rds.[[9]](#footnote-10)

2. ¿Cómo elaboré las gráficas?

Ahora voy a describir todo lo que hice antes de llegar al método final. Debo admitir que antes de encontrar la manera de probar mis hipótesis di muchos trompicones metodológicos. Quiero decir que hice varios experimentos basándome en corazonadas y no tanto en si me daban información respecto a mi hipótesis. Por ejemplo, creí que iba a ser buena idea ver las longitudes promedio anuales de los trabajos. Sin embargo, esto no sólo no me dio información que me acercara a comprobar o refutar mis hipótesis, sino que no me dio mucha información útil en general (resulta que los trabajos de titulación se han mantenido casi perfectamente igual de largos[[10]](#footnote-11)). Descubrí el paquete "stylo", que está hecho especialmente para generar información estilométrica en R, y me abrió muchas puertas. Después de mucho tiempo tratando de usar las gráficas que producía me di cuenta de que lo que tenía que hacer era replicar (o inspirarme en) el experimento de Jockers en su capítulo "Influence".

En ese capítulo, Jockers hace una matriz de distancias con todos los libros que está estudiando a partir de datos estilométricos y temáticos. Con esa matriz genera una linda gráfica en la que se puede ver cómo los libros publicados en años cercanos son más similares entre sí que libros publicados con más separación de tiempo, y también cómo los libros escritos por hombres son similares a otros escritos por hombres y los escritos por mujeres son similares a otros escritos por mujeres (figuras ["jockers\_date\_shading.png"] y ["jockers\_gender\_shading.png"]). Mis habilidades técnicas me limitan y no pude hacer una gráfica tan bonita como las suyas, pero leer sobre la matriz de distancias me dio una idea. Seguramente es posible promediar las distancias de los trabajos de titulación en un año. Y seguramente es posible comparar estas distancias promedio para encontrar una tendencia. Si las distancias disminuyen con los años, quiere decir que el estilo se ha ido homogeneizando, y si aumentan, que se ha ido diversificando. Más adelante explicaré qué es una matriz de distancias y el resto de los detalles de mi método.

Jockers dice que usa datos combinando estilo y tema, pero no es explícito en qué quiere decir esto. Por esta razón, y porque mi investigación separa el estilo del tema, decidí no juntar estilo con tema y tratar cada aspecto por separado. Ahora voy a explicar en qué consiste mi experimento sobre el estilo de los trabajos de titulación de la carrera de letras inglesas.

[["jockers\_date\_shading.png"]]

[["jockers\_gender\_shading.png"]]

2.1 Mi experimento sobre el estilo

2.1.1. Qué es la estilometría

La estilometría es un campo de estudio más antiguo de lo que se podría pensar. Más que nada se ha desarrollado con el objetivo de encontrar lxs autorxs "reales" de ciertos textos. Por ejemplo, son populares en las conversaciones respecto a lxs autorxs de textos religiosos, textos firmados con pseudónimos, o cuando se duda de que la persona que firma el texto realmente lo haya escrito. La idea fundamental es que cada persona tiene una "huella digital" estilística, y que ésta se puede revelar usando métodos estadísticos. La forma más rudimentaria (la que yo usé) de encontrar esta huella digital es calculando las frecuencias relativas de las palabras usadas en un texto (o en un corpus de textos). Ya resumí esto brevemente arriba pero lo voy a explicar con más profundidad.

Hay dos tipos de palabras, las funcionales y las de contenido.[[11]](#footnote-12) Las palabras de contenido suelen ser sustantivos, verbos, adjetivos y adverbios, mientras que las funcionales suelen ser preposiciones, pronombres, artículos y conjunciones. Lo que diferencia las palabras funcionales de las de contenido es que las palabras funcionales dan información gramatical mientras que las palabras de contenido dan información semántica. Es fácil pensar que las palabras que definen el estilo son las de contenido: a fin de cuentas, éstas son las que más notamos al leer. Sin embargo, a la estilometría le importan sobre todo las funcionales, pues tiene un entendimiento particular de qué es el estilo. Por ejemplo, nadie me va a disputar que los artículos de mecánica cuántica (no de divulgación) tienen un estilo técnico que es difícil de leer. Esto porque usan palabras (sustantivos, verbos, adjetivos y adverbios) propias del argot de la mecánica cuántica y que no son conocidas por un público amplio. En cambio, los artículos de divulgación científica suelen tener un estilo más sencillo, en gran medida porque usan un léxico menos especializado. Cuando digo que los artículos de mecánica cuántica tienen un estilo técnico y que los de divulgación uno sencillo, me estoy basando en las palabras de contenido que usan, y no es una aseveración errónea, simplemente no es el acercamiento al estilo de la estilometría.

