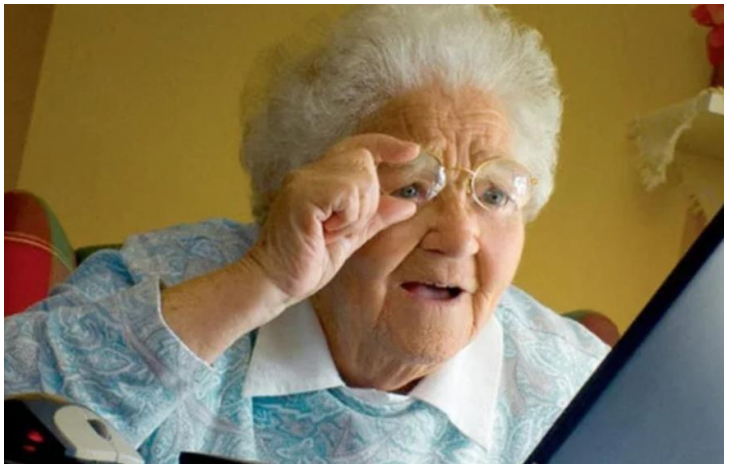
CONFERENCIA GRATUITA

AGENDA



- Minería de datos
- Clustering
- Análisis de sentimientos
- Una aplicación a Text Mining

Importancia de este tipo de charlas

ARTICULACIÓN CON NUESTRAS CARRERAS:

- *Lic. en Sistemas de la información*
- *Contadorxs*
- *Administradorxs*
- *Actuarixs*
- *Economistas*

Material disponible en:

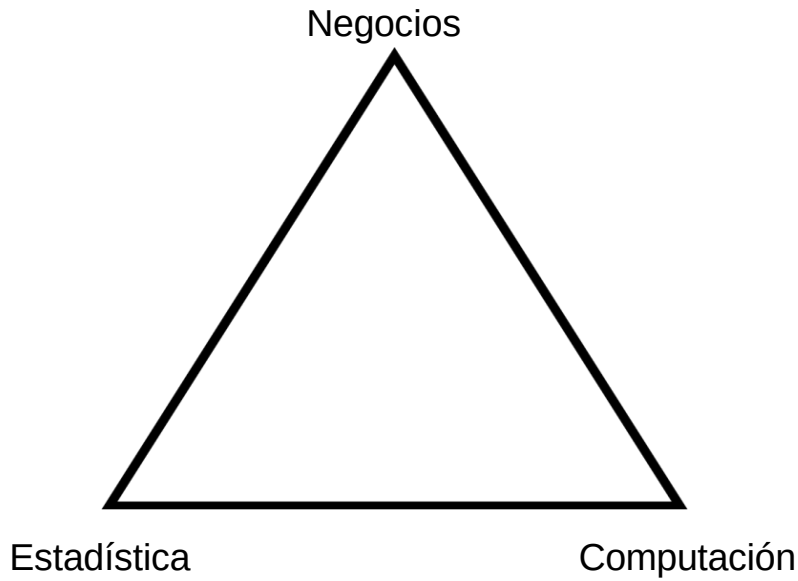
https://github.com/rodrigodelrosso/CECE-Nuevo-Espacio-Text-Mining

Minería de Datos

¿Qué es la Minería de datos?

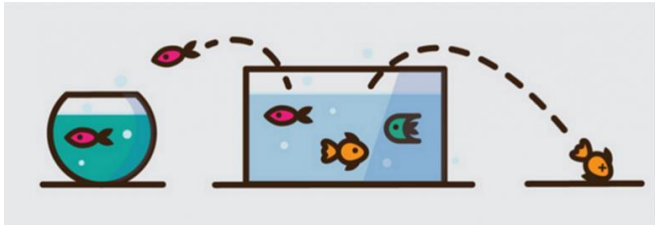
"*At a high level, **data science** is a set of fundamental principles that guide the extraction of knowledge from data. Data mining is the extraction of knowledge from data, via technologies that incorporate these principles.*" (Provost & Fawcett, 2013)

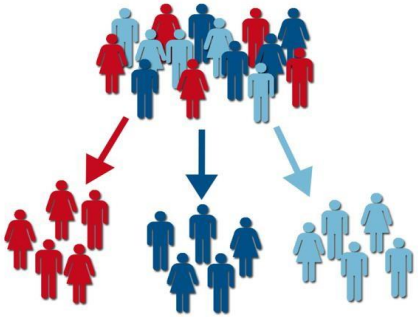
"*Data mining is a business process for exploring large amounts of data to discover meaningful patterns and rules.*" (Gordon & Berry, 2011). En este sentido, se requiere de habilidades en (al menos) tres campos:

