“Lectura distante y minería de corpus textuales con Python”

Romina De León
CONICET
rdeleon@conicet.gov.ar

Objetivo: Realizar una aproximación a las técnicas de minería de texto mediante Python en corpus de textos literarios o científicos.

Recorrido didáctico

- **Qué implica en la actualidad la ingesta masiva de datos y formas de análisis de estos:** Análisis textual estadístico y estilístico.
- **Acercamiento a las técnicas de exploración y análisis de corpus:** Data mining y text mining. Ejemplos de uso.
- **Empleo de Python para el análisis cualitativo y cuantitativo de grandes datos con métodos digitales:** Ejemplos de uso y presentación de scripts de análisis textual, para acceder a una construcción de narrativas y toma de decisiones: del dato al texto y viceversa.

Actividad (sincrónica)

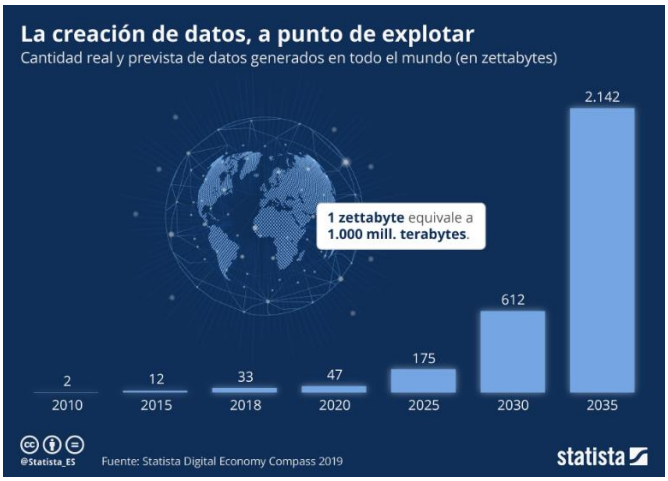
- Presentación sobre grandes datos, uso de los datos
- Lectura distante
- Minería de datos y de textos
- Análisis con métodos digitales.

Actividad (asincrónica)

- Introducción a Python para el análisis de textos de manera cualitativa y cuantitativa y tratamiento de análisis de textos para procesamiento automatizado.
- Ejercicio práctico de minería y procesamiento de lenguaje natural en textos, modelado de tópicos de interés.
- Ejercicio práctico de visualización y representación de la información obtenida, una vez procesados los textos.

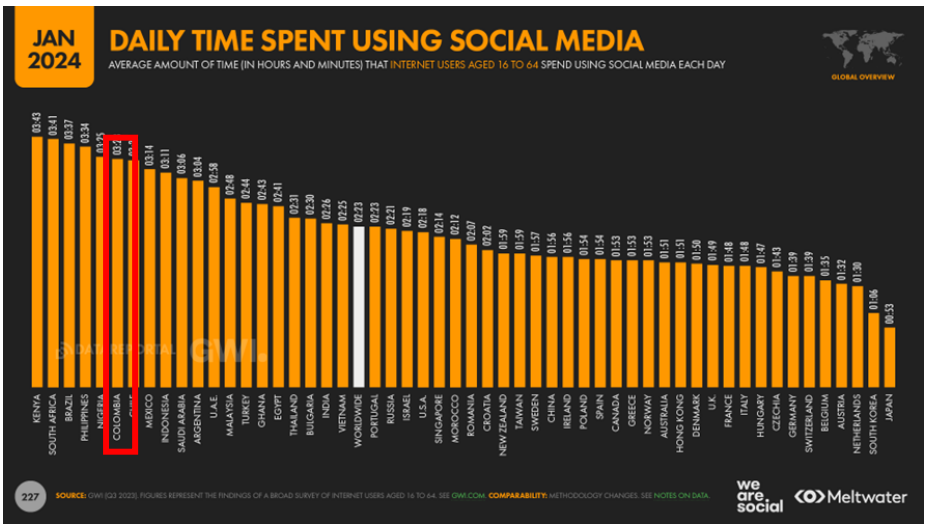
Los datos textuales, la lectura distante y cómo analizarlos

En los últimos 30 años, el surgimiento de las redes sociales y los avances tecnológicos han dado lugar a una generación masiva de información. Cada día se producen cantidades inimaginables de datos, con miles de terabytes (zettabytes) de información circulando a través de diferentes plataformas. Este fenómeno, en su esencia, está destinado a enriquecer nuestro conocimiento y comprensión del mundo a través de métodos avanzados de análisis. Pero ¿los datos son utilizados solo para conocimiento?

