

Departamento de LSI, (IXA Research Group) EHU
Topic Modeling

AitZiber AtutXa
jipatsaa@si.ehu.es

Contents

1 Introducción

2 Topic Modeling como Agrupamiento No Supervisado

2.1 Métodos estrictos

2.2 Métodos suaves

Definición

Aplicación de **algoritmos de agrupamiento/clustering no supervisados** con objeto de identificar automáticamente temas subyacentes en textos.

Se aplica cuando se dispone de una gran cantidad de textos y se pretende obtener una **agrupación** de estos **coherente** con respecto a los **argumentos** o hilos conductores que comparten entre ellos.

Referencia básica: Blei D. M. (2012) *Probabilistic Topic Models*. Communications of the ACM

Topic Modeling: Aspectos relevantes

① La salida consiste en:

- ① El agrupamiento de los documentos en K tópicos.
- ② La lista de los n términos mas relevantes asociados a cada tema.

② La salida **no** consiste en:

- ① Identificar automáticamente el número de tópicos. Por el contrario, es uno de los parámetros de entrada.
- ② No identifica automáticamente la descripción de los tópicos.

Topic Modeling

Definición

Aplicación de **algoritmos de agrupamiento/clustering no supervisados** con objeto de identificar automáticamente temas subyacentes en textos.

¿Qué es el clustering no supervisado?

Aprendizaje no supervisado

Reconocimiento de patrones sin exposición a las etiquetas reales en la fase de entrenamiento.

Clustering/Agrupamiento

Descubrimiento de grupos que contienen elementos similares dentro de los datos.

Aplicaciones en los negocios



Segmentación de clientes para campañas de marketing



Análisis de relaciones en redes sociales



Sistemas de recomendación

Clicka: [Amnistía Internacional](#)
Clicka: [Harvard Business School](#)

Tipos de Clustering no Supervisado

- ① Métodos estrictos (hard clustering): Un ítem solo puede pertenecer a un cluster.
 - K-means
 - Nearest Neighbors
 - clustering jerárquico
- ② Métodos suaves (soft clustering): Un ítem puede pertenecer a más de un cluster.
 - LDA: Latent Dirichlet Allocation

K-means

Se suele emplear en **segmentación de clientes**.

Mi empresa quiere establecer una estrategia de marketing para fidelizar a suscriptores. Primer paso, **identificar los tipos de suscriptores que tengo**.

- Edad
- Número de semanas suscrito



Edad: 42
Suscripción: 7



Edad: 18
Suscripción: 3



Edad: 23
Suscripción: 3



Edad: 49
Suscripción: 1



Edad: 37
Suscripción: 7



Edad: 51
Suscripción: 1



Edad: 40
Suscripción: 6



Edad: 20
Suscripción: 4

K-means

Origen:

							
Edad: 42 Suscripción: 7	Edad: 18 Suscripción: 3	Edad: 23 Suscripción: 3	Edad: 49 Suscripción: 1	Edad: 37 Suscripción: 7	Edad: 51 Suscripción: 1	Edad: 40 Suscripción: 6	Edad: 20 Suscripción: 4

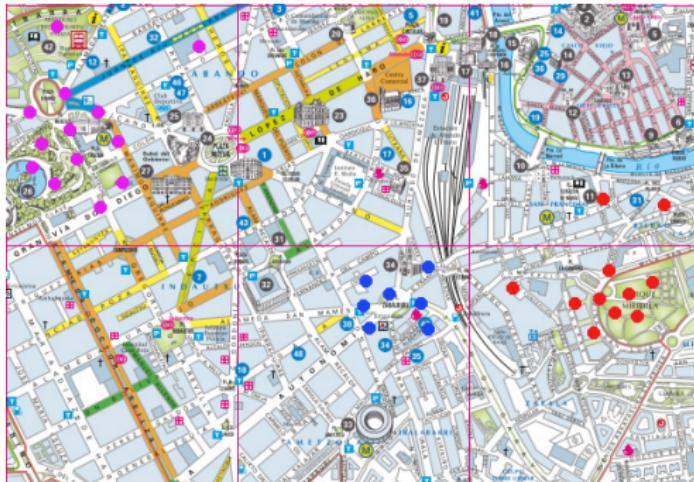
Objetivo:



Más segmentación de clientes...

Mi empresa de sushi quiere abrir 3 nuevas tiendas en Bilbao.

- **identificar las áreas con más clientes** (nota: las empresas solicitan el código postal al realizar una compra).



K-means

Los pasos del K-means:

- 0 Comenzaremos por determinar el **número de clusters K**.
- 1 Seleccionar al azar K elementos. **Centroides**

Repetir hasta convergencia:

- 2 Asignar a los elementos restantes un cluster. **distancia min**
- 3 Recalcular los nuevos centroides. **media**

K-means

0 paso: Determinar K (en este caso 3).

1 paso: Seleccionar al azar los centroides de entre los elementos.