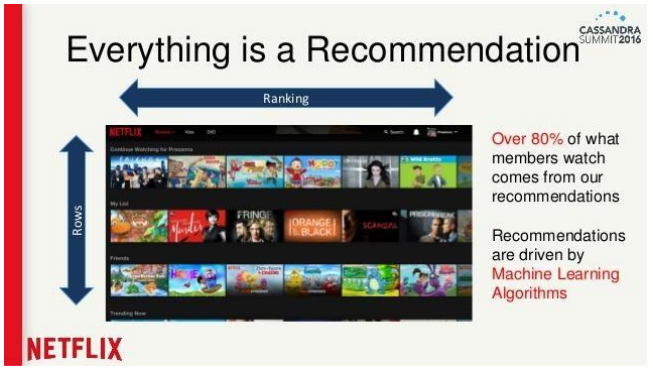


Aplicaciones en Negocios

- Churn (attrition) de clientes.
- Segmentación de clientes.
- Recomendación de productos.
- Publicidades personalizadas.
- Credit scoring.
- Predicción de valor de un cliente.
- Análisis de sentimiento.







Aprendizaje Automático (panorama)

Existen diversas técnicas para descubrir patrones en grandes volúmenes de datos (por ejemplo, a través de la exploración "manual").

En este curso estará centrado en utilizar técnicas de aprendizaje automático aplicados al ámbito de las Ciencias Actuariales.

Definición “casi canónica”:

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." (Mitchell, T., 1997, Machine Learning).

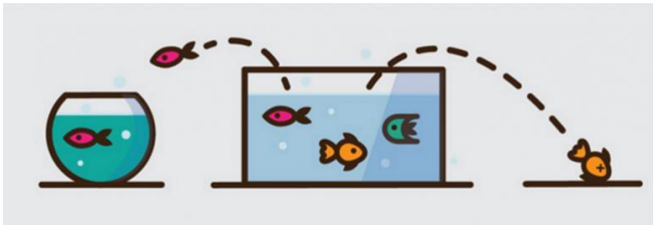
- *Computer program* → modelo estadístico.
- *Experience E* → datos.
- *Task T* → tarea (e.g., predecir el peso de un chico de acá a medio año).
- *Performance measure P* → medida de performance (e.g., *logloss*).

Aprendizaje Automático (panorama)

Existen tres grandes familias de algoritmos de Aprendizaje Automático (Supervisado, No Supervisado, Por Refuerzos).

1. Supervisado

"*For each observation of the predictor measurement(s)* x_i, $i = 1, ..., n$ *there is an associated response measurement* y_i".



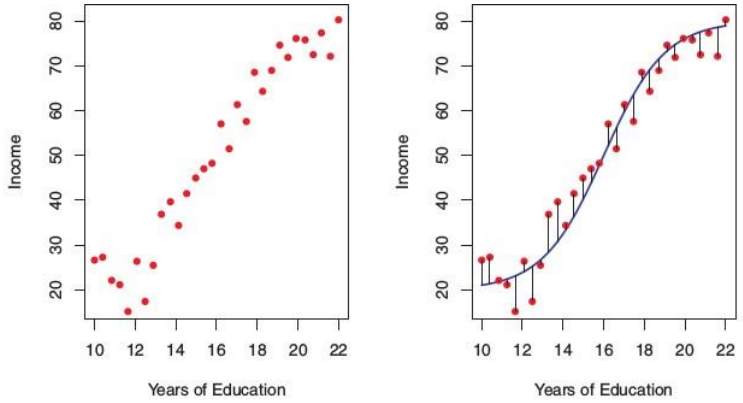
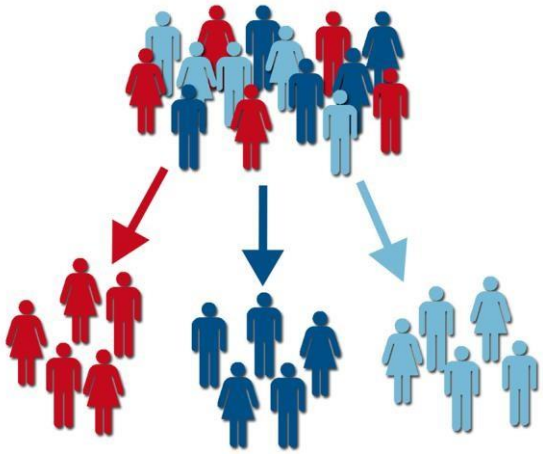


FIGURE 2.2. *The* `Income` *data set.* Left: *The red dots are the observed values of* `income` *(in tens of thousands of dollars) and* `years of education` *for* 30 *individuals.* Right: *The blue curve represents the true underlying relationship between* `income` *and* `years of education`*, which is generally unknown (but is known in this case because the data were simulated). The black lines represent the error associated with each observation. Note that some errors are positive (if an observation lies above the blue curve) and some are negative (if an observation lies below the curve). Overall, these errors have approximately mean zero.*

2. No Supervisado

"*Unsupervised learning describes the somewhat more challenging situation in which for every observation i = 1, ..., n, we observe a vector of measurements* x_i *but no associated response* y_i *… We can seek to understand the relationships between the variables or between the observations*".