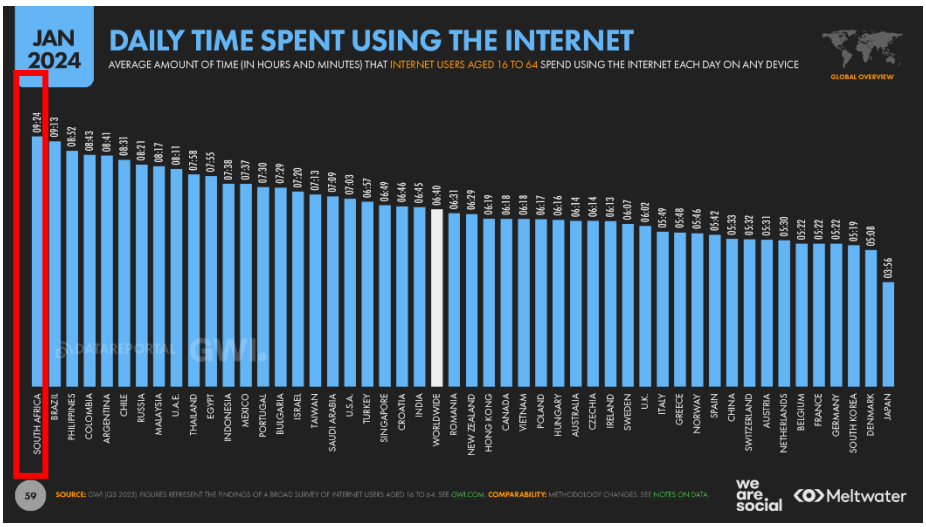


Fuente: https://es.statista.com/grafico/17734/cantidad-real-y-prevista-de-datos-generados-en-todo-el-mundo/

Este diluvio de datos hacer que vivamos inmersos en una explosión de información sin precedentes en la historia de la humanidad. Esto ha dado lugar a un nuevo paradigma conocido como ***datificación***. Esta refiere a la creciente interconexión de la sociedad moderna, donde las redes sociales no solo conectan a personas, sino también que vinculan a dispositivos móviles, máquinas y sistemas corporativos. Esta interconexión facilita la interoperabilidad y el intercambio de información a niveles antes inimaginables, lo que ha provocado la digitalización de prácticamente toda actividad humana.



Tiempo diario en redes sociales por países, usuarios entre 16 y 64 años. Datos a enero de 2024. Fuente: https://datareportal.com/reports/digital-2024-global-overview-report

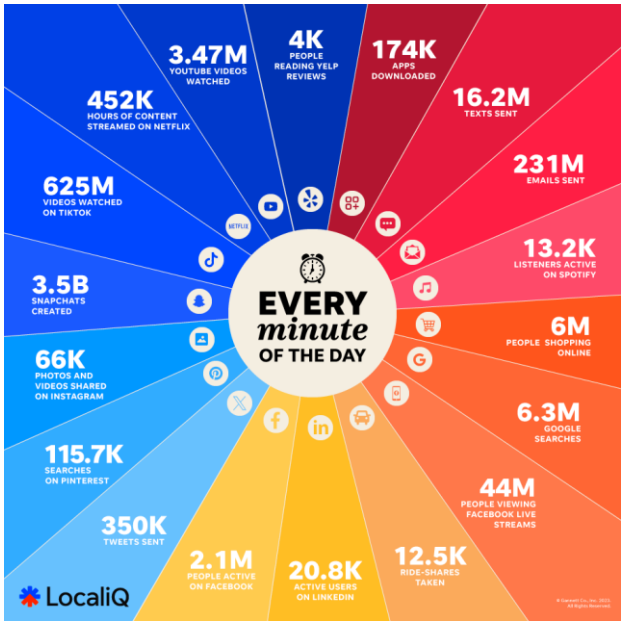


Tiempo diario en que se utiliza internet. Datos a enero de 2024. Fuente: https://datareportal.com/reports/digital-2024-global-overview-report.