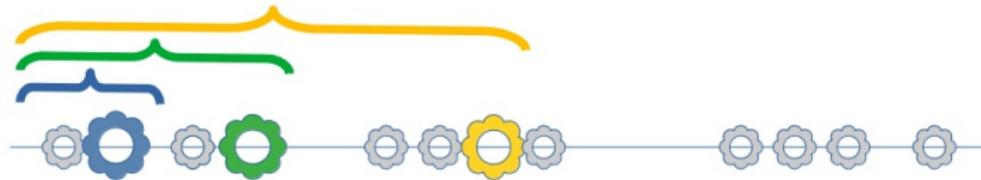


K-means

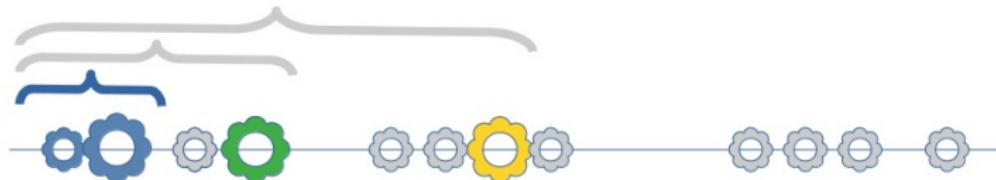
2 paso: Asignar a los elementos restantes un cluster.

Por cada elemento:

distancia euclídea con todos los centroides



cluster = **min distancia**



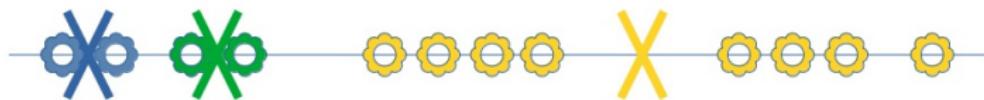
K-means

2 paso: Asignar a los elementos restantes un cluster.
distancia euclídea min



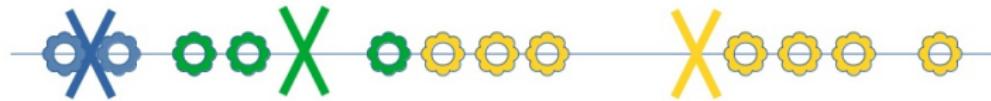
K-means

3 paso: Recalcular los nuevos centroides.
media



K-means

Repetir el proceso hasta convergencia



K-means

Convergencia:



<https://user.ceng.metu.edu.tr/~akifakkus/courses/ceng574/k-means/>

K-means

Convergencia:



K-means

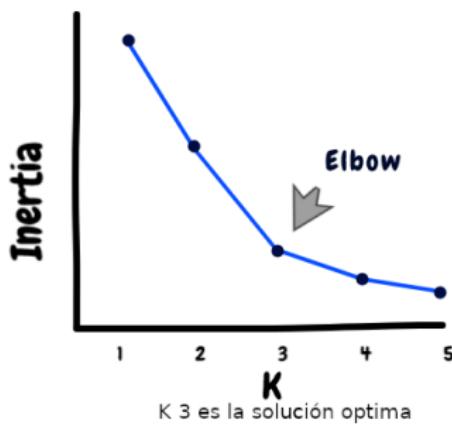
Variaciones (k-means++): Por ejemplo en el modo de seleccionar los primeros centroides.



K-means

Inercia: Intuitivamente, la inercia captura lo lejos que están los puntos dentro de un grupo. Por lo tanto, se busca minimizar la inercia. El rango de valor de inercia comienza desde cero y aumenta. Así, se prueban distintos K y se mide la **inercia**.

$$I = \sum_{i=1}^{\#elem} (\text{centroide} - \text{elem})^2$$



K 3 es la solución optima

K-means

¿Y si estamos en un plano con dos dimensiones? ¿Cómo se calcula la distancia euclídea?

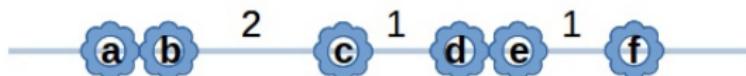


De Documentos a Vectores

	gol	votos	partido	algoritmo	evaluación	elecciones	resultado
doc1	8.0	0.0	9.0	0.0	0.0	0.0	6.0
doc2	0.0	1.0	0.0	8.0	6.0	0.0	5.0
doc3	0.0	10.0	8.0	0.0	0.0	18.0	7.0
doc4	0.0	9.0	8.0	0.0	1.0	8.0	8.0
doc5	8.0	2.0	15.0	0.0	0.0	0.0	0.0

Nearest Neighbors

Necesita un umbral o threshold.



umbral (threshold) = 1

Matriz de distancias

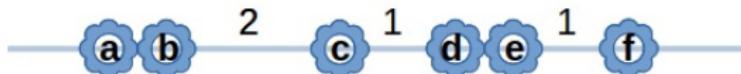
	a	b	c	d	e	f
a	0	1	3	5	6	8
b	1	0	2	4	5	7
c	3	2	0	1	2	3
d	5	4	1	0	1	2
e	6	5	2	1	0	1
f	8	7	3	2	1	0

Cluster1 = {a,b}

Cluster2 = {c,d,e,f}

Nearest Neighbors

Necesita un umbral o threshold.



