<http://rafalab.dfci.harvard.edu/dslibro/introducci%C3%B3n-a-machine-learning.html>

<https://rpubs.com/jboscomendoza/mineria-de-textos-con-r>

<https://arcruz0.github.io/libroadp/index.html> AnalizaR Datos Políticos. Francisco Urdinez y Andrés Cruz Labrín (Editores). Versión 1.1, compilada el 2021-04-17

<https://bookdown.org/gaston_becerra/curso-intro-r/>

Estos dos son para modificar el script

<https://towardsdatascience.com/rip-wordclouds-long-live-chatterplots-e76a76896098>

<https://cran.r-project.org/web/packages/ggwordcloud/vignettes/ggwordcloud.html> advanced feature

modificar para datificación: <https://www.statista.com/statistics/871513/worldwide-data-created/>

Se propone que les

asistentes 1) logren comprender algunos conceptos metodológicos

fundamentales para el preprocesamiento de datos textuales

(tokenizacion, lematizacion, stemming, etc.) y representación vectorial clásica de textos (Term-Frequency Matrix, tf-idf, bag of

words, n-gramas, etc.); 2) conozcan algunas técnicas de modelado y

detección de tópicos; 3) se introduzcan a algunas técnicas

modernas de representación vectorial de textos (word

embeddings); 4) sean capaces de identificar situaciones de

aplicación de estas técnicas en sus propias investigaciones. Se

utilizará como herramienta el lenguaje R para el procesamiento y

modelado de los datos. Vale aclarar que no se trata de un curso de

capacitación en R, sino que el lenguaje será utilizado como

herramienta para generar flujos de trabajo replicables.

**Objetivo del encuentro:** Realizar una aproximación a las técnicas de minería de texto mediante el lenguaje de programación R 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 RStudio en el análisis cualitativos y cuantitativos de grandes datos con métodos digitales. Ejemplos de uso y presentación de script de análisis textual, para acceder a una construcción de narrativas y toma de decisiones: del dato al texto y viceversa.

**Actividad (en clase)**

* Introducción a R 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.

**Referencias bibliográficas**

Grolemund, G., & Wickham, H. (s. f.). *R para ciencia de datos*. Recuperado 7 de julio de 2023, de <http://es.r4ds.hadley.nz/>

**Alternativas de lectura**

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59* (10), 1-23. <https://doi.org/10.18637/jss.v059.i10>

**Actividad 4.2**

* Trabajo en grupo para analizar ejemplos de data y text mining
* Presentación de script para text mining
* Trabajo en R con script

Maestría y Especialización en métodos cuantitativos para la gestión y análisis de datos en organizaciones

Buenas tardes, mi nombre es Romina De León, la clase de hoy “**Lectura distante y minería de corpus textuales con R**”

Para ello les comento como pensé el orden de la clase:  
\* parte teórica de datos y de big data

\* Lectura distante

\* que significa data mining o minería de datos y el text mining o minería de textos y estilometría, como para introducirlos un poco en el tema

\* ejemplos de minería de texto mediante el uso de R.

## FILMINA 2

Para preparar esta clase comencé pensando en el surgimiento de las redes sociales y de la tecnología, la información que se genera día a día es una cosa absolutamente inaudita, hay miles de teras de información (zettabytes), repensando, en forma general, que todo ello tiene como propósito que nos enriquezcamos, à métodos para analizarlas

Además como Vivimos inmersos en un diluvio de datos, una explosión de información sin precedentes en la historia de la humanidad que ha dado lugar a un **nuevo paradigma** --> **datificacion**

à es decir que la interconexión de la sociedad actual, donde las redes sociales ponen en contacto a las personas, es decir que las personas están siempre conectadas con dispositivos móviles, además las máquinas y los sistemas corporativos también se conectan entre sí para interoperar o intercambiar información, à ha provocado la digitalización de toda la actividad humana.