Cuando se trata de problemas de atribución autoral, las palabras de contenido dan muy poca información útil. Voy a poner un ejemplo hipotético. Supongamos que me llega una carta de amor anónima y que yo tengo dos personas sospechosas —llamémoslas persona A y persona B―. Con ambas personas he mantenido una larga correspondencia, por lo que tengo un grupo de cartas que sé con certeza que ha escrito A y otro grupo que sé que ha escrito B. Tanto A como B son personas muy cercanas a mí, por lo que nos escribimos de temas varios: sucesos en nuestras vidas, sucesos públicos, libros, cine, música, etcétera. Si calculara las frecuencias de las palabras de contenido en las cartas de A, podría encontrar que la palabra "película" aparece unas dos veces, la palabra "canción" unas diez, "concierto" tres, "café" cuatro, etcétera; y de las cartas de B que "película" aparece nueve veces, "canción" tres, "concierto" no aparece y "café" aparece tres veces. Más que el estilo de escritura, estas palabras revelan de qué escriben A y B. En otras palabras, me dan información sobre los temas, no sobre el estilo. Algo que podría asegurar tentativamente de este ejemplo es que a A le gusta escribir más de música que a B, mientras que B escribe de cine más que A. Sin embargo, la carta anónima que recibí no habla ni de cine ni de música. Lo que tengo que cuantificar entonces es el estilo, y para cuantificar el estilo lo que hay que contar son las palabras funcionales, no las de contenido. Esto es porque las palabras funcionales son muchísimo más frecuentes que las de contenido y aparecen independientemente del tema del que se esté escribiendo. Para encontrar quién escribió la carta anónima, habría que calcular las frecuencias relativas de las palabras funcionales de las cartas de A, las de las cartas de B y las de la carta anónima y compararlas. Por dar un ejemplo muy burdo, si la frecuencia relativa de la palabra "de" en las cartas de A es de 4%, en las de B de 8% y en la carta anónima de 7%, se puede deducir que B es quien escribió la carta anónima (claro que los métodos y los problemas estadísticos que usa la estilometría real son mucho más complejos que este ejemplo, pero la idea es la misma).

Hay varios ejemplos de métodos estilométricos que se han usado para atribución autoral. Uno es el artículo "Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution", en el que José Nilo G. Binongo se da a la tarea de determinar quién escribió el decimoquinto libro en la serie de Oz. Tiene dos candidatxs, Lyman Frank Baum, quien sabemos que escribió los primeros 14, y Ruth Plumly Thompson, quien sabemos que escribió del 16 al 33. Para hacerlo, cuenta las 50 palabras funcionales más comunes del corpus y crea una gráfica utilizando el método estadístico llamado análisis de componentes principales. En la gráfica se ven dos aglomeraciones diferentes de puntos (cada punto representa un pedazo de una novela): uno en el que están las novelas de Baum y otro en el que están las de Thompson. La gráfica muestra de manera convincente que Thompson escribió la novela número 15, pues todos los pedazos de la novela 15 están en el grupo de Thompson. Quien quiera ver una lista mucho más extensa de trabajos de atribución autoral estilométrica puede consultar "Computational Methods in Authorship Attribution", de Moshe Koppel, Jonathan Schler y Slomo Argamon, en el que hacen un recuento de la historia de los estudios de atribución autoral.

[[ "pca\_oz.png"]]