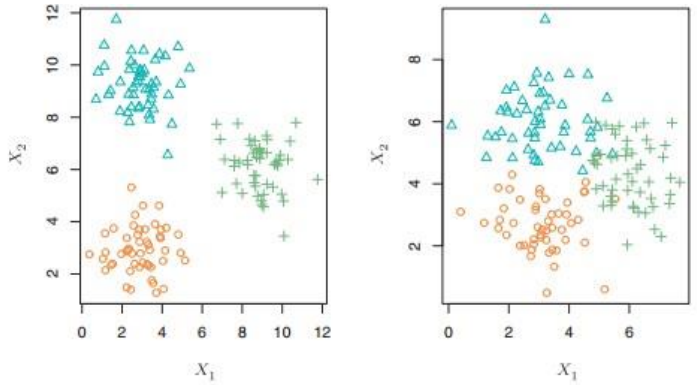
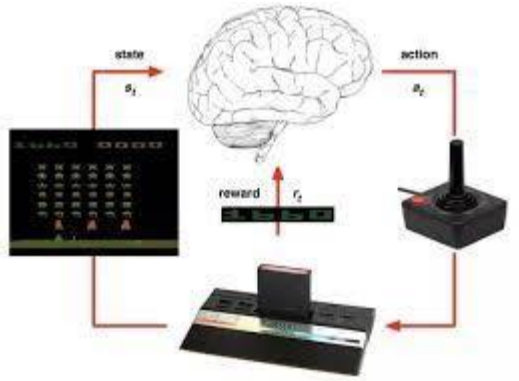


FIGURE 2.8. *A clustering data set involving three groups. Each group is shown using a different colored symbol.* Left: *The three groups are well-separated. In this setting, a clustering approach should successfully identify the three groups.* Right: *There is some overlap among the groups. Now the clustering task is more challenging.*

3. Por Refuerzos

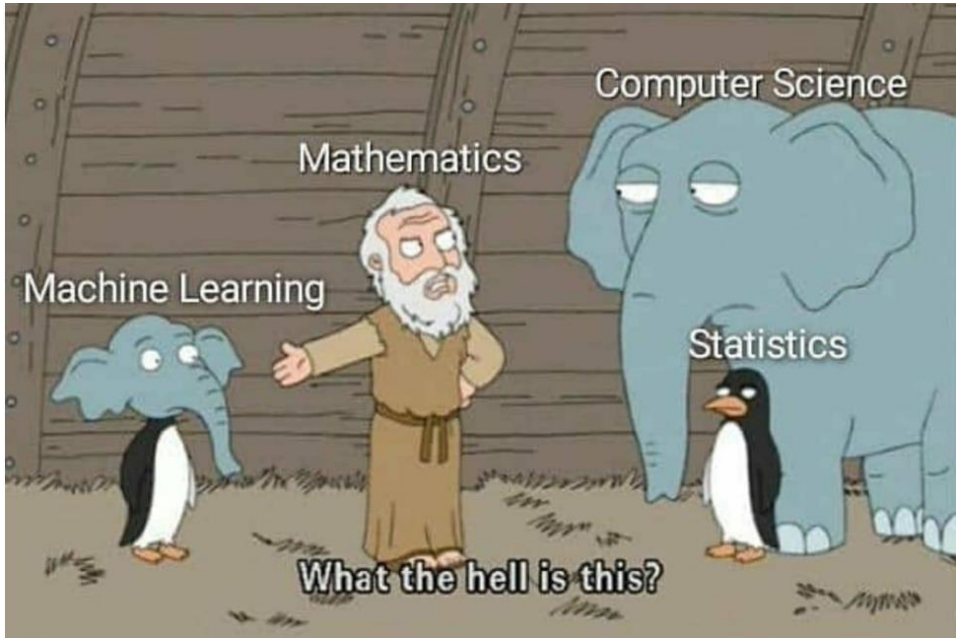
A grandes rasgos, se enfoca en lograr que agentes maximicen un beneficio esperado mediante la interacción con un ambiente (el cual muchas veces no conocen)... Muchos problemas de la realidad se ajustan a este esquema.



ADVERTENCIA

Desde nuestra humilde apreciación, aquí debemos parar un segundo y aclarar algo.

Necesitas tomar riendas de los procesos y GENERAR VALOR ORGANIZACIONAL con los modelos que vamos a utilizar.



También, repensar y construir RESPONSABLEMENTE un concepto de gobernanza de la información y la ética en la utilización de datos "públicos".