umbral (threshold) = 2

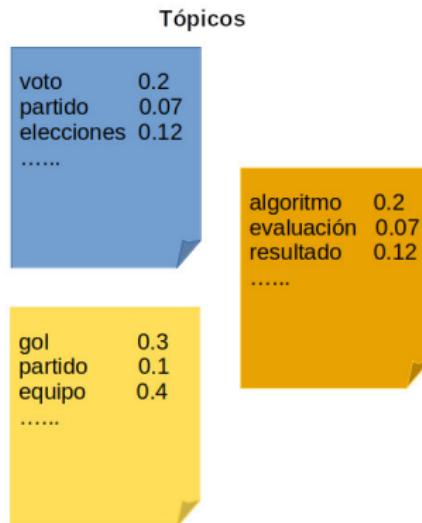
Matriz de distancias

	a	b	c	d	e	f
a	0	1	3	5	6	8
b	1	0	2	4	5	7
c	3	2	0	1	2	3
d	5	4	1	0	1	2
e	6	5	2	1	0	1
f	8	7	3	2	1	0

Cluster1 = {a,b,c,d,e,f}

LDA: Latent Dirichlet Allocation

Un **Tópico** es considerado como una **distribución de probabilidad** de las **palabras** del diccionario seleccionado (eliminando stop-words, palabras genéricas, etc).



LDA: Latent Dirichlet Allocation

LDA es un **modelo generativo** donde el proceso de identificar los tópicos se modela como:

El proceso que **maximiza la probabilidad de la generación de documentos** a partir de palabras tal que cada palabra se genere bajo un determinado tópico teniendo en cuenta la distribución de tópicos del documento y la distribución de palabras que muestra ese tópico.

Formalmente:

Probabilidad de un documento

$$P(\mathbf{W}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \alpha, \beta) = \prod_{j=1}^M P(\theta_j; \alpha) \prod_{i=1}^K P(\varphi_i; \beta) \prod_{t=1}^N P(Z_{j,t} | \theta_j) P(W_{j,t} | \varphi_{Z_{j,t}})$$

LDA: Latent Dirichlet Allocation

¡Claves!

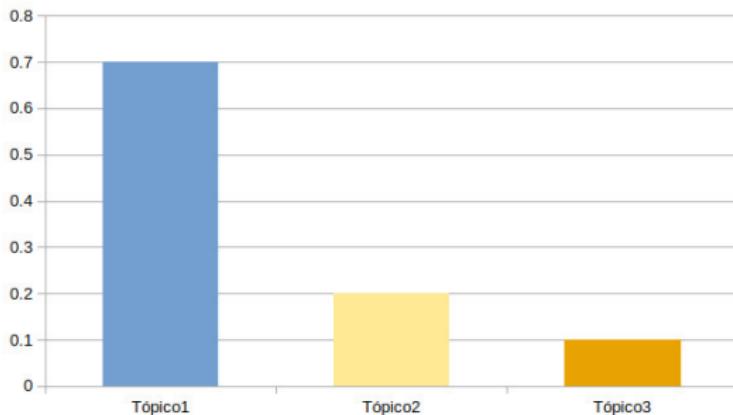
- **modelo generativo** (Naive Bayes) $P(\mathcal{H}|\mathcal{E})^5$, pero no supervisado. No se puede calcular recogiendo frecuencias en base al tópico....
- **maximiza la prob. de la generación de documentos** teniendo en cuenta la **distribución de tópicos (θ) del documento (Z)** y la **distribución (φ) de palabras (W) que muestra ese tópico.**

Formalmente

$$P(Z, \theta, W, \varphi; \alpha, \beta) = \prod_{j=1}^M P(\theta_j; \alpha) \prod_{t=1}^N P(Z_{jt} | \theta_j) \prod_{i=1}^K P(\varphi_i; \beta) \prod_{t=1}^N P(W_{it} | \varphi_{Z_{jt}})$$

LDA: Latent Dirichlet Allocation

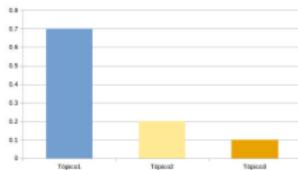
Un Documento es considerado como una **distribución de probabilidad** de los **tópicos** que lo conforman.⁶



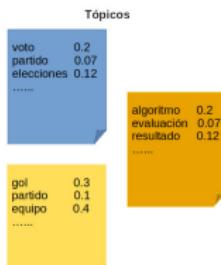
LDA: Latent Dirichlet Allocation

Recordad que LDA es un **modelo generativo** y para crear un documento sintéticamente modela generar sus palabras así:

- ① seleccionar el tópico en base a la probabilidad de este dentro del documento



- ② seleccionar la palabra en base a probabilidad de esta dentro del tópico seleccionado en el paso 1



LDA: Latent Dirichlet Allocation

La distribución de **probabilidad de los tópicos** en un **documento** sería una Distribución Multinomial.

Fórmula que define una distribución multinomial es:

$$\frac{\#pal!}{\#palTopic1! \dots \#palTopicK!} Prob_{topic1}^{\#palTopic1} \dots Prob_{topicK}^{\#palTopicK}$$

donde $\#pal$ es el número de palabras en el doc y $\#palTopicK$ sería el número de palabras asociadas al tema k . $Prob_{topicK}$ sería la probabilidad del tema K para el documento d.