Un nuevo horizonte de análisis ha surgido con el término *Big Data*, que refiere a la capacidad de recopilar y analizar, a gran velocidad, volúmenes enormes de datos que son variados en forma y estructura. El objetivo es establecer relaciones significativas que aporten conocimiento a partir de información que, por su complejidad, no podría ser interpretada de manera convencional.

Los métodos que definen las técnicas de Big Data están fundamentalmente vinculados a áreas como la estadística computacional, el aprendizaje automático y la minería de datos. Estos campos permiten transformar el diluvio de datos en conocimiento útil, facilitando la toma de decisiones en diversas disciplinas.

En esta era emergente, la de la datificación, la transformación del mundo está marcada por un crecimiento continuo y exponencial de los datos, generados a nivel global segundo a segundo. Sin embargo, estos datos son crudos y están cargados de información, lo que plantea el desafío de cómo procesarlos y analizarlos efectivamente.



Lo que sucede en internet en 60 segundos. Fuente: https://localiq.com/blog/what-happens-in-an-internet-minute/.

Como podemos observar, la información generada es mayormente textual. De hecho, se estima que aproximadamente el 80% de los datos que se generan de manera constante son de naturaleza textual. Este tipo de datos, conocidos como no estructurados, son más complejos de analizar en comparación con los datos numéricos.

Los datos textuales tienen una importancia relativa crucial, ya que las palabras que los componen forman una parte fundamental de cualquier tipo de análisis. El significado que otorguemos a los resultados derivados de estos análisis influye directamente en el valor y la relevancia del estudio o proyecto en cuestión.

Para abordar el análisis de estos datos no estructurados, es imprescindible recurrir a herramientas avanzadas como el Big Data, el Machine Learning o la Inteligencia Artificial. Estas tecnologías permiten que los grandes volúmenes de datos sean filtrados y analizados de forma efectiva, facilitando la diseminación de conocimientos más allá de su ámbito original de estudio. Incluso, podemos observar que ofrecen soluciones innovadoras a dilemas que, anteriormente, parecían insalvables.

Ahora bien, ¿cómo se podría analizar esa enorme cantidad de datos? Para abordar esta cuestión, reflexionemos cómo incorporamos los datos textuales; la respuesta es sencilla: a través de la lectura. Sin embargo, dentro de la lectura diferenciamos la lectura cercana, es decir, la que practicamos cotidianamente en nuestra vida, lo que están haciendo con este texto, o al momento de leer un libro. Luego, tenemos una lectura actual, digital, enriquecida con elementos como hipertextos, la que hacer al visitar una página web. O en este texto, con los hipervínculos que les comparto.

Volviendo a la datificación, podemos considerar que ha generado nuevas formas de lectura y de interpretación. Varios humanistas han analizado estas metodologías, entre ellos el investigador italiano Franco Moretti, quien ha publicado diversos trabajos con formulaciones novedosas que le han otorgado visibilidad en el campo de los estudios histórico-literarios. Moretti propuso trascender las metodologías de lectura cercana (Close Reading), como el análisis de contenido al que todos estamos acostumbrados, y llevar los estudios literarios e históricos hacia la lectura de grandes volúmenes de texto de manera no literal, sino cuantitativa. Esta aproximación permite superar la interpretación tradicional, resaltando estructuras como repeticiones, regularidades y patrones, que se hacen visibles al percibir la literatura desde una perspectiva de *longue dureé* (Moretti, 2000, 2013; Braudel, 1980)[1].

Por otro lado, es importante comprender que, para una historiadora como es mi caso, y en general para cualquier humanista, tratar con datos masivos no tiene la misma relevancia ni orientación que cuando son explotados por un analista de mercado, un economista o un ingeniero. Si seguimos la explicación de Shawn Graham, arqueólogo digital, los datos masivos pueden entenderse como una cantidad de información mayor a la que un investigador podría abarcar en su trabajo cotidiano, en ciertos casos en toda su vida. Es decir, que sería humanamente imposible leer e interpretar de manera tradicional ese tipo de cantidad (Graham, 2022).