Aprendizaje Supervisado

Lo que caracteriza al aprendizaje supervisado es que uno quiere predecir una variable. Puede haber dos tipos de variables a predecir:

Continua → regresión
Categórica → clasificación (binaria, multiclase)

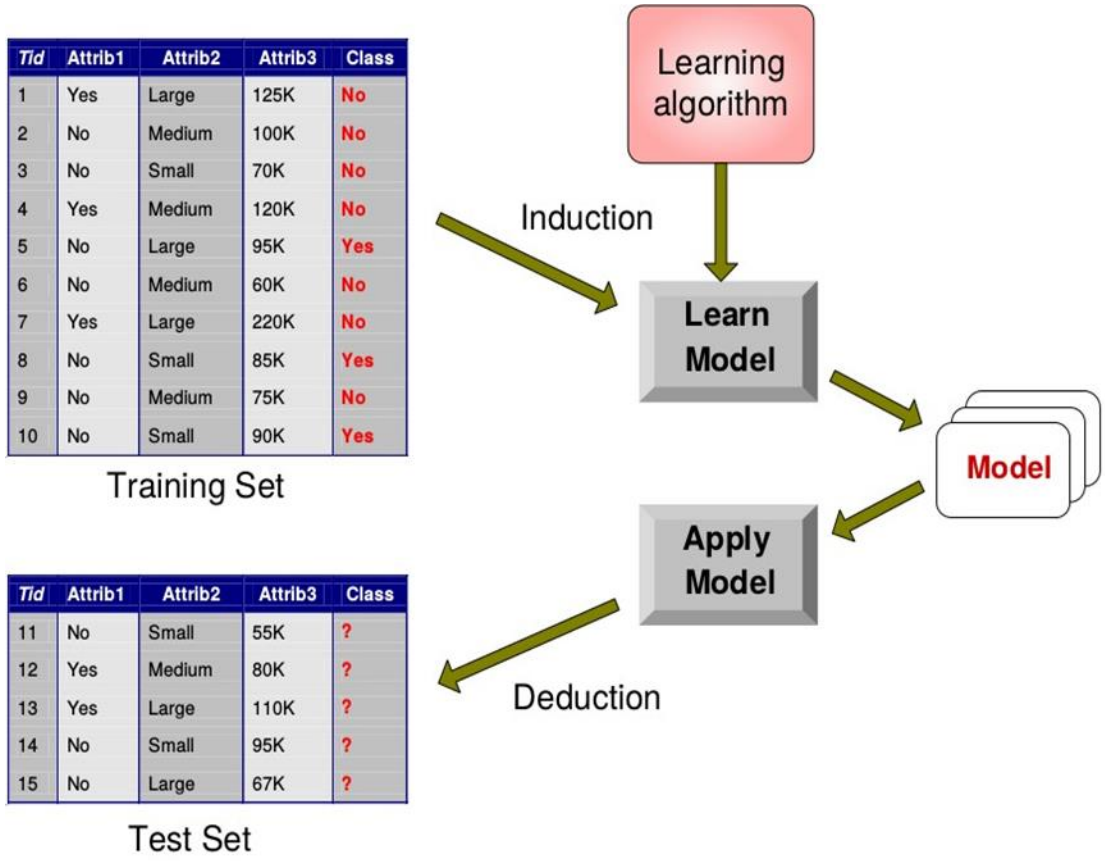
$$Y = f(X)$$

Distintos problemas se atacan con distintos algoritmos (algunos sirven tanto para regresión como clasificación).

¿Qué tipo de problema son los siguientes?

- Predecir la cantidad de ventas de líneas telefónicas en un día dado.
- Predecir si un cliente en particular se dará de alta dada nuestra campaña de marketing.
- Predecir si un día determinado tendremos o no más de 200 líneas nuevas vendidas.

Aprendizaje Supervisado



$$Y = f(X)$$

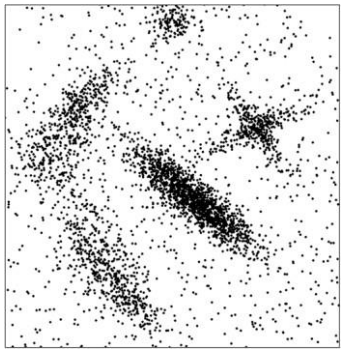
Clustering

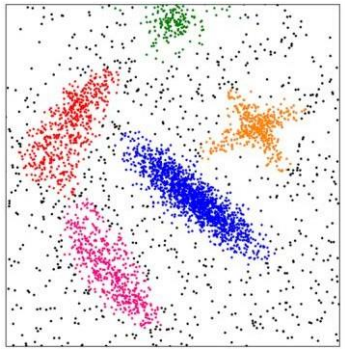
Clustering se refiere a una serie de técnicas cuyo objetivo es encontrar subgrupos o "*clusters*" en un conjunto de datos.