LDA: Latent Dirichlet Allocation

La distribución de **probabilidad de los tópicos** en un **documento** sería una Distribución Multinomial.

Fórmula que define una distribución multinomial es:

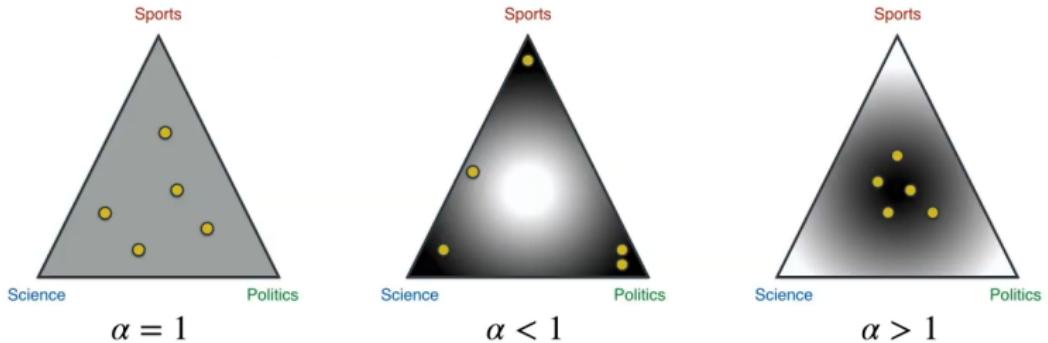
$$\frac{\#pal!}{\#palTopic1! \dots \#palTopicK!} Prob_{topic1}^{\#palTopic1} \dots Prob_{topicK}^{\#palTopicK} \text{ donde}$$

$\#pal$ es el número de palabras en el doc y $\#palTopicK$ sería el número de palabras asociadas la clase k . $Prob_{topicK}$ sería la probabilidad de la clase K para el documento d.

LDA: Latent Dirichlet Allocation

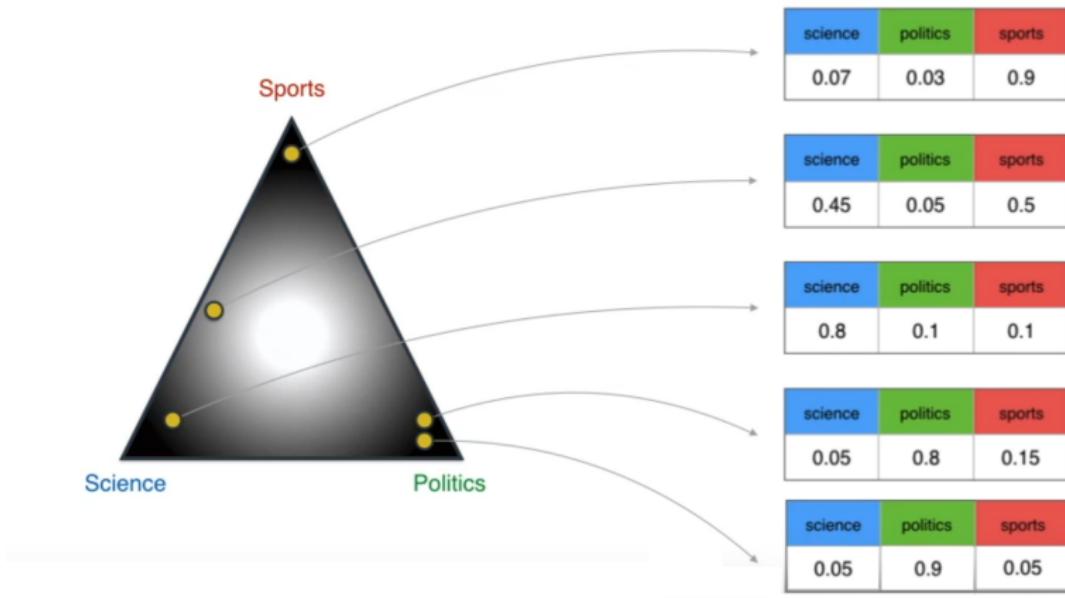
- Aquí es donde entra en juego la distribución de Dirichlet que una forma intuitiva de interpretarla es como *una generadora de probabilidades* en base a **K** (en nuestro caso el **número de tópicos**) y un parámetro α (**parámetro de concentración**).
- Así mayor α , más concentrados estarán las probabilidades, menor α menor concentración, es decir, valores más esparsos.