Con todo lo anterior, se puede comprender que quienes investigamos y trabajamos en Humanidades o Ciencias Sociales, la información suele encontrarse de forma semiestructurada. Esta se encuentra catalogada

[1] Para una interesante entrevista a Franco Moretti que recorre sus trabajos y formación, véase Hackler y Kirsten, 2016.

de forma tal que es posible identificar el tipo de documento, las características físicas del original, el lugar de procedencia, autores, e incluso un resumen o descriptor del contenido. Todo esto conforma lo que llamamos *metadatos*, información que permite ubicar, ordenar y guardar datos de manera automatizada. Sin embargo, también existe mucha información no estructurada, como ya he mencionado, que consiste en una inmensa colección de unidades de lenguaje cuyo análisis no puede depender completamente de la automatización. Por esta razón, en ocasiones se prefiere construir bases de datos antes que "minar" recursos web para recolectar información. Las humanidades digitales, campo que aúna Humanidades con tecnología, como habrán visto en la clase de la Dra. del Rio, se han enfrentado a lo que podría considerarse una nueva necesidad de lectura, un nuevo paradigma de interpretación de textos literarios, históricos, etc. mediante la mencionada *lectura distante*.

Por ello, la tesis de Moretti (2013) refiere a que las grandes escalas a las que nos enfrenta el medio digital hacen que necesitemos cuantificar la literatura y leerla (o hacer que las máquinas la lean por nosotros) en base a nuevas disposiciones provenientes de otras disciplinas científicas. Por lo que argumenta que la única manera de analizar estos inmensos corpus es mediante la *lectura distante*, es decir, no prestando atención a los detalles, sino delimitando ciertas características, ciertos patrones, que se comparan a través del procesamiento computacional.

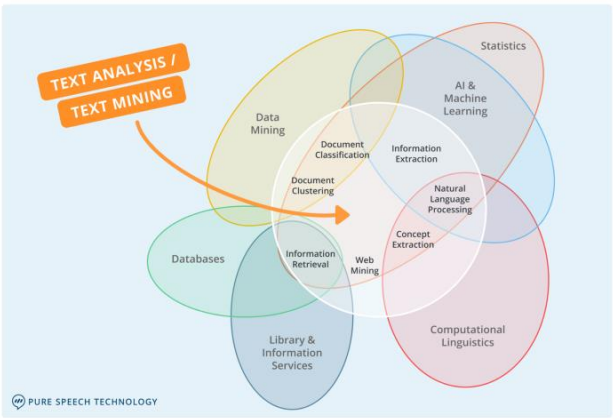
De manera similar, otro humanista digital, Matthew Jockers (2013), ha preferido hablar de macroanálisis para describir métodos cuantitativos y computacionales similares, considerando este término más intuitivo. En síntesis, la consigna es: los humanos ya sabemos leer textos, ahora aprendamos a no leerlos. Es decir, dejemos que las máquinas procesen por nosotros esas grandes cantidades de datos, de big data, de big corpus que solo ellas pueden manejar. Entonces, debemos realizar una *macrolectura* o lectura distante de los grandes corpus de datos, y pensar en cómo analizarlos.

En primer lugar, si estos datos fuesen en formato estructurado o numérico, podríamos utilizar bases de datos, realizar análisis numéricos o estadísticos, comprendiendo la estadística como ciencia, método y técnica, pues nos permite contabilizar elementos, calcular su probabilidad, proponer descripciones de la distribución, etc. Algo similar puede hacerse con los datos no estructurados, o textos, a través de la textometría o estadística textual, donde los elementos a contabilizar pueden ser caracteres, palabras, clases de palabras, etc., y sus descripciones pueden abarcar una gran variedad de temas, como el crecimiento del vocabulario, el establecimiento de series textuales cronológicas, el cálculo de segmentos repetidos (para detectar énfasis del autor o autora), la identificación de fijaciones léxicas, o su contrapartida, las alteraciones discursivas.

En este ámbito se incluye la minería de texto, tema que nos convoca, que se considera más cercana a la *ciencia de datos* que a la estadística, algo que exploraremos en detalle. Esta metodología se adapta a donde haya datos no estructurados, es decir, en forma de texto. Para profundizar en el análisis de texto, partimos de la ciencia de datos, que se define como el estudio de datos con el objetivo de extraer información significativa para determinados estudios científicos, intereses de empresas, etc. Cuenta con un enfoque multidisciplinario, combinando principios y prácticas de matemáticas, estadística, inteligencia artificial e ingeniería informática para realizar análisis de grandes cantidades de datos.