La idea es particionar los datos de tal manera que:

1. Las observaciones que pertenecen a un grupo sean similares entre ellas.
2. Las observaciones de grupos distintos sean distintas entre ellas.

Esto plantea el "problema" de definir cuando dos observaciones son similares o distintas entre sí. (Algo que en gran medida puede depender del dominio en donde se esté trabajando.)





Clustering

Existen múltiples familias de algoritmos de clustering:

1. *Partitioning*
2. *Hierarchical*
3. *Density-based*
4. *Grid-based*
5. *Model-based*

Nosotros veremos los dos algoritmos más tradicionales:

- *K-means clustering* (partitioning)
- *Hierarchical*

Clustering

Divide al conjunto de datos en K subconjuntos distintos sin solapamiento. Uno debe fijar el valor de K antes de correr el algoritmo de K medias.

Los clusters deben cumplir las siguientes condiciones:

1. $C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$. In other words, each observation belongs to at least one of the K clusters.

2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$. In other words, the clusters are non-overlapping: no observation belongs to more than one cluster.

Clustering

K-means asume que una buena asignación es aquella que dado un valor de K minimiza lo más posible la variabilidad intra cluster (*within-cluster variation*).

Si $W(C_j)$ es una medida que indica cuánto las observaciones de un cluster j difieren entre ellas, el problema se puede escribir como:

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

Si uno usa la distancia euclídea como medida de disimilaridad $W(C_j)$ puede escribirse de la siguiente manera:

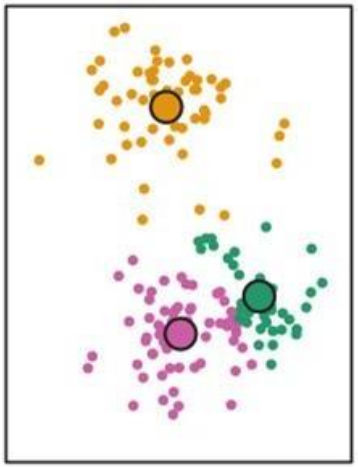
$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

De esta forma el problema se puede reescribir como:

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

En donde $W(C_j)$ se puede expresar como la distancia a un centroide:

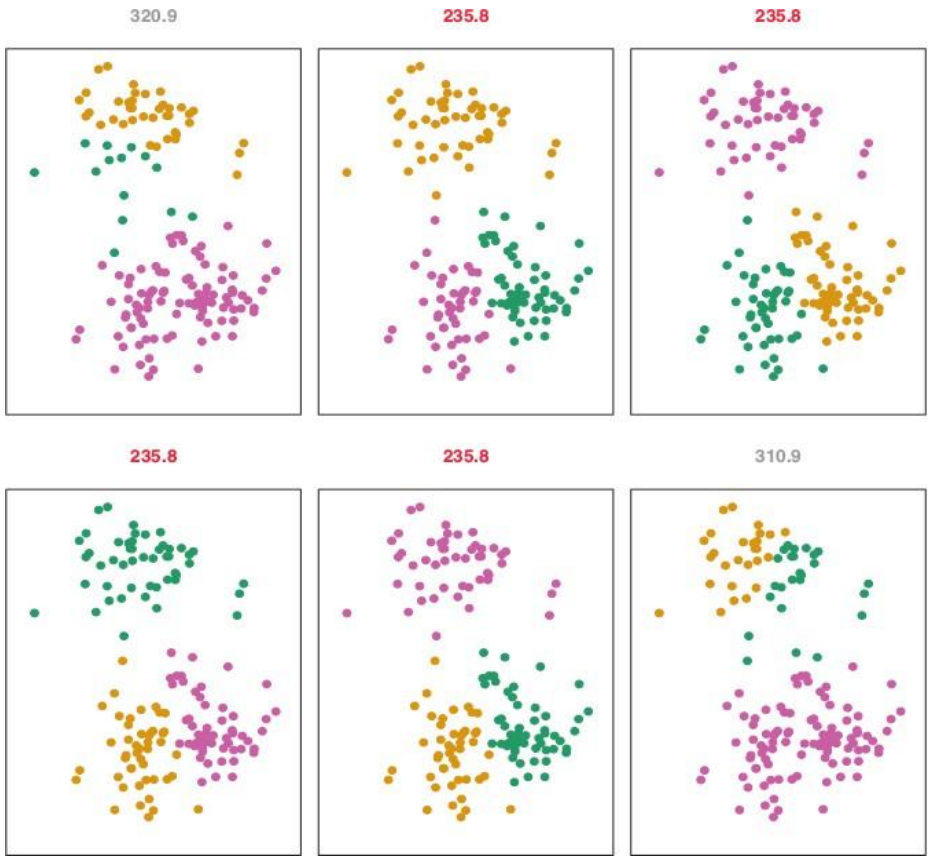
$$\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$



Clustering

La solución obtenida por K-means depende en gran medida de los valores iniciales de asignación a clusters.

