Minería de Datos

UTDT

Master in Management + Analytics

Sesión 6

¿Qué vimos hasta ahora?

- Qué es la minería de datos.
- Distintos tipos de aprendizaje automático.
- Aprendizaje supervisado.
- Naïve bayes.
- K-nearest neighbors.
- Noción de overfitting & underfitting
- Árboles de decisión
- Selección de modelos
- Introducción a Caret
- Ingeniería de atributos (peligro de data leaking)
- Métricas de performance de clasificadores
- Introducción a XGboost
- K-means
- Clustering jerárquico

Reglas de asociación

MERCADONA S.A.

AVDA. CADIZ, 93 - C/ RONDA
CORDOBA
Teléfono 957481935
CIF: A-46103834
P.V.P. I.V.A. INCLUIDO
01/10/2013 09:35 OP: 078910
N: 969951 CAJA:010 SUP: 2145

1	1	ROLLOS COCINA	2,35
2	1	C.POLLO/PAVO/L	4,25
3	1	PAN BLANCO 2B	1,00
4	1	ESPINACAS	1,19
5	1	LECHUGA CORAZN	1,55
6	1	QUESO RALLADO	1,15
7	1	NATA COCINAR	1,05

TOTAL.....EUROS 12,54
ENTREGA.....EUROS 20,00

DEVOLUCIÓN..EUROS 7,46

DETALLE

IVA	Base Imponible	Cuota
4 %	3,74	0,15
10 %	5,73	0,57
21 %	1,94	0,41
TOTAL:	11,41	1,13

LE ATENDIÓ: TRINI

Reglas de asociación

Dado un conjunto de transacciones se quiere encontrar reglas que predigan la ocurrencia de un(nos) ítem(s) a partir de otros ítems de la transacción.

Transacciones de una cesta de compras
(Market-Basket transactions)

TID	*Items*
1	Pan, Leche
2	Pan, Pañales, Cerveza, Huevos
3	Leche, Pañales, Cerveza, Coca
4	Pan, Leche, Pañales, Cerveza
5	Pan, Leche, Pañales, Coca

Ejemplo de Reglas de asociación:

{Pañales} $\rightarrow$ {Cerveza}
{Leche, Pan} $\rightarrow$ {Huevos, Coca}
{Cerveza, Pan} $\rightarrow$ {Leche}

Reglas de asociación

- ***Itemset***
  - Colección de uno o más items
    - Ejemplo: {Leche, Pan, Pañales}
  - k-itemset
    - Itemset con k items
- ***Support count* (σ) de un itemset**
  - Cantidad de transacciones que contienen al itemset
  - Ejemplo σ({Leche, Pan, Pañales}) = 2
- ***Support* de un itemset**
  - Transacción que contienen el itemset / Total de transacciones
  - Ejemplo: s({Leche, Pan, Pañales}) = 2/5
- ***Frequent Itemsets* (dado un *minsup*)**
  - Itemsets cuyo soporte es mayor o igual que un umbral *minsup (minimo soporte)*

TID	*Items*
1	Pan, Leche
2	Pan, Pañales, Cerveza, Huevos
3	Leche, Pañales, Cerveza, Coca
4	Pan, Leche, Pañales, Cerveza
5	Pan, Leche, Pañales, Coca

Reglas de asociación

Regla de asociación

- Implicación de la forma X → Y, donde X e Y son itemsets
- Ejemplo: {Leche, Pañales} → {Cerveza}

Dos métricas asociadas a una regla

- Soporte (*Support*) (s)
  - Porcentaje de transacciones que contienen a X e Y sobre el total
- Confianza *(Confidence)* (c)
  - Cantidad de transacciones que contienen a X e Y sobre las que contienen a X. Mide la frecuencia de ocurrencia de los ítems de Y en las transacciones que contienen a X.

TID	*Items*
1	Pan, Leche
2	Pan, Pañales, Cerveza, Huevos
3	Leche, Pañales, Cerveza, Coca
4	Pan, Leche, Pañales, Cerveza
5	Pan, Leche, Pañales, Coca

$$\{\text{Leche}, \text{Pañales}\} \Rightarrow \text{Cerveza}$$

$$s = \frac{\sigma(\text{Leche}, \text{Pañales}, \text{Cerveza})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Leche}, \text{Pañales}, \text{Cerveza})}{\sigma(\text{Leche}, \text{Pañales})} = \frac{2}{3} = 0.67$$

Reglas de asociación

TID	Items
1	Pan, Leche
2	Pan, Pañales, Cerveza, Huevos
3	Leche, Pañales, Cerveza, Coca
4	Pan, Leche, Pañales, Cerveza
5	Pan, Leche, Pañales, Coca

Ejemplos de reglas:

{Leche, Pañales} → {Cerveza} (s=0.4, c=0.67)
{Leche, Cerveza} → {Pañales} (s=0.4, c=1.0)
{Pañales, Cerveza} → {Leche} (s=0.4, c=0.67)
{Cerveza} → {Leche, Pañales} (s=0.4, c=0.67)
{Pañales} → {Leche, Cerveza} (s=0.4, c=0.5)
{Leche} → {Pañales, Cerveza} (s=0.4, c=0.5)

Observaciones:

• Todas estas reglas son particiones binarias del mismo itemset: {Leche, Pañales, Cerveza}

• Las reglas que se originan del mismo itemset tienen el mismo soporte pero pueden tener diferente confianza

Algoritmo apriori

Problema:

¿Cómo encontrar las reglas que cumplan un soporte mínimo y una confianza mínima?