LDA: Latent Dirichlet Allocation



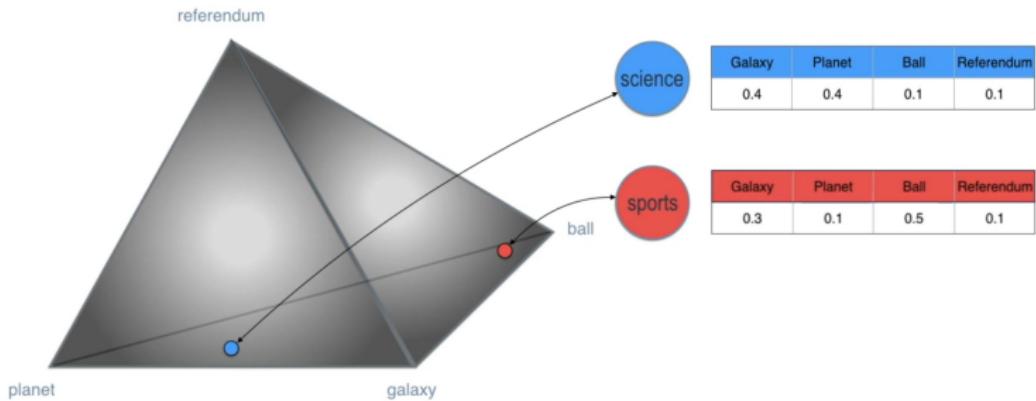
LDA: Latent Dirichlet Allocation

La distribución de Disrichlet nos permite obtener para documentos probabilidades de tópicos.



LDA: Latent Dirichlet Allocation

La distribución de Disrichlet se puede aplicar también para obtener para un tópico las probabilidades de las palabras en dicho tópico.



LDA: Latent Dirichlet Allocation

La distribución de **probabilidad de las palabras** en un **tópico** sería también una Distribución Multinomial.

Fórmula que define una distribución multinomial es:

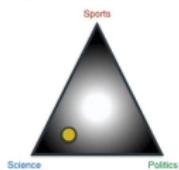
$$\frac{\#palTopic1!}{\#palTopic1! \dots \#palTopicK!} Prob_{pal1}^{\#palTopic1} \dots Prob_{paln}^{\#palTopic1}$$

donde $\#palTopic1$ es el número de palabras del topic1 y $\#pal1Topic$ sería el número de veces que la pal1 está asociada al Topic1. $Prob_{pal1}$ sería la probabilidad de la pal1 asociada al tópico1.

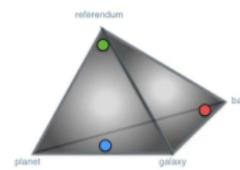
LDA: Latent Dirichlet Allocation

Sumando todas las partes, el modelo de Latent Dirichlet Allocation se definiría de la siguiente forma:

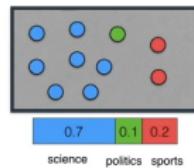
$$\prod_{j=1}^M P(\theta_j; \alpha)$$



$$\prod_{i=1}^K P(\varphi_i; \beta)$$



$$\prod_{t=1}^N P(Z_{j,t} | \theta_j)$$



$$P(W_{j,t} | \varphi_{Z_{j,t}})$$

galaxy	galaxy	planet
galaxy	planet	planet
galaxy	planet	referendum
planet	ball	ball
planet	referendum	referendum
referendum	referendum	referendum
galaxy	referendum	referendum
referendum	referendum	referendum
galaxy	ball	ball
ball	ball	galaxy
planet	referendum	

science	politics	sports
0.7	0.1	0.2

Galaxy	Planet	Ball	Referendum
0.4	0.4	0.1	0.1
Galaxy	Planet	Ball	Referendum
0.1	0.1	0.1	0.7

Galaxy	Planet	Ball	Referendum
0.3	0.1	0.5	0.1

Topics
science
science
sports
science
science
politics
sports
sports
science

LDA: Latent Dirichlet Allocation

El modelo es computacionalmente intratable ... demasiado costoso...



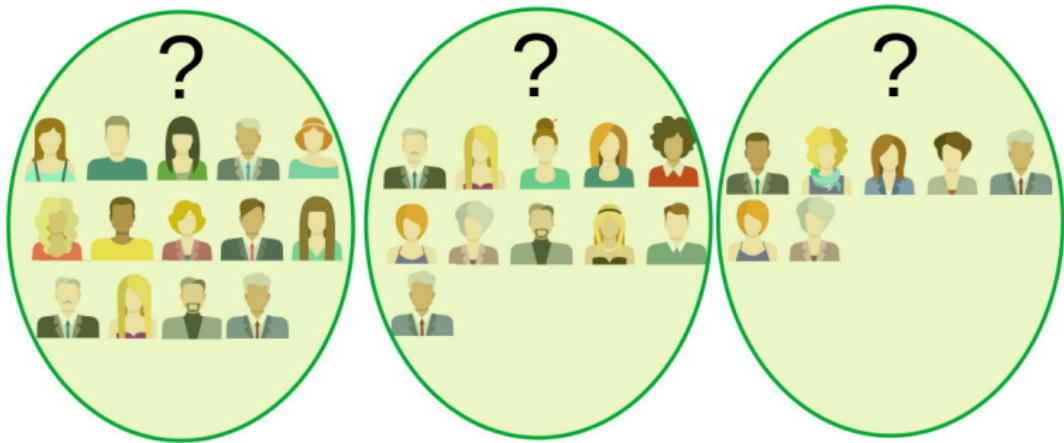
LDA: Gibbs Sampling

Una aproximación al **LDA** (Gibbs Sampling) para realizar el **Topic Modeling** se basa en la idea de que si calcular $p(x,y)$ es complicado, una alternativa es calcular $p(x|y)$ y luego $p(y|x)$. Es decir, calcular probabilidades condicionadas fijando el resto de las variables.