Por este motivo se suele correr el algoritmo muchas veces y quedarse con la mejor solución.



Clustering

Los algoritmos jerárquicos requieren más tiempo…

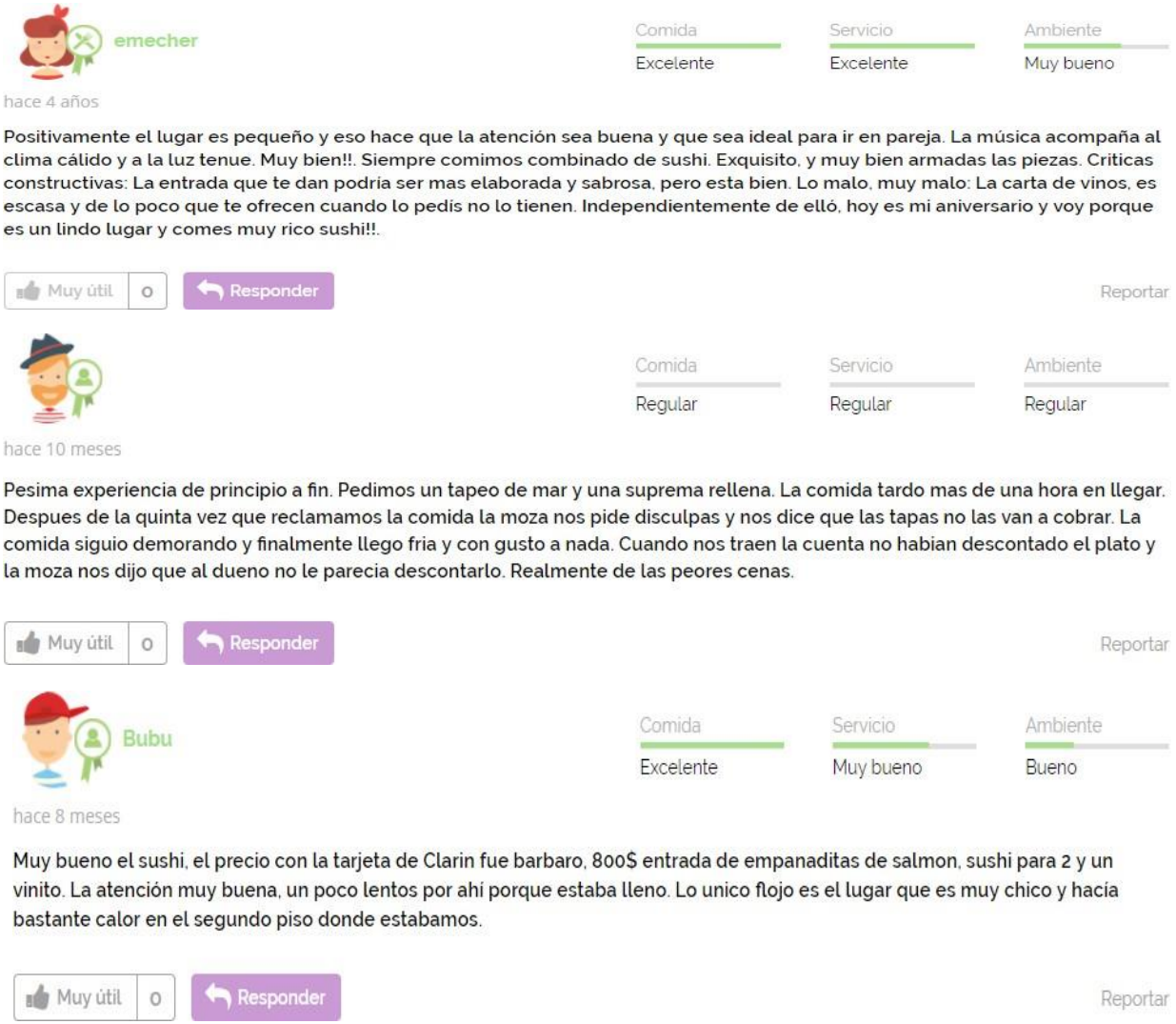
¡Lo dejamos para otra charla que nos inviten desde el CECE-Nuevo Espacio!

Análisis de Sentimiento

Análisis de sentimiento

Caso de estudio:

Guía Óleo



emecher

hace 4 años

Comida	Servicio	Ambiente
Excelente	Excelente	Muy bueno

Positivamente el lugar es pequeño y eso hace que la atención sea buena y que sea ideal para ir en pareja. La música acompaña al clima cálido y a la luz tenue. Muy bien!!. Siempre comimos combinado de sushi. Exquisito, y muy bien armadas las piezas. Criticas constructivas: La entrada que te dan podría ser mas elaborada y sabrosa, pero esta bien. Lo malo, muy malo: La carta de vinos, es escasa y de lo poco que te ofrecen cuando lo pedís no lo tienen. Independientemente de elló, hoy es mi aniversario y voy porque es un lindo lugar y comes muy rico sushi!!.