¿Esto es aprendizaje supervisado o no supervisado?

Hacer todos las combinaciones posibles (fuerza bruta) es muy costoso (reglas posibles: $3^d-2^{d+1}+1$).

Vamos a dividir al problema en dos grandes pasos:

1. Descubrir todos los itemsets frecuentes.
2. Descubrir las reglas con suficente confianza contenidas en esos itemsets.

1. Descubrimiento de todos los itemsets frecuentes

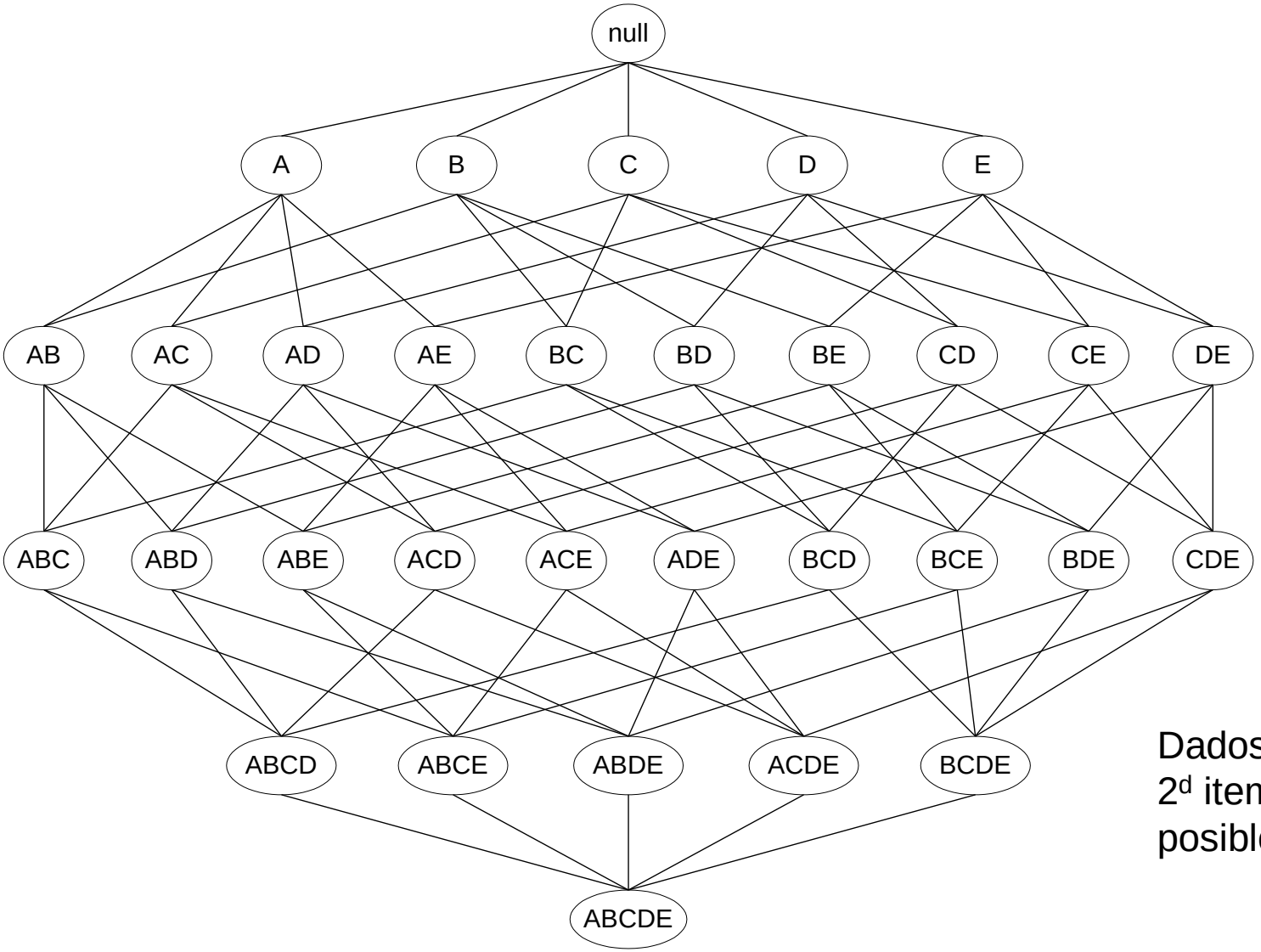
- Se manejan 2 conjuntos de itemsets
  - Candidatos (C_k) y Frecuentes (L_k)