LDA: Gibbs Sampling

Comencemos por un ejemplo:

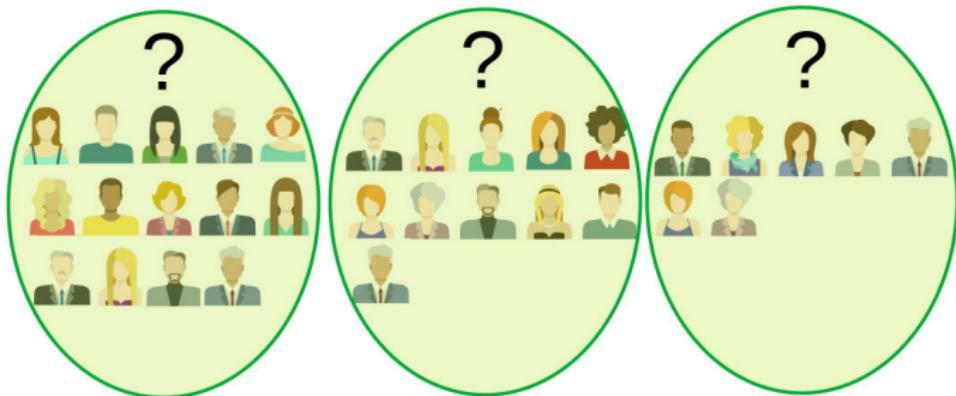
- Jornadas informativas para alumnos organizadas por la universidad.



Cada aula está abierta a cualquier alumno (de 1º, 2º, 3º), pero principalmente dirigidas a temas de un curso, es decir, cada aula hablará de la carrera (1,2,3) y ahondará en un curso.

LDA: Gibbs Sampling

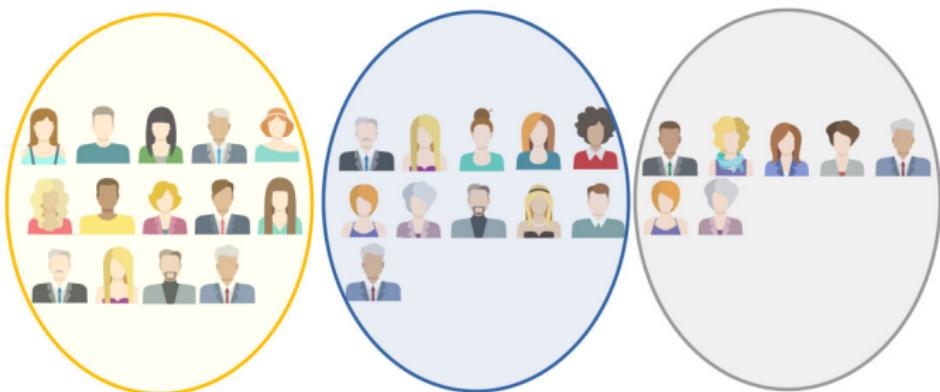
- Pero desconozco la distribución. Se han sacado fotos del aula que tengo que colgar en la web, pero desconozco que aula es de 1º, 2º o 3º



LDA: Gibbs Sampling

Pero puedo hacer ciertas asunciones:

- En **cada jornada** → alumnos de **un determinado curso** principalmente.



LDA: Gibbs Sampling

Pero puedo hacer ciertas asunciones:

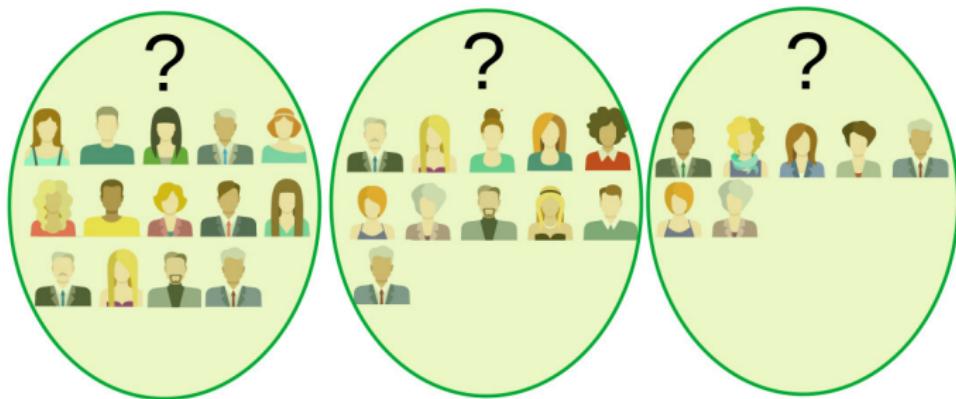
- **Cada alumno** → matriculado en **un curso** principalmente.
- **Cada alumno** → sentado **junto a sus similares**.



Primero	Segundo	Tercero
10 %	10%	80%
100 %	0%	0%
20 %	80%	0%

LDA: Gibbs Sampling

- Pero desconozco la distribución.