à Esta revolución digital genera montañas de datos en forma de registros web, videos, grabaciones de voz, fotografías, correos electrónicos, posts, etc., como dijimos a la espera de poder ser explotados.

Con todo lo anterior entendemos que el término Big Data refiere a la capacidad de recoger y analizar, a gran velocidad, enormes volúmenes de datos, variados en forma y estructura, para establecer relaciones que aporten conocimiento a partir de una información que es demasiado compleja para ser captada a simple vista.

Ee decir que los métodos que definen las técnicas Big Data, fundamentalmente vinculados a la estadística computacional, el aprendizaje automático y la minería de datos

## FILMINA 3

Además, en esta emergente **era de datificación**, la transformación del mundo implica un crecimiento de datos en el mundo, segundo a segundo. Pero estos son crudos y cargados de información, como podemos apreciar en estos dos sitios que recolectan estadísticas de datos:

* Worldometers.info
* LocalIG.com FILMINA 4

Como podemos ver la información generada es mayormente textual, aproximadamente un 80% de los datos generados constantemente son textuales, y estos no son tan sencillos de analizar como los numéricos. Es decir, que este tipo de información generada es no estructurada.

Aún más entendemos que las palabras de los datos textuales, implican una **importancia relativa** à es decir que forman parte fundamental para cualquier tipo de análisis, de la misma manera que el significado que le demos a los resultados, siendo el sabor total de un estudio o proeycto.

Para estos análisis serán necesarios los aportes del big data, machine learning o inteligencia artificial, pues podremos observar que al filtrarse en los acervos comunes van a permitir diseminaciones más allá de su ámbito original de estudio. Y proporcionaran formas de solventar disyuntivas antes consideradas ineludibles.

FILMINA 5

Ahora bien ¿cómo se podrían analizar esa enorme cantidad de datos?

Para ello los invito a pensar en como incorporamos datos textuales, y es la respuesta sencilla, desde la lectura. Pero dentro de ella, diferenciaremos la ***cercana***, es decir la que practicamos constantemente en nuestra vida. Aqu[i tienen un claro ejemplo de lo que podemos entender como lectura cercana, la típica en un libro y la actual, digital con todos lo enriquecimientos que ello implica, hipertextos, por lo cual les trajo un proyecto en el cual participo, en el que desarrollamos esta nueva forma de lectura digital:

<https://hdlab.space/argentina-y-conquista-del-rio-de-la-plata/>

Volviendo a la datificación, podemos entender que ha generado nuevas formas de lectura así como de interpretaciones. Es por ello que distintos humanistas han analizado estas metodologías, como el investigador italiano Franco Moretti, que ha publicado diversos trabajos, con nuevas formulaciones, que le han otorgado visibilidad dentro del campo de los estudios histórico-literarios, pues este investigador pretendió trascender las propuestas de lectura cercana (Close Reading), como el análisis de contenido al que todos estamos acostumbrados, y procuró llevar los estudios literarios e históricos, hacia los grandes volúmenes de texto, al aplicarles una lectura no literal sino ***cuantitativa*** de grandes volúmene, que permitirían superar la interpretación, hacer evidentes estructuras, entendidas como repeticiones, regularidades, órdenes y patrones que se hacen visibles al percibir la literatura desde la *longue dureé*.

Por otro lado, Hay que entender asimismo que, para un historiador, como es mi caso y en general para cualquier humanista, tratar con datos masivos no significa lo mismo que para un analista de mercado u otro ingeniero que trabaja con big data.

Siguiendo la explicación dada por Shawn Graham, arqueologo digital, los datos masivos pueden comprenderse como una cantidad de información más grande de lo que normalmente un investigador podría abarcar en su trabajo cotidiano; es decir, que es humanamente imposible leer e interpretar de manera tradicional.

Para el científico de las humanidades, el contenido de la información suele encontrarse de **forma semiestructurada**, es decir, catalogada de tal manera 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 ello es lo que se **denomina como metadatos**, es decir, información que permite ubicar, ordenar y guardar datos de manera automatizada.