- Se aprovecha una propiedad que cumple el soporte:
  - Propiedad anti-monótona de f:

$$X \subseteq Y \rightarrow f(X) \geq f(Y)$$ (siendo X e Y conjuntos de ítems)

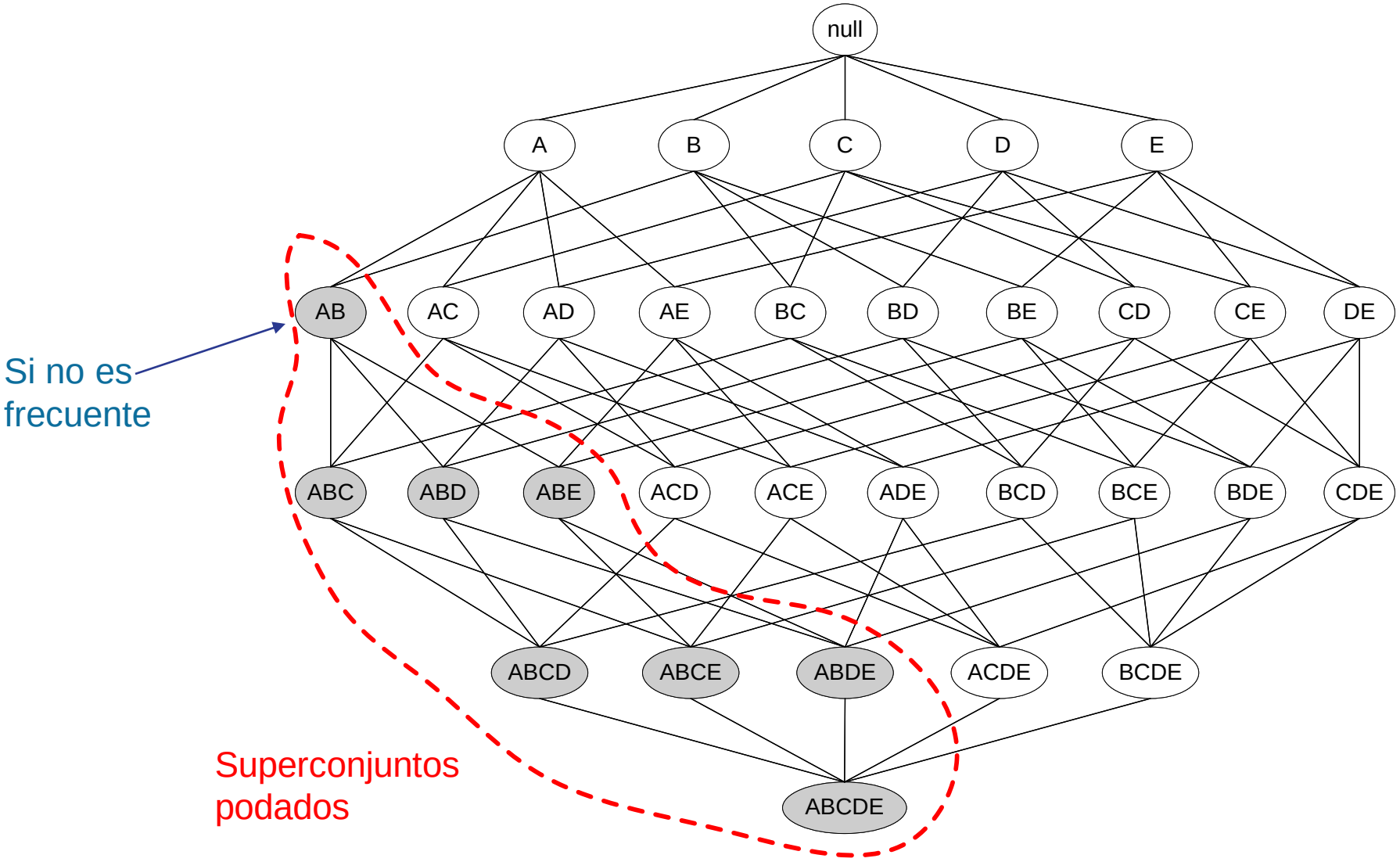
  - Support cumple con la propiedad anti-monótona.
  - ¿Qué implica? un itemset no puede ser frecuente si algún itemset contenido en él no lo es.