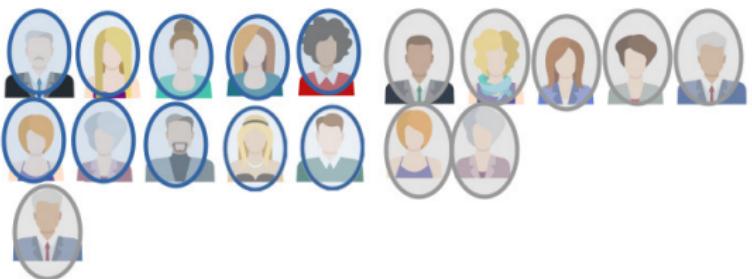


LDA: Gibbs Sampling

Primero? Segundo? Tercero?



0.71
0.21
0.07

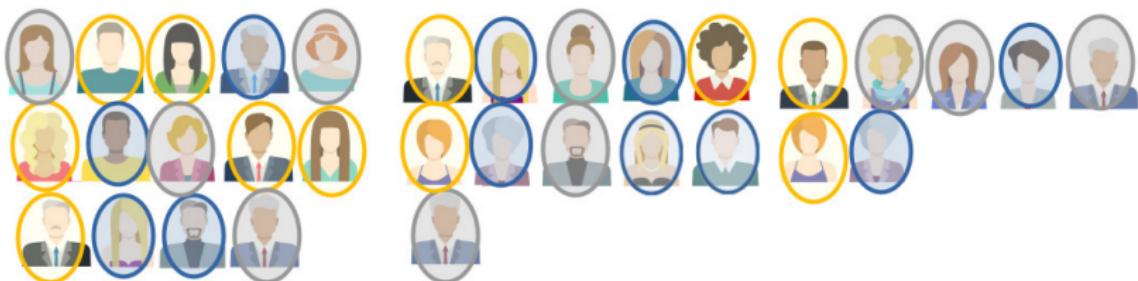


0.90
0.10
0.00

1.00
0.00
0.00

LDA: Gibbs Sampling

Primero? Segundo? Tercero?



inicializo aleatoriamente, asignandole a cada persona un grupo aleatorio.

LDA: Gibbs Sampling



- Comienzo por recalcular a la primera persona de la primera aula.
- Asumo que las demás personas tienen una asignación correcta.

LDA: Gibbs Sampling

¿De qué color creéis que hay que ponerle a la primera chica?

Primero? Segundo? Tercero?



LDA: Gibbs Sampling

¿De qué color creéis que hay que ponerle a la primera chica?

Primero? Segundo? Tercero?



- Basandonos solo en el aula 1 (documento): $\frac{6}{13}$, $\frac{4}{13}$ y $\frac{3}{13}$
- Basandonos solo en la persona (palabra): $\frac{0}{0}$, $\frac{0}{0}$ y $\frac{0}{0}$

$$\left(\frac{6}{13} \times \frac{0}{0}, \frac{4}{13} \times \frac{0}{0}, \frac{3}{13} \times \frac{0}{0} \right)$$

LDA: Gibbs Sampling

¿De qué color creéis que hay que ponerle a la primera chica?
Primero? Segundo? Tercero?



La asignación será el máximo entre ($\frac{6}{13} \times \frac{0}{0}$, $\frac{4}{13} \times \frac{0}{0}$, $\frac{3}{13} \times \frac{0}{0}$)

- Pero los ceros son demasiado **drásticos** especialmente si el random inicial fue muy equivocada.
- ¿Cómo evito el problema?

LDA: Gibbs Sampling

¿De qué color creéis que hay que ponerle a la primera chica?

Primero? Segundo? Tercero?



- Basandonos solo en el aula 1: $\frac{6}{13} + \alpha$, $\frac{4}{13} + \alpha$ y $\frac{3}{13} + \alpha$
- Basandonos solo en la persona: $\frac{0}{0} + \beta$, $\frac{0}{0} + \beta$ y $\frac{0}{0} + \beta$

Estos valores generan una distribución de probabilidades

$$\left(\frac{6}{13} + \alpha \times \frac{0}{0} + \beta, \frac{4}{13} + \alpha \times \frac{0}{0} + \beta, \frac{3}{13} + \alpha \times \frac{0}{0} + \beta \right)$$

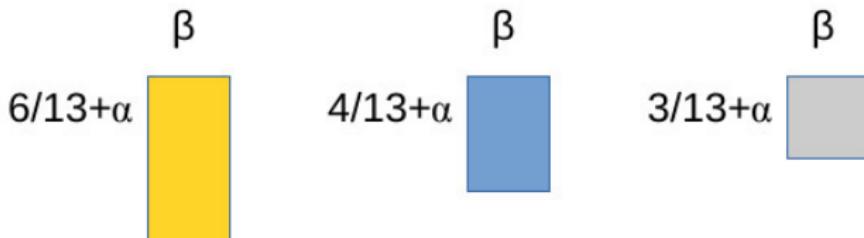
LDA: Gibbs Sampling

¿De qué color creéis que hay que ponerle a la primera chica?

Primero? Segundo? Tercero?