Pero hay mucha **información que no está estructurada**, como ya he mencionado, y consiste en una inmensa colección de unidades de lenguaje para cuyo análisis no es posible fiarse de la automatización, por ello en ocasiones se prefiere construir las bases de datos antes que “minar” los recursos Web para recolectar información. Así, las humanidades digitales, este campo que auna las humanidades con la tecnología, se han enfrentado a lo que puede considerarse una **nueva necesidad de lectura** en el sentido de un nuevo paradigma de interpretación de textos históricos por medio de esta lectura distante.

FILMINA 6

Es por ello, que les comparto **la tesis de Moretti**: ***las grandes escalas a las que nos enfrenta el medio digital hace 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***. Y finalmente refiere que la única manera de analizar estos inmensos corpus es mediante la *lectura distante*. Es decir, no prestar atención a los detalles, sino delimitar ciertas características, ciertos patrones, que serán comparados, a través del procesamiento computacional.

De manera similar, otro humanista digital, Matthew Jockers (2015) ha preferido hablar de *macroanálisis* para referirse a 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 lean por nosotros esas grandes cantidades de datos, de big data, de big corpus que solo ellas pueden procesar.

Veamos el siguiente ejemplo de otro proyecto en el que he trabajado, <https://hdlab.space/Relacion-de-las-cosas-sucedidas/>

Entonces debemos hacer una macrolectura o lectura distante de los grandes corpus de datos que venimos mencionando, y respecto a ellos

debemos pensar como analizarlos.

Filmina 7

En primer lugar, si estos datos fuesen en formato estructurado o numérico, podríamos utilizar bases de datos, realizar análisis numéricos o estadísticos, pues entendemos la estadística, como ciencia, método y técnica, ya que nos permite contabilizar elementos, calcular su probabilidad, proponer descripciones de la distribución, etc.

Algo similar se puede hacer con los datos no estructurados, o textos, por medio de la ***textometría o estadística textual***, donde los elementos a contabilizar pueden ser caracteres, palabras, clases de palabras, etc. y sus descripciones pueden involucrar una gran variedad de temas. Como ser 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, etc.

Allí podemos incluir a la minería de texto à que a su vez, podemos considerarla más cercana a la ciencia de datos que a la estadística, algo que ya veremos, puesto que esta metodología se ajusta donde haya datos no estructurados, es decir, en forma de texto.

Ahora bien, para poder profundizar en el análisis de texto, partamos de la **ciencia de datos, esta se define** como el estudio de **datos** que tiene como finalidad extraer información significativa para determinados estudios científicos, intereses de empresas, etc.

Asimismo posee un enfoque multidisciplinario puesto que combina principios y prácticas de las matemáticas, la estadística, la inteligencia artificial y la ingeniería de computación para poder realizar análisis de grandes cantidades de **datos**.

FILMINA 8

No vamos a detenernos en ello, pero si ahondaremos un poco más, pensando en la minería de datos, que como podrán entender se trata de proceso que realizan búsquedas de anomalías, patrones y correlaciones en grandes conjuntos de **datos** para predecir resultados. Sin embargo, con algunas diferencias y otras similitudes en el trabajo con bases de datos, estructuras, tablas, etc., la minería de textos hace uso de estos pero con datos no estructurados, como ya se mencionó, para lo cual, la idea es trasladar los textos a una base de datos semi estructurada.

Es decir que el text mining se encuentra relacionada con distintas áreas, como ser data mining, machine learning, lingusitica computacional, como pueden ver en la diapositiva.

Ahora bien, voy a agregar otras cuestioens que debemos tener en cuenta para llevar adelante análisis textual, y servirán para comprender porque realizamos determinados cálculos al realizar los script.

En término general entendemos que un texto está compuesto por palabras, la cantidad dependerá del tipo de texto, puede ser un párrafo, un tweet, un libro, un comentario de una red social, todos los escritos de un autor o libros de una biblioteca.