Ciencia de datos. Fuente: https://commons.wikimedia.org/wiki/File:Data-Science-Landscape.jpg

No nos detendremos mucho más allí, sin embargo, si profundizaremos en la minería de datos, que como podrán entender, busca anomalías, patrones y correlaciones en grandes conjuntos de datos para predecir resultados. Con algunas diferencias y similitudes respecto al trabajo con bases de datos, estructuras, tablas, etc., la minería de textos utiliza datos no estructurados, trasladando los textos a una base de datos semi estructurada. Así, el *text mining* se relaciona con áreas como *data mining*, *machine learning* y lingüística computacional.



Fuente: https://towardsdatascience.com/a-guide-text-analysis-text-analytics-text-mining-f62df7b78747

Ahora bien, voy a sumar algunas consideraciones, que creo importantes, para llevar adelante un análisis textual, y que serán útiles para comprender por qué realizamos determinados cálculos al desarrollar *scripts* (una secuencia de instrucciones, en nuestro caso para un lenguaje de programación).

En términos generales, entendemos que un texto está compuesto por palabras, cuya cantidad dependerá del tipo de texto: puede ser un párrafo, un tuit, un libro, un comentario en redes sociales, todos los escritos de un autor o los libros de una biblioteca. No obstante, mediante esta metodología podremos analizar palabras, estructuras semánticas y sintácticas, lo que nos permitirá extraer información útil de cualquier tipo de texto. Es decir, podremos observar cómo y qué dice un texto, lo que influirá en otros resultados, por ejemplo, en el análisis de una encuesta o en el desarrollo de un *chatbot*.

Pero ¿qué debemos tener en cuenta para comprender cómo se realizan estas técnicas? Sabemos que el lenguaje es nuestra facultad para expresarnos y comunicarnos, ya sea mediante el habla u otros signos. Este comprende estilos, modos y normas en todas sus formas, lo que lo convierte en uno de los sistemas más complejos que existen, dada la multitud de lenguas y estructuras que lo conforman. Se suma, además, que el lenguaje no solo depende del contexto, sino que se encuentra en continua evolución, por lo que podría intuirse que su modelado o la realización de predicciones específicas podría ser una tarea casi imposible.

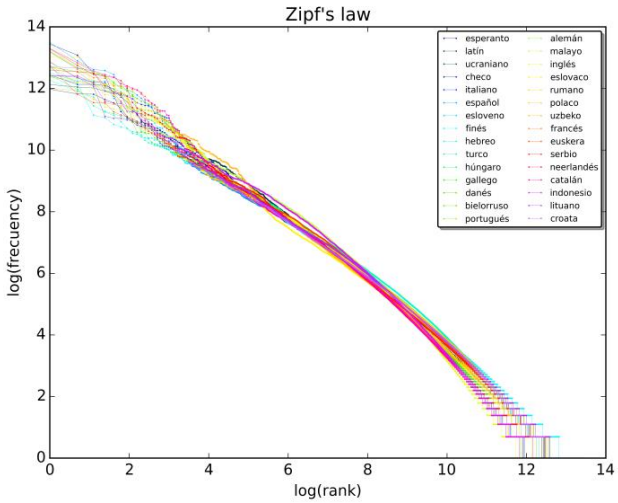
Sin embargo, se sabe que aproximadamente el 50% del contenido de cualquier libro, artículo o conversación está compuesto por las mismas 50-100 palabras, mientras que aproximadamente la otra mitad serán palabras que aparezcan solamente una o dos veces. Estos argumentos fueron estudiados de forma empírica y matemática por George Zipf, lingüista de la Universidad de Harvard, durante la década de 1940. Zipf observó que la mayoría de las palabras se repetían siempre en cualquier idioma, en un libro o en un artículo. Lo interesante es que identificó un patrón en la frecuencia de uso de cada palabra. El patrón que descubrió es que la frecuencia de aparición de una palabra es proporcional al inverso de la posición que ocupa dicha palabra según su número de apariciones. Esto se conoce como la Ley de Zipf, y es observable en todos los idiomas. Por ejemplo, en español, las diez palabras más frecuentes según la RAE son las que aparecen en mayor tamaño en esta nube de palabras. Esto significa que la palabra «de» será la más frecuente en casi todos los textos.