2.1.2. Graficando el estilo

Al final de mi investigación quedé con una gráfica que muestra la evolución del estilo de los trabajos de titulación del año 2006 al año 2023. Mejor dicho, muestra qué tan similares o diferentes entre sí son los trabajos de cada año en cuanto a estilo. El proceso para obtenerla fue el siguiente. El paquete de R "Stylo" tiene muchas funciones. Una de ellas es que produce una tabla con las frecuencias relativas de las 5,000 palabras más comunes de un corpus de textos; cada fila es una palabra y cada columna es un texto. [["table\_frequencies.png"]] Hice esto con los 492 trabajos que conseguí y resultó la tabla que está en <https://github.com/dieguchofodiego-g-fonte/analisis2/blob/main/comprobacion_hipotesis/estilo/table_with_frequencies.txt>[.](https://github.com/dieguchofo/analisis2/blob/main/comprobacion_hipotesis/estilo/table_with_frequencies.txt) Reduje esas 5,000 palabras a sólo las 150 más comunes, pues me pareció que después de 150, las palabras eran en su mayoría de contenido. Lo que hice después fue dividir la tabla por año de publicación, de manera que ahora tenía 18 tablas, una por cada año. Para cada una de esas tablas, calculé lo que se llama una matriz de distancias, que es una tabla con las distancias entre cada uno de los puntos. Un ejemplo muy sencillo es éste: hay un trabajo de titulación en el que "de" es el 4.42% de todas las palabras y "la" el 3.35%; otro trabajo tiene estas mismas palabras con frecuencias relativas de 5.01% y 4.39% respectivamente. La distancia entre los dos trabajos, tomando en cuenta solamente esos dos aspectos, sería calculada así: (con el famoso teorema de Pitágoras, el cual se puede adaptar a mayores dimensiones). Ése sería solamente un dato dentro de la matriz, la cual tiene dimensiones de N x N, dando por hecho que N es el número de trabajos de titulación publicados en un año. Si la distancia es mucha entre un trabajo y otro, estoy interpretando que sus estilos son muy diferentes, y si la distancia es corta, que tienen un estilos similares.[[12]](#footnote-13)La matriz nos dice qué tan similares o diferentes son los estilos de cualquier par específico de trabajos, pero lo que a mí me interesa es la similitud o diferencia de los textos entre sí a través de los años. Mi solución a esto fue muy sencilla: promediar las distancias de las 18 matrices para así tener 18 distancias promedio. Estas 18 distancias representan qué tan similar fue el estilo de los trabajos publicados en cada año. Están graficadas en la figura ["estilo150.png"]. Ésta gráfica es muy ruidosa y difícil de interpretar. Si usamos una función que calcule una curva de tendencia, obtenemos una gráfica mucho más clara (figura ["estilo150\_smooth.png"]).

[[figura "estilo150.png"]]

[[figura "estilo150\_smooth.png"]]

# 2.2. Mi experimento sobre los temas

Para calcular las diferencias entre temas, seguí el mismo método que para el estilo, pero obtuve los datos de manera diferente. Se podría decir que obtuve la materia prima de una manera diferente pero la procesé igual que la del estilo. El modelado de temas (*topic modeling*) es una técnica que analiza las palabras de contenido dentro de un corpus de textos y encuentra los temas que contienen.[[13]](#footnote-14) A grandes rasgos lo que hacen los programas de modelado de temas es, primero que nada, eliminar todas las palabras funcionales de los textos que van a analizar. Después, calculan las distancias entre las palabras de contenido de un texto, y crean grupos de palabras que repetidas veces aparecen cerca las unas de las otras.

Usé MALLET (MAchine Learning for LanguagE Toolkit), un programa que tiene muchísimas funciones no sólo de modelado de tópicos. Sin embargo, para mi investigación solamente necesité usar una. Con MALLET es muy fácil producir tablas parecidas a las que describí en la sección anterior, pero de temas. Estas tablas contienen los puntajes de cada tema que tiene cada texto. [["tabla\_mallet.png"]] En la tabla no es evidente cuáles son los temas, pero para esta investigación eso en realidad no importa. Al final es lo mismo si los temas son "feminismo", "traducción" y "literatura irlandesa" que si son "tema 1", "tema 2" y "tema 3" porque lo que me interesa no son los temas en sí, sino las similitudes entre trabajos de titulación —tomando como valores, por supuesto, los temas―.[[14]](#footnote-15) Como funciona MALLET es que asume que cada texto tiene cada tema en proporciones diferentes. Un ejemplo hipotético podría ser con textos científicos. Supongamos que dentro de un grupo de textos científicos MALLET descubre dos temas: plantas y rocas. Un texto de biología describiendo una nueva especie de arbusto podría tener puntajes de 0.9 para plantas y 0.1 para rocas. MALLET da puntos del 0 al 1, y siempre todos los puntajes de un texto van a sumar 1. Esto es porque es un modelo probabilístico, y en el campo de las probabilidades se piensa de algo con probabilidad de 0 como algo imposible y de algo con probabilidad de 1 como algo certero. Una moneda tiene una probabilidad de 0.5 de caer águila y de 0.5 de caer sol. Lo que querría decir el puntaje hipotético de MALLET es que cualquier segmento del texto de biología tiene una probabilidad de 0.9 de ser de plantas y de 0.1 de ser de rocas (en mi ejemplo imaginario, el texto de biología puede hablar de los minerales de la tierra en la que vive el arbusto y cosas parecidas).