Sin embargo, mediante esta metodología podremos analizar palabras, estructuras semánticas y sintácticas, permitiéndonos tomar información útil de cualquier tipo de texto. Es decir, que podremos observar ¿Cómo y qué dice un texto? Entonces, esto permitirá influenciar en otros resultados, por ejemplo, como analisis de una encuesta, chatbot, etc.

FILMINA 9

Pero que debemos tener en cuenta para comprender como se realizan estas técnicas

sabemos el lenguaje es nuestra facultad para expresarnos y comunicarnos, en forma de habla u otros signos; además Comprende estilos, modos y normas en todas sus formas convirtiéndolo en uno de los sistemas más complejos que existe, dada la multitud de lenguas y estructuras que lo conforman.

Sumado que el lenguaje no solo depende del contexto sino que se encuentra en continua evolución. Por lo cual se podría intuir que su modelado o realizar determinadas predicciones podría ser una tarea, cuasi imposible.

Sin embargo, se conoce que aproximadamente el 50% del contenido de cualquier libro, artículo o conversación será de no más de las mismas 50-100 palabras, mientras que aproximadamente la otra mitad serán palabras que aparezcan solamente una o dos veces.

**Uds han trabajado en una clase anterior con Voyant Tools o AntCOnt?**

En Voyant Tools si observaron en la parte inferior, se encontraba la densidad de vocabulario, es decir que cuando más cercano a 1 es, el índice de densidad indicará que el vocabulario tiene mayor variedad de palabras, es decir, que es más denso.

Estas cuestioens fueron estudiadas de forma empírica y maetmaticamente --> por **George Zipf,** lingüista de la Universidad de Harvard, durante la década de 1940,

Él observó que la mayoría de las palabras se repetían siempre en cualquier idioma, en un libro o en un artículo. Lo interesante, es que comprendió que parecía existir 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. Está será definida como la **Ley de Zipf**

Además, como mencioné, lo más interesante es que no solo ocurre en inglés, como descubrió Zipf, sino que es observable para 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 quiere decir que la palabra «de» será la más frecuente en casi todos los textos.

Vale mencionar que La curva de Zipf es representada 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 que permiten caracterizarlo mejor y establecer comparaciones con otros textos, ya que típicamente las palabras de la cima de la curva se corresponden con las stopwords o palabras vacías (es decir que son demasiado generales, aparecen en cualquier texto) y las palabras hacia final del eje X tienen única ocurrencia (son particulares, ocurren únicamente en ese texto).

FILMINA 10

Una última cuestión sobre la parte teórica del análisis textual, es traerles muy a vuelo de pájaro, La ***estilometría*** que es el análisis estadístico de textos escritos, trata de identificar las semejanzas y diferencias que existen entre ellos para agruparlos de acuerdo con sus características lingüísticas con el objetivo de detectar señales estilísticas que puedan servir para establecer su autoría, su ubicación genérica, sus orígenes, estilo, etc.

El término fue acuñado por el filósofo polaco Wincenty Lutosławski (1863–1954) en Principes de stylométrie. Sin embargo, los primeros análisis, sin conexión con Lutosławski, son de Thomas Corwin Mendenhall, autodidacta y físico norteamericano, (1841-1924) en el trabajo «A Mechanical solution of a literary problem».

El análisis de textos automatizado, como he mencionado, es una serie de técnicas de investigación que utiliza el análisis computacional para desvelar patrones en amplios conjuntos de textos. Así se pueden obtener multitud de resultados léxicos –índices léxicos, colocaciones, concordancias– y numéricos que después hay que procesar con paquetes estadísticos, y para ello utilizaremos R, en particular para realizar este tipo de análisis se han desarrollado diversas librerías, y para estilometría en particular, Stylo, que es el paquete con la que se ha determinado estos gráficos. Contar sobre PCA de jane austen y las hermanas Bronte, recordemos que Charlotte creó “Jane Eyre”; Emily, “Cumbres Borrascosas”, y Anne, “Agnes Grey”. Y la imagen a la derecha es el de discursos. Explicar los significados.