1. Descubrimiento de todos los itemsets frecuentes



Dados d items, existen 2^d itemsets como posibles candidatos

1. Descubrimiento de todos los itemsets frecuentes



1. Descubrimiento de todos los itemsets frecuentes

Algoritmo:

Create L_1 = set of supported itemsets of cardinality one
Set k to 2
while $(L_{k-1} \neq \emptyset)$ {
    Create C_k from L_{k-1}
    Prune all the itemsets in C_k that are not
        supported, to create L_k
    Increase k by 1
}
The set of all supported itemsets with at least two members is $L_2 \cup \cdots \cup L_{k-2}$

1. Descubrimiento de todos los itemsets frecuentes

Ilustración:

Item	Cantidad
Cerveza	**3**
Coca	**2**
Huevos	**1**
Leche	**4**
Pan	**4**
Pañales	**4**

1-itemsets

Itemset	Cantidad
{Cerveza, Leche}	**2**
{Cerveza, Pan}	**2**
{Cerveza, Pañales}	**3**
{Leche, Pan}	**3**
{Leche, Pañales}	**3**
{Pan, Pañales}	**3**

2-itemsets

(No es necesario generar los candidatos que involucran Coca o Huevos)

conteo de soporte mínimo = 3

3-itemset

Itemset	Cantidad
{Leche, Pan, Pañales}	**3**

¿Por qué no {Cerveza, Leche, Pan}?

1. Descubrimiento de todos los itemsets frecuentes

Factores que influyen en la complejidad computacional del algoritmo:

- Elección del umbral mínimo de soporte.
- Número de ítems en el conjunto de datos.
- Cantidad de transacciones.
- Cantidad promedio de ítems por transacción.

2. Descubrimiento de las reglas con suficiente confianza

- Dado un itemset frecuente L, encontrar todos los subconjuntos $f \subset L$ tales que $f \rightarrow L - f$ satisface el requerimiento mínimos de confianza.

- Si {A,B,C,D} es un itemset frecuente, las reglas candidatas son:

$ABC \rightarrow D$,	$ABD \rightarrow C$,	$ACD \rightarrow B$,	$BCD \rightarrow A$,
$A \rightarrow BCD$,	$B \rightarrow ACD$,	$C \rightarrow ABD$,	$D \rightarrow ABC$
$AB \rightarrow CD$,	$AC \rightarrow BD$,	$AD \rightarrow BC$,	$BC \rightarrow AD$,
$BD \rightarrow AC$,	$CD \rightarrow AB$,		

- Si $|L| = k$, entonces existen $2^k - 2$ reglas de asociación candidatas (ignorando $L \rightarrow \varnothing$ y $\varnothing \rightarrow L$).

2. Descubrimiento de las reglas con suficiente confianza

Propiedad:

$$\text{confianza}(AB \rightarrow C) \geq \text{confianza}(A \rightarrow BC)$$

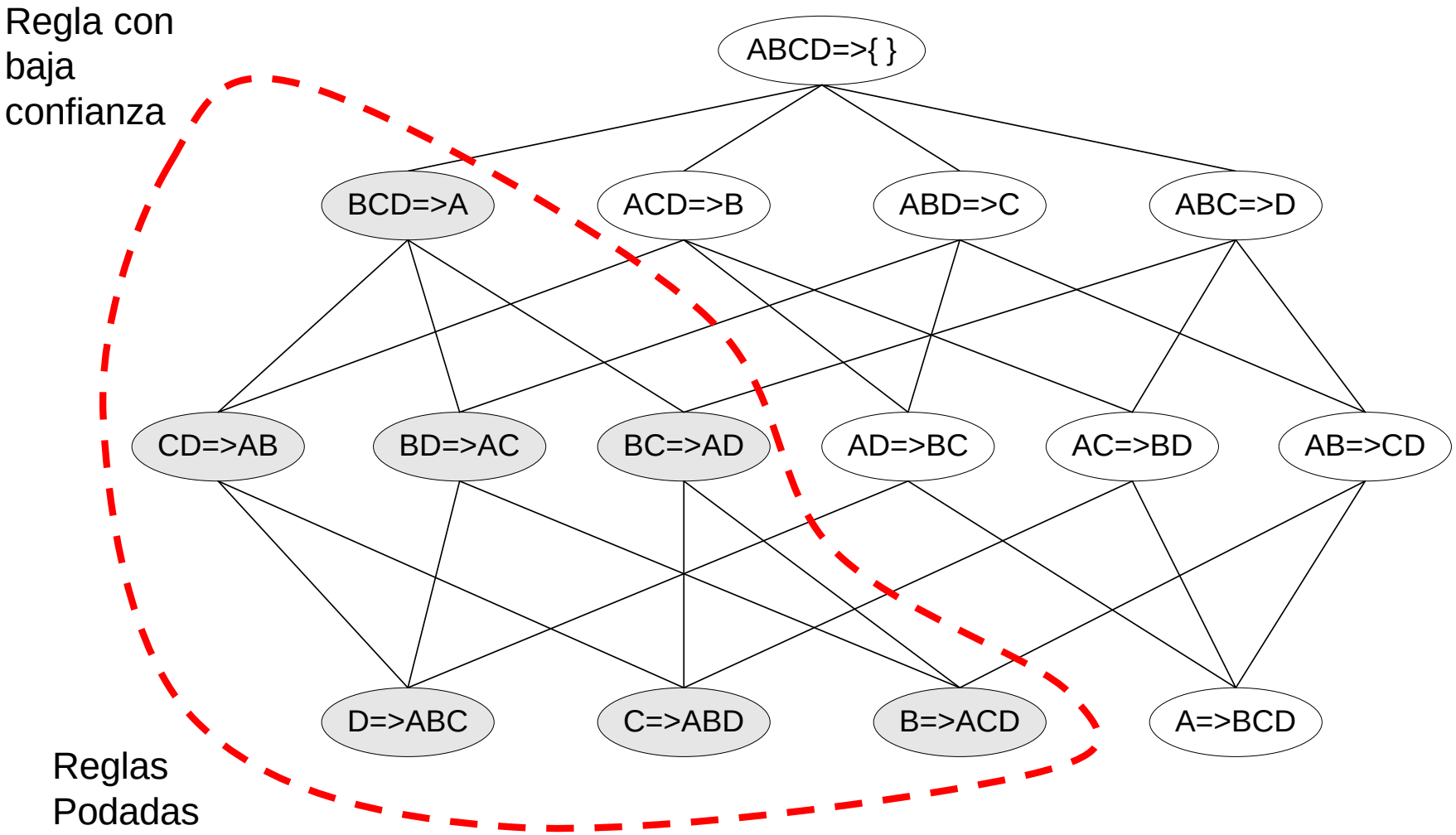
Ver que:

$$\text{confianza}(AB \rightarrow C) = \sigma(ABC) / \sigma(AB) = \text{soporte}(ABC) / \text{soporte}(AB)$$

$$\text{confianza}(A \rightarrow BC) = \sigma(ABC) / \sigma(A) = \text{soporte}(ABC) / \text{soporte}(A)$$

$$\text{soporte}(A) \geq \text{soporte}(AB)$$

2. Descubrimiento de las reglas con suficiente confianza



Medidas de interés

Los algoritmos de reglas de asociación tienden a producir muchas reglas.

Se pueden utilizar medidas de interés para podar u ordenar las reglas. Las medidas más comunes son:

$$support(L \rightarrow R) = count(L \cup R) / |T|$$

$$confidence(L \rightarrow R) = count(L \cup R) / count(L) = support(L \cup R) / support(L)$$

$$lift(L \rightarrow R) = support(L \cup R) / (support(L) * support(R)) = lift(R \rightarrow L)$$

$$leverage(L \rightarrow R) = support(L \cup R) - support(L) * support(R)$$

Medidas de interés

¿Supongan que están analizando 2000 transacciones y que obtiene los conteos que muestran la tabla de abajo, cuánto valen support, confidence, lift y leverage?

$\text{count}(L)$	$\text{count}(R)$	$\text{count}(L \cup R)$
220	250	190

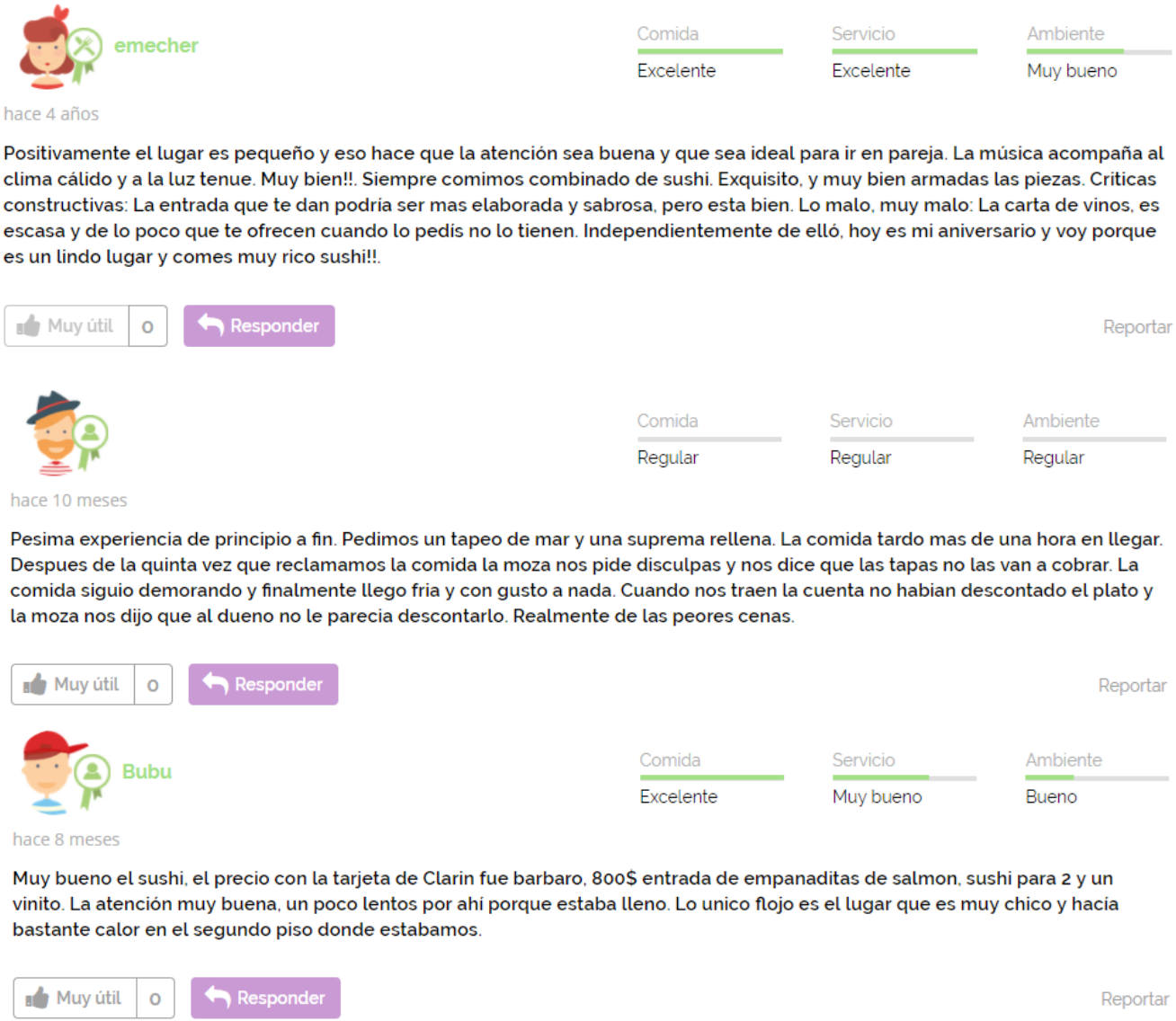
¿Qué implica un lift menor a 1?

Arules en R

Probémoslo en R!

Análisis de sentimiento

Caso de estudio: Guía Óleo



emecher

Comida: Excelente | Servicio: Excelente | Ambiente: Muy bueno

hace 4 años

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

Comida: Regular | Servicio: Regular | Ambiente: Regular

hace 10 meses

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

Comida: Excelente | Servicio: Muy bueno | Ambiente: Bueno

hace 8 meses

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.

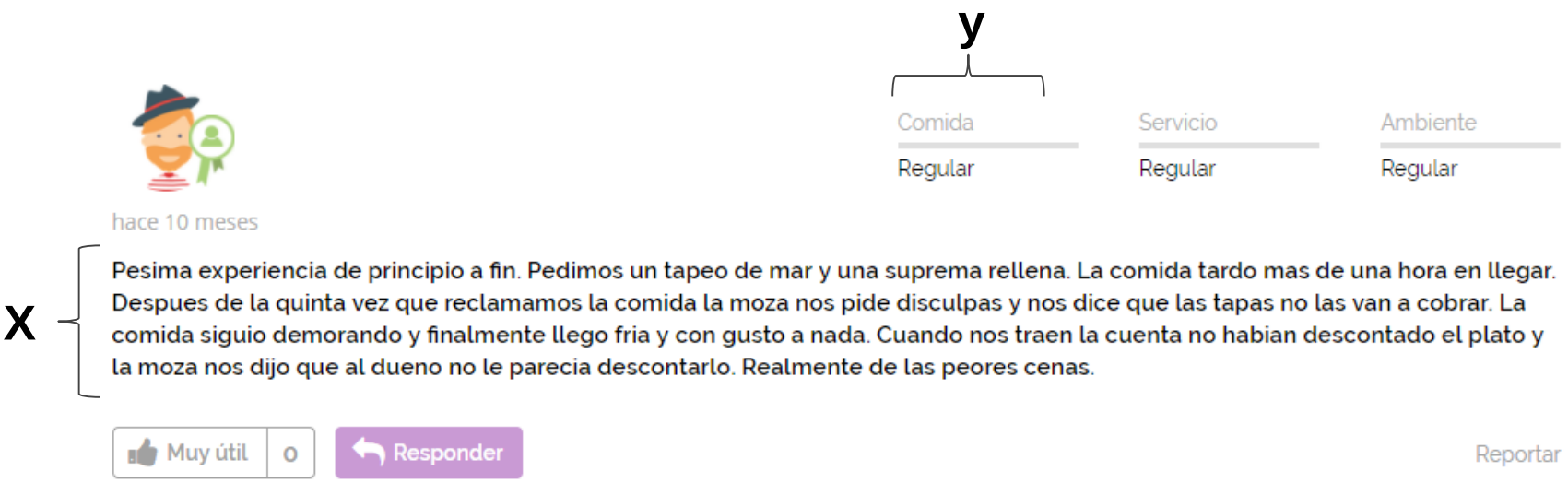
Muy útil 0 | Responder | Reportar

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

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



Análisis de sentimiento

¿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*.