Muy útil 0 Responder Reportar

hace 10 meses

Comida	Servicio	Ambiente
Regular	Regular	Regular

Pesima experiencia de principio a fin. Pedimos un tapeo de mar y una suprema rellena. La comida tardo mas de una hora en llegar. Despues de la quinta vez que reclamamos la comida la moza nos pide disculpas y nos dice que las tapas no las van a cobrar. La comida siguio demorando y finalmente llego fria y con gusto a nada. Cuando nos traen la cuenta no habian descontado el plato y la moza nos dijo que al dueno no le parecia descontarlo. Realmente de las peores cenas.

Muy útil 0 Responder Reportar

Bubu

hace 8 meses

Comida	Servicio	Ambiente
Excelente	Muy bueno	Bueno

Muy bueno el sushi, el precio con la tarjeta de Clarin fue barbaro, 800$ entrada de empanaditas de salmon, sushi para 2 y un vinito. La atención muy buena, un poco lentos por ahí porque estaba lleno. Lo unico flojo es el lugar que es muy chico y hacía bastante calor en el segundo piso donde estabamos.

Muy útil 0 Responder Reportar

Dr. Martín E. Masci – Mg. Rodrigo Del Rosso

Análisis de sentimiento

¿Podremos armar un sistema que aprenda a detectar cuando se está hablando bien o no de la comida de un restaurant?

Análisis de sentimiento

Pensémoslo como Aprendizaje Supervisado:

y

Comida	Servicio	Ambiente
Regular	Regular	Regular

hace 10 meses

X

Pesima experiencia de principio a fin. Pedimos un tapeo de mar y una suprema rellena. La comida tardo mas de una hora en llegar. Despues de la quinta vez que reclamamos la comida la moza nos pide disculpas y nos dice que las tapas no las van a cobrar. La comida siguio demorando y finalmente llego fria y con gusto a nada. Cuando nos traen la cuenta no habian descontado el plato y la moza nos dijo que al dueno no le parecia descontarlo. Realmente de las peores cenas.

Muy útil 0 Responder Reportar

bag-of-words model

¿Cómo podemos incorporar texto a modelos como los que vimos hasta ahora?

Terminología:

- A una colección de textos se la llama corpus.
- A un elemento del corpus se lo llama documento.
- Los documentado pueden tener metadatos asociados (e.g., la clase que queremos predecir).

Objetivo: representar al corpus como una matriz de números (de modo de
poder aplicar las técnicas vistas en la materia).

Para ello vamos a usar lo que se conoce como el *bag-of-words model*.

Tokenización:

Un texto no es más una secuencia de caracteres. Nosotros entendemos a un texto como una secuencia de palabras.

Tokenizar es la acción se dividir un conjunto de caracteres en una secuencia de palabras (o tokens).

En español bien escrito es simple: separar por caracteres que no sean alfanuméricos.

Sin embargo, la realidad es más cercana a esto:

"felicidadees!! k t lo pases muy bien!! =)Feeeliiciidaaadeeess !! (:

Felicidadesss!!pasatelo genialll :DFeliicCiidaDesS! :D Q tte Lo0 pases bN! ;) (heart)"

¿Emojis/emoticones tendrán información relacionada al sentimiento hacia algo?

Problemas

Problemas de bag-of-words model (algunos):

- Perdemos toda información referida al orden de las palabras (¿Toy Dog == Dog Toy?).
- Da lugar a una matriz rala. En en el caso dtm, muchas columnas que en general valen 0 para la mayoría de los documentos (aun así, la podemos trabajar).
- Ignora el contexto de las palabras (e.g., que estén negadas).
- Ignora similitud semántica entre palabras (¿Auto != Automóvil?).
- No contempla la polisemia (¿qué significa banco, gato o cresta?)

Pre-procesamiento

Pre-procesamiento que se suele hacer en bag-of-words:

- Pasar todo a minúsculas.
- Quitar stopwords (palabras que suelen no tener significado).
- Ignorar palabras poco frecuentes.
- Asignar part-of-speech tags a las palabras.
- Reducir formas inflecionales de palabras a su forma común:
  - Stemming: son heurísticas que quitan el final de la palabra para lograr este objetivo (having → hav)
  - Lematización: busca hacer esto de manera apropiada usando un vocabulario y análisis morfológico (having → have).