Una peculiaridad de los modelos de temas es que requieren que alguien les especifique el número de temas que deben encontrar. Es decir, pueden generar dos temas o mil. Algo curioso que encontré es que si le pedía dos, tres o cuatro temas, encontraba que se dividían en trabajos de traducción, de didáctica y de investigación (tesis y tesinas): dos temas resultaban en investigación y el resto; tres en investigación, traducción y el resto; y cuatro en investigación de narrativa, investigación de poesía, traducción y el resto (el resto siempre era sobre todo de didáctica).[[15]](#footnote-16) Escribe Jockers sobre la elección de la cantidad de temas al usar programas de modelado de temas, “[t]here is neither consensus nor conventional wisdom regarding a perfect number of topics to extract, but it can be said that the ‘sweet spot’ is largely dependent upon and determined by the scope of the corpus, the diversity of the corpus, and the level of granularity one is seeking in terms of topic interpretability and coherence” (128). Como dije, la “interpretabilidad” de los temas no era mi prioridad. Noté que cuando pedía 50 temas, más o menos la mitad me eran inteligibles y la mitad no. Esto junto con que me pareció adecuado tener un tema por cada 10 textos fue lo que me llevó a elegir 50.[[16]](#footnote-17) Con los 50 temas que identificó MALLET, me quedé con una tabla de 50 filas y 492 columnas. Esta tabla la procesé igual que la de frecuencias relativas y resultó la gráfica de la figura ["tema\_y\_real.png"].

[["tema\_y\_real.png"]]

Bibliografía

Binongo, José Nilo G. "Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution". *CHANCE*, vol. 16, no. 2, Mar. 2003, pp. 9–17.

Blei, David M. "Probabilistic Topic Models". *Communications of the ACM*, vol. 55, no. 4, Apr. 2012, pp. 77–84.

"Converting a PDF file to text". *Pdfminersix*,

pdfminersix.readthedocs.io/en/latest/topic/converting\_pdf\_to\_text.html, accesado el 12 septiembre 2024.

Hallebeek, Jos. "Las palabras funcionales del español". *Boletín AEPE,* no. 34-35, pp. 205-216.

Jockers, Matthew. *Macroanalysis*. University of Illinois Press, 2013.

Koppel, Moshe, et al. "Computational Methods in Authorship Attribution". *Journal of the American Society for Information Science and Technology*, vol. 60, no. 1, Jan. 2009, pp. 9–26.

1. Lo llamé "crawler" ingenuamente; es un scraper. La diferencia no importa para este trabajo, pero quiero que, si por alguna razón alguien que sepa del tema lee esta tesina, sepa que estoy consciente de mi error. [↑](#footnote-ref-2)
2. Un n-grama es nada más una agrupación de palabras consecutivas. Si tomamos la primera frase de este capítulo, "antes que nada", separado por unigramas tendríamos "antes", "que", y "nada". Separado por bigramas, "antes que", y "que nada", y por trigramas sólo tendríamos "antes que nada". También se pueden entender los n-gramas como agrupaciones de caracteres, de manera que, por ejemplo, la palabra "n-grama" separada en tetragramas de caracteres daría "n-gr", "-gra", "gram" y "rama". [↑](#footnote-ref-3)
3. R es un lenguaje de programación diseñado para análisis estadísticos. [↑](#footnote-ref-4)
4. MALLET es un programa para hacer modelados de temas. Más adelante lo describo más a detalle. [↑](#footnote-ref-5)
5. No voy a entrar en detalles respecto a JavaScript ni sus objetos porque no escribí nada en JavaScript para este proyecto, y no soy experto en ese lenguaje. [↑](#footnote-ref-6)
6. A lo largo de este trabajo, voy a diferenciar el texto escrito por máquinas usando una tipografía monoespaciada como ésta. [↑](#footnote-ref-7)
7. Estos datos serían valiosos para ver tendencias estilísticas que no dependen de la elección de palabras. Entre más datos, más acertados son los resultados; se podrían haber sumado las longitudes de las oraciones, de los párrafos y las frecuencias de signos de puntuación a los datos que procesé para así tener un conjunto de datos más rico.