Esta parte no la analizaremos dentro de nuestro script, pero me parece muy interesante que vean hasta que punto se puede trabajar con R y las librerías para análisis textual.

FILMINA 11

Bueno ahora para ir terminando con esta parte teórica, vamos a ver cuestiones relacionadas a aplicaciones, utilidad y resultados que podrá generar con miniería de texto:

Va a ser utilizada en diversas áreas como el marketing, sociología, humanidades, etc, por medio de herramientas como indica en la parte inferior, a través de metodologías como las que se encuentran a la izquierda para llegar a resultados como se ve a la derecha.

En resumen

FILMINA 12

Las cinco principales aplicaciones del Text Mining que podemos encontrar son las siguientes:

    \* Búsqueda de Información. Son las búsquedas que se hacen a partir de una pregunta, palabra concreta o contenido interesante. IMAGEN CLICK \*\*\*\*Por ejemplo, podemos encontrar esta aplicación en Google y su motor de búsqueda del contenido que podemos encontrar en Internet.\*\*\*\*

    \* Reconocimiento de entidades mencionadas y referencias. Se utilizan estadísticas, nombres de lugares, personas, entidades, etcétera. Con este análisis se pretende agrupar una cantidad de textos que tengan la misma palabra o información

IMAGEN CLICK \*\*\*\*como sería el caso de los Trending Topic (Tema de Tendencia) de Twitter.\*\*\*

   \* Clustering. Se agrupan los textos que reúnen criterios similares que no pueden ser vistos directamente

\*\*\*\*Sería el caso de reunir los libros por estilo de redacción o vocabulario utilizado, como se mostró en el caso de estilometría\*\*\*\*

    \* Clasificación. Se trata de la etiquetar los textos para dividirlo en categorías de forma más eficaz.

IMAGEN CLICK \*\*\*\*Por poner un ejemplo podemos ver los documentos de texto de Facebook o los vídeos de YouTube que están agrupados por ofensivos o no ofensivos, contenido violento, etcétera.\*\*\*\*

    \* Análisis de sentimientos. En base a los resultados estadísticos podremos saber si un determinado texto es positivo, negativo o incluso neutral.

IMAGEN CLICK \*\*\*\*En este caso podríamos hablar de Tripadvisor, una app donde podemos acceder a los comentarios sobre restaurantes, hoteles etcétera. A través del análisis de texto podemos ver cuántos comentarios son positivos y cuántos negativos.\*\*\*\*

FILMINA 13

Acá les comparto una lista de ***donde utilizar text mining*** que salen de las aplicaciones que recién les enumeraba***.***

FILMINA 14

Por último, y este será el recorrido metodológico que seguiremos, tendremos primero:

1. Recolección: Recopilación de datos de diferentes recursos, tales como sitio web, correos electrónicos, comentarios de clientes, archivo de documentos. Dependiendo de la aplicación, este proceso puede ser completamente automatizado o guiado por una persona encargada de realizar este proceso.

2. Preprocesamiento: La identificación del contenido y la extracción de características representativas

\* Limpieza de textos: eliminación de cualquier información innecesaria o no deseada, como los anuncios de las páginas.

\* Tokenización: un ordenador sólo «ve» una cadena de caracteres, sin poder identificar, por ejemplo, párrafos, frases o palabras. La Tokenización divide el texto en entidades significativas (palabras, oraciones, etc.) dados los espacios en blanco presentes y las puntuaciones.

\* Extracción de características (también llamada selección de atributos): es el proceso de caracterización.

3. Luego la creación de índice de ciertos términos, sus ubicaciones y números. Esto permite un acceso rápido y una estructuración de los datos procesados.

4. Minería, propiamente dicha, el texto ha sido preprocesado y ahora se puede "extraer". Para ello, aplicamos diferentes técnicas de exploración de datos para revelar nuevos conocimientos.