Text Mining

Text Mining

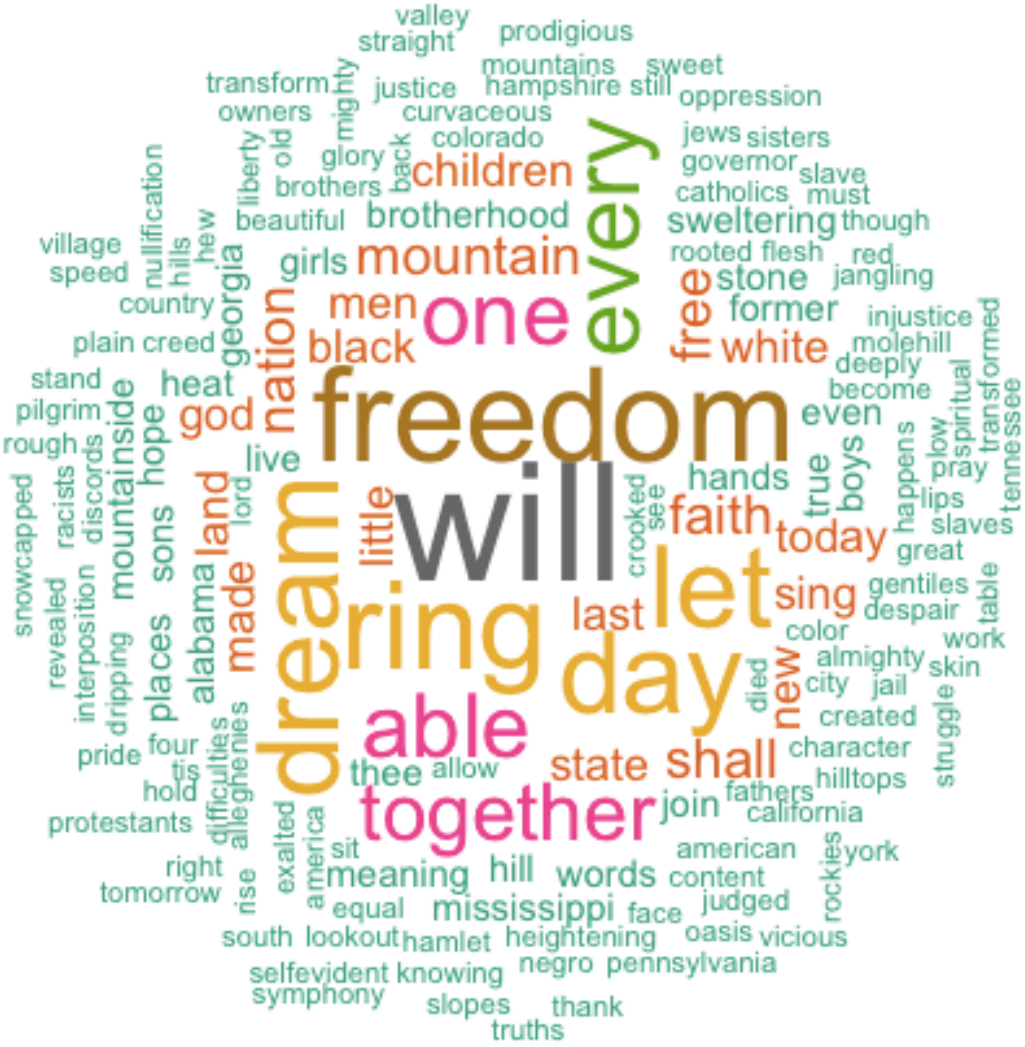
Los métodos de minería de texto (text mining) nos permiten resaltar las palabras clave más utilizadas en un párrafo de textos.

Se puede crear una nube de palabras (word cloud), también conocida como nube de texto (text cloud) o nube de etiquetas (tag cloud), que es una representación visual de los datos de texto.

El procedimiento para crear nubes de palabras es muy simple en R si conoce los diferentes pasos a ejecutar.

El paquete de minería de texto (text mining) (tm) y el paquete generador de nube de palabras (wordcloud) están disponibles en R para ayudarnos a analizar textos y visualizar rápidamente las palabras clave como una nube de palabras.

Nube de palabras



Ventajas

3 razones por las que deberías usar nubes de palabras para presentar tus datos de texto:

1. **Las nubes de palabras agregan simplicidad y claridad.**

2. **Las palabras clave más utilizadas se destacan mejor en una nube de palabras.**

3. **Las nubes de palabras son una potente herramienta de comunicación. Son fáciles de entender, de compartir y son impactantes.**

Las nubes de palabras son visualmente atractivas que los datos de una tabla

¿Quién está usando nubes de palabras?

- Investigadores: para informar datos cualitativos
- Comercializadores: para resaltar las necesidades y los puntos débiles de los clientes
- Educadores: para apoyar temas esenciales
- Políticos y periodistas
- Sitios de redes sociales: para recopilar, analizar y compartir los sentimientos de los usuarios

¿Dudas?

“Lo importante es no dejar de hacerse preguntas”

Albert Einstein

martinmasci@economicas.uba.ar

rdelrosso@economicas.uba.ar