Estos valores generan una distribución de probabilidades

$$\left(\frac{6}{13} + \alpha \times \frac{0}{0} + \beta, \frac{4}{13} + \alpha \times \frac{0}{0} \beta, \frac{3}{13} + \alpha \times \frac{0}{0} + \beta \right)$$



Lanzo un random y veo dónde cae. Como lanzar un dardo sobre estas areas.

LDA: Gibbs Sampling



- Basandonos solo en el aula 1: $\frac{8}{13} + \alpha$, $\frac{2}{13} + \alpha$ y $\frac{3}{13} + \alpha$
- Basandonos solo en la persona: $\frac{0}{1} + \beta$, $\frac{0}{1} + \beta$ y $\frac{0}{1} + \beta$

Estos valores generan una distribución de probabilidades

$$\left(\frac{8}{13} + \alpha \times \frac{0}{1} + \beta, \frac{2}{13} + \alpha \times \frac{1}{1} + \beta, \frac{3}{13} + \alpha \times \frac{0}{1} + \beta \right)$$

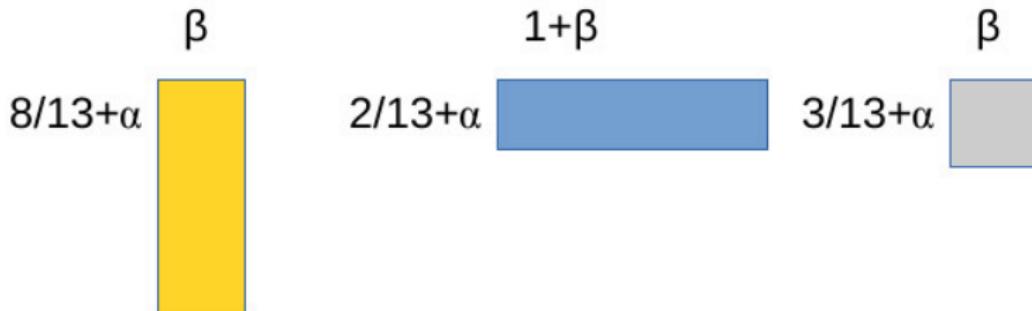
Lanzo un random y veo dónde cae. Como lanzar un dardo sobre estas areas.

LDA: Gibbs Sampling

Primero? Segundo? Tercero?

Estos valores generan una distribución de probabilidades

$$\left(\frac{8}{13} + \alpha \times \frac{0}{1} + \beta, \frac{2}{13} + \alpha \times \frac{1}{1} + \beta, \frac{3}{13} + \alpha \times \frac{0}{1} + \beta \right)$$



Lanzo un random y veo dónde cae. Como lanzar un dardo sobre estas areas.

LDA: Gibbs Sampling

Ejercicio:



Dar un par de vueltas al algoritmo, suponiendo que el dardo caiga donde mayor es la probabilidad.

LDA: Gibbs Sampling

Cuestiones prácticas:

- Para la implementación del LDA podéis emplear la librería Gensim:
<https://radimrehurek.com/gensim/models/ldamodel.html>
- Para seleccionar el codo, consultar las medidas de coherencia que implementa gensim <https://radimrehurek.com/gensim/models/coherencemodel.html>

LDA: Gibbs Sampling

Cuestiones prácticas:

- uci:

$$C_{UCI} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N = PMI(w_i, w_j)$$

$$PMI = \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}$$

Suponiendo en tópico palabras: *game*, *sports*, *ball* y *team*:

$$C_{UCI} = \frac{1}{6} (PMI(game, sport) + PMI(game, ball) + PMI(game, team) + PMI(sport, ball) + PMI(sport, team) + PMI(ball, team))$$

$\log(0,0002 \div (0,5 \times 0,02))$	=	-1,698970004
$\log(0,0013 \div (0,5 \times 0,02))$	=	-0,886056648

LDA: Gibbs Sampling

Cuestiones prácticas:

- umass:

$$C_{UMASS} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N \log \frac{P(w_i, w_j) + \epsilon}{P(w_i)}$$

Suponiendo en tópico palabras: *game*, *sports*, *ball* y *team*:

$$C_{UMASS} = \frac{1}{6} (PMI(game|sport) + PMI(game|ball) + PMI(game|team) + PMI(sport|ball) + PMI(sport|team) + PMI(ball|team))$$

$\log(0,0002/(0,5))$	=	-3,397940009
$\log(0,0013/(0,5))$	=	-2,585026652

LDA: Gibbs Sampling

Cuestiones prácticas:

- c_v:

$$C_{C_V} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N = NPMI(w_i, w_j)$$

$$NPMI = \frac{\log \frac{P(w_i, w_j) + \epsilon}{P(w_i)P(w_j)}}{-\log(P(w_i, w_j) + \epsilon)}$$

Bibliografía

- Latent Dirichlet Allocation. Journal of Machine Learning Research 3 (2003) 993-1022, Autores: David M. Blei, Andrew Ng, Michael I. Jordan
- Grokking Machine Learning, Luis Serrano
- youtube.com/c/LuisSerrano
- Introduction to the Dirichlet Distribution and Related Processes, UWEE Technical Report Number UWEETR-2010-0006, Autores: Bela A. Frigyik, Amol Kapila, and Maya R. Gupta

Frame Title