   Además de esto, podrían haber sido valiosos para ver tendencias diferentes a las que yo medí. Sólo por poner un ejemplo, si pudiéramos medir la cantidad de texto citado en los trabajos de titulación a través de los años, podrían emerger líneas de investigación interesantes. Si aumentara la cantidad de texto citado podríamos conjeturar que es por un aumento en el rigor de los trabajos. Si disminuyera podría ser por un aumento al valor que se le da a la voz de lx autorx. Otro ejemplo es que si no me hubiera deshecho de los signos de puntuación, podría haber tenido la longitud de las oraciones como dato adicional, además de la frecuencia de cada signo. Entre más datos hay mayores son las posibilidades de investigación y de interpretación. [↑](#footnote-ref-8)
8. No es que los signos de puntuación sean imposibles de preservar cuando se pasa un PDF a texto simple, es que los trabajos de titulación suelen usar signos de puntuación para comunicar información que no es la puntuación de un texto. Por ejemplo, usar puntos en el índice entre el nombre del capítulo o sección y su número de página, o usar muchos guiones o guiones bajos para dividir secciones (o notas al pie de página del texto normal). Alguien con más tiempo que yo podrá investigar los signos de puntuación en los trabajos de titulación de letras inglesas del 2006 al 2023. [↑](#footnote-ref-9)
9. Este documento ya no es un documento JSON, sino un documento RDS (R Data Serialization). Esto es un tipo de documento parecido a JSON en cuanto que es una manera de transportar objetos de R (así como JSON es una manera de transportar objetos de JavaScript). Las diferencias son que estos documentos no son fáciles de leer por personas (se tiene que usar R para leerlos) ni son fáciles de leer con otros lenguajes de programación. Es decir, sirven para transportar objetos de R de un script a otro, que es justo lo que yo necesitaba. [↑](#footnote-ref-10)
10. Me imagino que esto se debe a que, en el reglamento de las opciones de titulación, las extensiones mínima y máxima no se han modificado. Sí hay una muy ligera tendencia a que sean menos largos, pero es muy pequeña y no me siento con la seguridad de llegar a ninguna conclusión. [↑](#footnote-ref-11)
11. Jos Hallebeek, en "La palabras funcionales del español", las llama palabras funcionales y palabras lexicales. A las primeras también se les llaman palabras gramaticales, palabras vacías o *stopwords* en inglés. A las segundas también se les dice palabras de contenido o palabras plenas. Yo, como alguien que aprendió esta distinción leyendo textos en inglés, prefiero llamar a las primeras palabras funcionales (*function words*) y a las segundas palabras de contenido (*content words*). [↑](#footnote-ref-12)
12. Para calcular las distancias usé la función de R "dist". El código que genera las gráficas de estilo está en https://github.com/diego-g-fonte/analisis2/blob/main/comprobacion\_hipotesis/estilo/2.r. La manera de entender las gráficas es que si el punto es alto, los trabajos de ese año son diferentes, y si el punto es bajo, son similares. [↑](#footnote-ref-13)
13. Los programas de modelado de temas emplean fórmulas matemáticas que exceden mi conocimiento actual. David M. Blei tiene un artículo bastante didáctico ("Probabilistic Topic Models" en el número 4 de la revista Communications of the ACM) que explica cómo funcionan sin necesariamente entender los modelos matemáticos. [↑](#footnote-ref-14)
14. Una posible línea de investigación podría ser sobre la evolución de los temas a través del tiempo. Por ejemplo, según lo que logré divisar, el feminismo se ha hecho un tema más presente en los trabajos recientes. Lamentablemente, esta línea excede los límites de esta tesina. [↑](#footnote-ref-15)
15. Otra línea de investigación que decidí no perseguir sería sobre la posibilidad de categorizar los trabajos de titulación justamente siguiendo este esquema. Se podrían usar los modelos de temas para automáticamente categorizar a los trabajos como tesis/tesina, traducción comentada, propuesta de material didáctico, etcétera. [↑](#footnote-ref-16)
16. Me parece importante notar que el número 50 es resultado de una decisión que bien podría haber sido diferente. Por la naturaleza de mi investigación, la cantidad específica de temas es de muy poca importancia; más importante es lo que se hace después con esos datos. [↑](#footnote-ref-17)