Análisis de sentimiento

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?

Análisis de sentimiento

Bag-of-words model:

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	5	117	0	0

Figure 15.2 The term-document matrix for four words in four Shakespeare plays. The red boxes show that each document is represented as a column vector of length four.

Noten que esta matriz tendrá tantas filas como tokens distintos tenga el corpus (vocabulario) y columnas como documentos.

A la traspuesta de esta matriz se la llama document-term matrix (dtm). La misma podría ser usada para entrenar modelos.

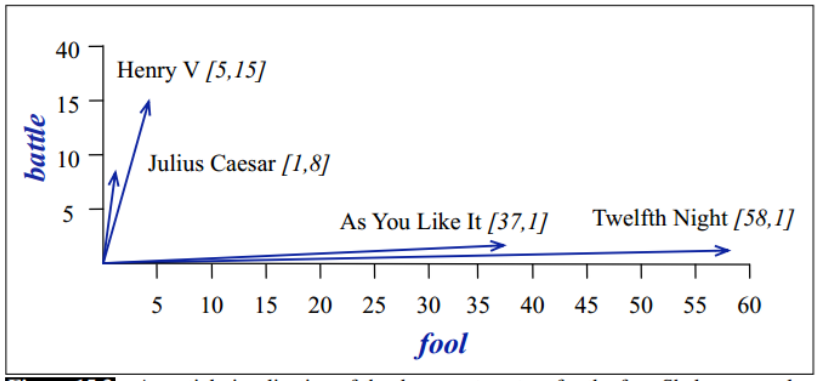


Figure 15.3 A spatial visualization of the document vectors for the four Shakespeare play documents, showing just two of the dimensions, corresponding to the words *battle* and *fool*. The comedies have high values for the *fool* dimension and low values for the *battle* dimension.

Análisis de sentimiento

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?)

Análisis de sentimiento

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 inflecciones de palabras a su forma común:
  - Stemming: son heurísticas que quitan el final de la palabra para lograr este objetivo (having $\rightarrow$ hav)
  - Lematización: busca hacer esto de manera apropiada usando un vocabulario y análisis morfológico (having $\rightarrow$ have).

Clustering de documentos

Teniendo el corpus representado en una dtm. Podemos usar la similitud coseno para medir distancia entre documentos.

$$\vec{a} \cdot \vec{b} = |\vec{a}||\vec{b}| \cos\theta$$

$$\frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|} = \cos\theta$$

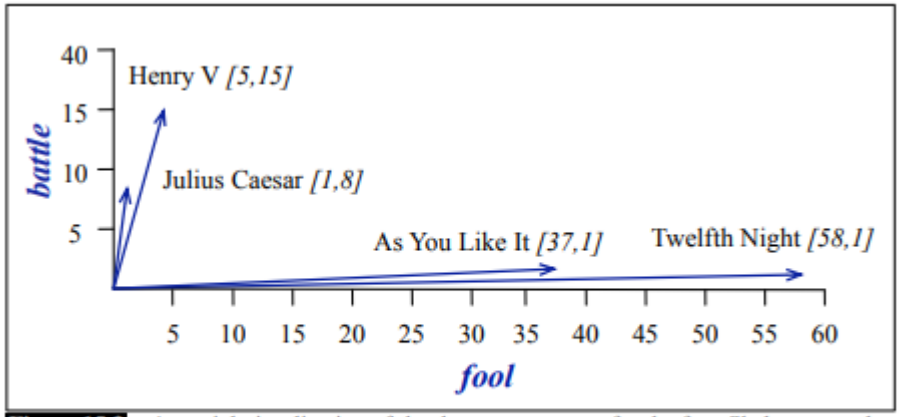
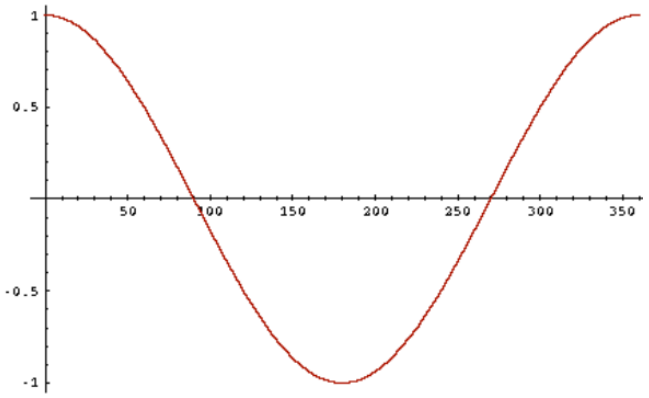