Forma	Frec. gral.	Frec. norm.
de	20670985	61827.534
la	12120610	36253.107
que	10460665	31288.162
y	8847001	26461.645
el	8405150	25140.055
en	8368164	25029.429
a	6766329	20238.292
los	5036895	15065.504
se	4032356	12060.897
del	3663174	10956.663

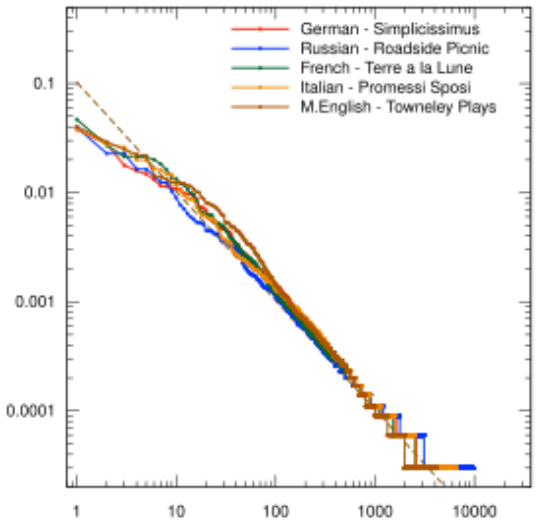
Listado de palabras más frecuentes en el CORPES (Corpus del Español del Siglo XXI). Fuente: https://www.rae.es/noticia/conozca-algo-mas-el-corpes-listados-de-frecuencias.

Es importante mencionar que la curva de Zipf se representa generalmente utilizando el logaritmo de las frecuencias y el ránking de los elementos, lo que permite apreciar mejor la parte central de los datos (las palabras de frecuencia media). Los elementos centrales de la curva son los que mejor representan un texto y permiten caracterizarlo y establecer comparaciones con otros textos, ya que típicamente las palabras en la cima de la curva corresponden a *stopwords* o palabras vacías (es decir, demasiado generales y que aparecen en casi cualquier texto), mientras que las palabras hacia el final del eje X tienen una única ocurrencia (son

particulares y ocurren únicamente en ese texto). Este comportamiento es clave para el análisis textual, ya que permite diferenciar entre palabras que forman parte del núcleo común del lenguaje y aquellas que son específicas de un texto o un autor en particular, proporcionando así información valiosa sobre el estilo, el contenido, y las características distintivas del texto analizado.



Gráficas de ranking versus la frecuencia de las primeras 10 millones de palabras en 30 Wikipedia (descargas de octubre del 2015) en una representación logarítmica en los dos ejes. Fuente: https://upload.wikimedia.org~.png.



Textos en alemán (1669), ruso (1972), francés (1865), italiano (1840), e inglés medieval (1460)[2]. Fuente: https://en.wikipedia.org/wiki/Zipf~.svg.

Parte práctica:

Ahora, vamos a realizar una revisión de algunas cuestiones sobre las aplicaciones, la utilidad, y los resultados que se pueden obtener con la minería de textos. Esta técnica se utiliza en diversas áreas, como el marketing, la Sociología, las Humanidades, entre otras, mediante herramientas y metodologías específicas.

En resumen, las cinco principales aplicaciones del *Text Mining* que podemos destacar son:

1. **Búsqueda de Información:** Se refiere a las búsquedas realizadas a partir de una pregunta, palabra clave o contenido específico.
2. **Reconocimiento de Entidades Mencionadas y Referencias:** Este análisis agrupa textos que contienen la misma palabra o información, utilizando estadísticas sobre nombres de lugares, personas, entidades, etc.
3. **Clustering:** Agrupa textos que cumplen con criterios similares, aunque estos no sean evidentes a simple vista.
4. **Clasificación:** Consiste en etiquetar textos para dividirlos en categorías de manera más eficaz.
5. **Análisis de Sentimientos:** A través de resultados estadísticos, este análisis permite determinar si un texto es positivo, negativo o neutral.

Cada una de estas aplicaciones estará relacionada a una enorme cantidad de áreas donde son utilizadas, como ser gestión de riesgo, prevención de delitos cibernéticos, enriquecimiento de contenido, etc.