5. finalmente el análisis: Los pasos de la minería producen resultados brutos. Estos deben evaluarse y visualizarse para que puedan interpretarse con respecto a las preguntas que el minero de texto quiere investigar.

Bien, ahora que pudimos ver las utilidades, lo que yo traigo como práctica son algunos script para poder entender el método estadístico, el método computacional y como leerlos, porque es necesario entender que hay detrás así como comprender el contexto donde estamos trabajando

Es decir, no vamos a hacer un desarrollo de script ni tampoco de análisis estadísticos sino que buscaremos comprender como utilizar la metodología general para poder llegar a un análisis estilístico textual.

Entonces la idea es aprender a manejar las herramientas, y evaluar si el modelo que utilizamos se ajusta a nuestros requerimientos.

### # Tidyverse y Tidytext

Estas librerías serán fundamentales para la parte más interesante de este módulo, donde aprenderemos a limpiar y transformar datos en R. Este trabajo es indispensable para el text mining, que realizado en R, contará con varias ventajas: no requiere muestreo para extraer información, permite aplicar varias metodologías de forma rápida y genera procedimientos auditables y reproducibles. El principio básico de un dataframe 'tidy' es que cada columna representa una variable y cada línea, una observación.

La librería Tidytext permite realizar minería de texto con el uso de herramientas tidy, generando datos estructurados y tidy a partir de datos textuales.

Nos serviremos de la librería tidytext para 'tokenizar', es decir, sustituir algunos datos por sustitutos sin alterar el tipo o la longitud de los datos.

# un poco de análisis textual a nivel morfosintáctico, y realizaremos un etiquetado gramatical automático (PoS Tagging)

# etiquetado gramatical es una forma de anotar las palabras de un texto con la categoría gramatical o clase de palabras (sustantivo, preposición, verbo, determinante…) correspondiente con una serie de etiquetas, es muy desarrollado y trabajado por expertos de NLP (Natural Language Processing)

Vamos a utilizar para ello librería UdPipe, desarrollada por el [Instituto de linguistica formal y aplicada de la Universidad de la República Checa](https://ufal.mff.cuni.cz/udpipe), que tiene un modelo para procesar texto en castellano

Luego, deberemos descargar el modelo del idioma que nos interesa.

Vamos a empezar a parsear nuestro corpus de oraciones, y anotar qué tipo de componente es cada palabra

pre-procesamiento de texto:

* tokenización: la unidad mínima del análisis es ahora cada palabra (notá que por eso hay muchas mas filas que antes). El doc\_id nos permitirá volver a unir las piezas cuando hagamos tareas por oraciones;
* para cada palabra se ha anotado el tipo en upos;
* se ha convertido la palabra a su raíz en lemma, La lemmatización es un procedimiento que busca reducir las palabras a su forma no flexionada o conjugada

primero

* modelo en español
* vamos a utilizar un vector con todo nuestro corpus, que ya tenemos realizado, al que le pasaremos el modelo en español, a su resultado lo convertiremos en tibble, para separar los tipos de palabras
* en dicho tibble, vamos a poder observar los tipos de palabras
* para visualizar los resultados primero generamos un lista con las frecuencias y tipo de palabras, para ello agruparemos con usar groupyby() y con summarise()contará la cantidad de tokens por categoría gramatical. Y mediante arrange()ordenara la tabla decrecientemente
* Los NA los eliminaremos usando drop\_na()
* Visualizaremos la frecuencia absoluta de cada categoría con un gráfico de barras
* .
* Ahora vamos a realizar otro tipo de grafico con cada tipo de palabras para ello modelaremos nuevamente nuestro corpus, y lo etiquetaremos con udpipe\_annotate()
* Primero haremos un filtro por sustantivo, agruparemos por token, e indicaremos que nos grafique los primeros 30, por orden, e invertiremos el orden para poder apreciar mejor la lectura de las palabras
* En el caso de verbos y adjetivos graficaremos pero por lemma, para poder considerar la forma canónica de la palabras