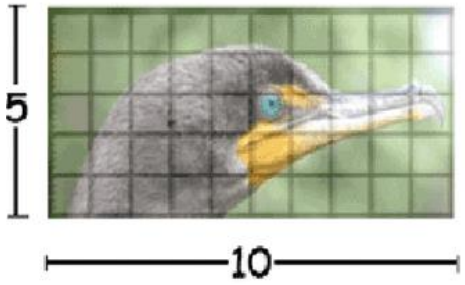
Figure 15.3 A spatial visualization of the document vectors for the four Shakespeare play documents, showing just two of the dimensions, corresponding to the words *battle* and *fool*. The comedies have high values for the *fool* dimension and low values for the *battle* dimension.



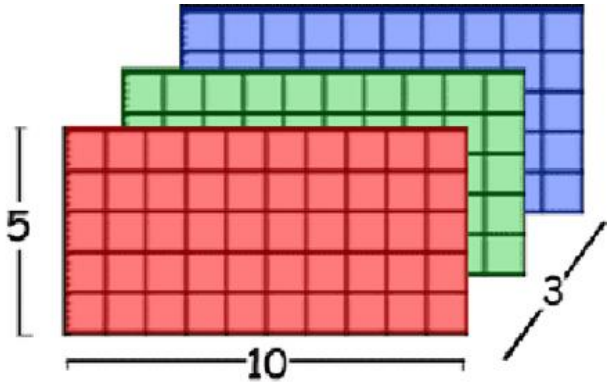
Clustering de documentos en R

Probémoslo en R!

Datos no estructurados

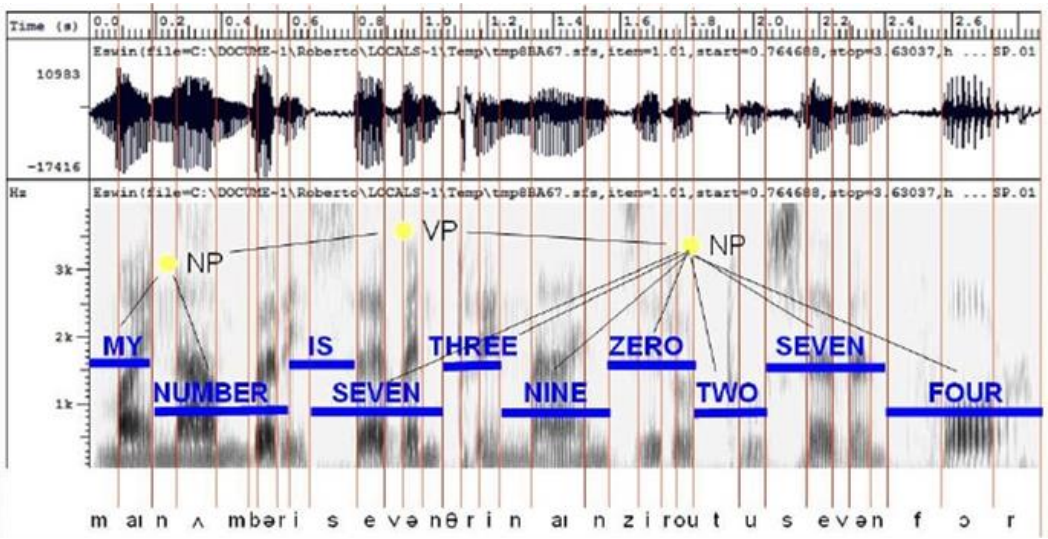


Original Color Image



RGB Matrix

¿Qué hay en la foto?



Lecturas recomendadas para los temas vistos hoy

Reglas de asociación:

- Bramer (Cap 17)

Análisis de sentimiento (no se va a evaluar)

- Martin, J.H. and Jurafsky, D., 2009. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall.

Práctica de laboratorio

Los datos de la competencia Instacart Market Basket Analysis (https://www.kaggle.com/c/instacart-market-basket-analysis) contienen datos de transacciones similares a las que analizamos.

En base a los mismos, replique el análisis que en este post (https://www.kaggle.com/msp48731/frequent-itemsets-and-association-rules/data) se hace.

Proponga alguna manera de mejorar dicho análisis.

Práctica teórica

Suponga que tiene las siguientes cuatro transacciones con los siguientes itemsets:

T1: {K, A, D, B}
T2: {D, A C, E, B}
T3: {C, A, B, E}
T4: {B, A, D}

Encuentre todas las reglas que tienen un soporte (*support*) de al menos 60% y una confianza (*confidence*) de al menos 80%. Para cada una de las reglas resultantes, calcule su *lift*.