[2] Los textos son: Alemán, *The Adventures of Simplicius Simplicissimus* (*Der Abenteuerliche Simplicissimus Teutsch*), una novela de Hans von Grimmelshausen (~1669). Ruso, novela *Roadside Picnic* (*Piknik na obochine*) de Arkady and Boris Strugatsky. Francés, de Jules Verne, *De la Terre à la Lune* (1865). Italiaano, de Alessandro Manzoni, la novela *I Promessi Sposi* (*The Bethrothed*) (1840). Inglés medieval de *Wakefield Mystery Plays* aka *Towneley Mystery Plays* (1460).

Y, por último, les comparto el recorrido metodológico que seguiremos en nuestro análisis:

1. **Recolección:** Recopilación de datos
2. **Preprocesamiento:** Identificación del contenido y extracción de características representativas.
    - **Limpieza de textos:** Eliminación de información innecesaria o no deseada
    - **Tokenización:** los programas solo "ven" una cadena de caracteres sin poder identificar párrafos, frases o palabras. La tokenización divide el texto en entidades significativas (palabras, oraciones, etc.) según los espacios en blanco y las puntuaciones presentes.
    - **Extracción de características (o selección de atributos):** Proceso de caracterización de los textos.
3. **Indexación:** Creación de un índice de ciertos términos, sus ubicaciones y frecuencias, lo que permite un acceso rápido y estructurado a los datos procesados.
4. **Minería de Texto:** Aplicación de diferentes técnicas de exploración de datos para extraer conocimiento a partir del texto preprocesado.
5. **Análisis:** Este último punto es fundamental, y siempre será realizado por el científico, analista o el interesado que esté llevando adelante el proceso, pues los resultados de la minería de texto deben evaluarse y visualizarse para poder interpretarlos en función de las preguntas que se desean responder.

Para concluir esta parte, les compartiré un sitio de donde descargarán el material[3], con las instrucciones a seguir, que nos permitirán entender tanto el método estadístico como el computacional, y cómo interpretarlos. Es crucial comprender qué hay detrás de estas técnicas, así como el contexto en el que estamos trabajando. No realizaremos un desarrollo exhaustivo de scripts ni análisis estadísticos, sino que nos centraremos en comprender cómo utilizar la metodología general para llevar a cabo un análisis estilístico textual. El objetivo es aprender a manejar las herramientas y evaluar si el modelo que empleamos se ajusta a nuestros requerimientos.

Referencias bibliográficas

Amat Rodrigo, J. (2020). *Análisis de texto (text mining) con Python*. https://cienciadedatos.net/documentos/py25-text-mining-python

Braudel, F. (1980). "History and the Social Sciences: The Longue Durée". En *On History (Trad. Sarah Matthews)*. University of Chicago Press.

Caballero Roldán, R., Martín Martín, E., & Riesco Rodríguez, A. (2019). *Big Data con Python*. En Pubhtml5. https://pubhtml5.com/nkpo/lcwg/

Graham, S. (2022). The joys of big data for historians. In S. Graham, I. Milligan, S. B. Weingart, & K. Martin (Eds.), Exploring Big Historical Data (2nd ed.). World Scientific Publishing Co. Pte. Ltd. https://doi.org/10.1142/9789811243042_0001

[3] Material de trabajo: https://github.com/rominicky/mineria-texto-python.

Hackler, R. M., & Kirsten, G. (2016b). Distant Reading, Computational Criticism, and Social Critique: An Interview with Franco Moretti. *Le Foucaldien*, *2*(*1*), 7. https://doi.org/10.16995/lefou.22

Jockers, M. L. (2013). MACROANALYSIS. In *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

Kokensparger, B. (2018). Chapter 7 "Textual Analysis: Frequencies and Stop Words in Dirty Text". En *Guide to Programming for the Digital Humanities: Lessons for Introductory Python*. Springer. https://doi.org/10.1007/978-3-319-99115-3_7

Laramée, F. D. (2018). Introduction to stylometry with Python. *The Programming Historian*, *7*. https://doi.org/10.46430/phen0078

Moretti, F. (2000). Conjeturas sobre la literatura mundial. *New Left Review*, *3*. https://newleftreview.es/issues/3

Moretti, F. (2013). *Distant reading*. Verso Books.

Stover, J. A., Winter, Y., Koppel, M., & Kestemont, M. (2015). Computational authorship verification method attributes a new work to a major 2nd century African author. *Journal of the Association for Information Science and Technology*, *67*(1), 239-242. https://doi.org/10.1002/asi.23460