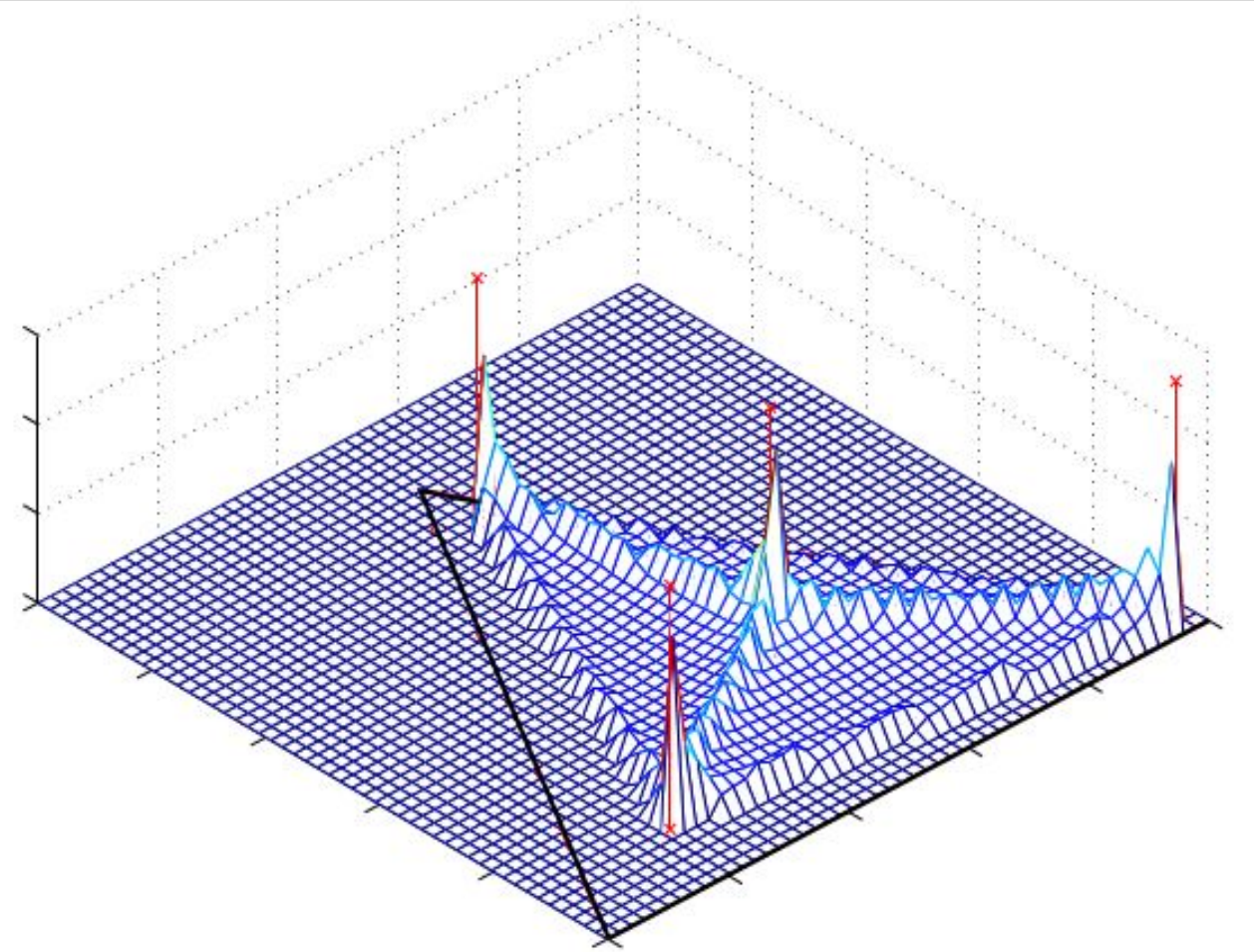


DigitalHouse > Coding School

INTELIGENCIA ARTIFICIAL

UNIDAD 3
MÓDULO NLP

Topic Modeling

DigitalHouse > Coding School

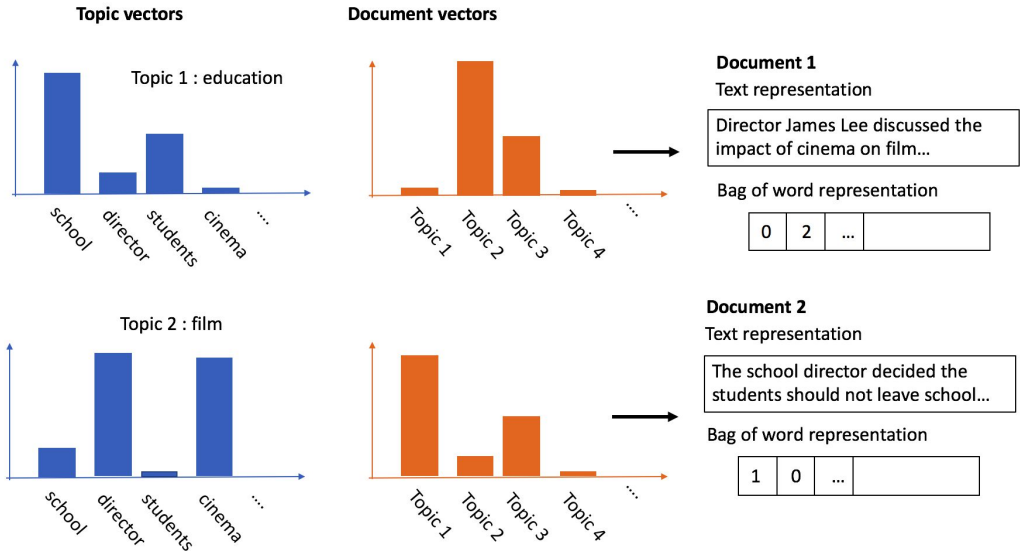
1. Vamos a ver para qué sirve el modelado de tópicos
2. Estudiaremos cómo es el proceso generativo propuesto por Latent Dirichlet Allocation
3. Veremos extensiones del modelo básico de LDA para incorporar la dimensión temporal
4. Analizaremos de qué forma modificar el modelo clásico de LDA para establecer correlación entre tópicos

INTRODUCCIÓN

- El objetivo en Topic Modeling es encontrar los temas de los que habla un documento.

- Topic Modeling se puede utilizar para:
  - Dimensionality Reduction
  - unsupervised learning
  - Tagging
  - Paso previo para clasificación de textos

- El método más difundido para esto es el Latent Dirichlet Allocation (LDA), un método con profundos fundamentos en la *inferencia bayesiana*, que a su vez se puede re-utilizar para otros objetivos que comparten ciertos supuestos



LATENT DIRICHLET ALLOCATION

Comencemos por definir la terminología del problema:

- una ***palabra*** es la unidad básica discreta de los datos, se define como un ítem de un *vocabulario*, indexado como {1…,V}. Las palabras se representan como vectores unitarios con un único componente igual a 1, y los demás iguales a 0. Definimos el superíndice *i* del vector como la i-ésima palabra del vocabulario y el i-ésimo elemento del vector. La V-ésima palabra del vocabulario es el vector w, tal que $w^v=1$ y $w^u=0$. $u \neq v$

- un ***documento*** es una secuencia de N palabras, definidos como $\mathbf{w} = (w_1, w_2, \ldots, w_N)$

- un **corpus** es una colección de M documentos definida como $D = (\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_M})$

- un **tópico** es una dimensión latente del corpus, y suponemos una cantidad fija k de los mismos.

- Dado un corpus, como resultado de aplicar LDA obtendremos 2 elementos:
  - Una distribución de tópicos sobre cada documento.
  - Una distribución de las palabras sobre los tópicos.

- De esta forma, podemos tomar las palabras más importantes de cada tópico para definirlo, a la vez que decimos que cada documento habla de los tópicos que más pesan en su distribución.

- También podemos ver cuáles son los tópicos más importantes de corpus, y en un tema más avanzado (dynamic topic modeling), cómo evolucionan los tópicos en el tiempo

PROCESO GENERATIVO

- LDA es un **modelo generativo probabilístico del corpus**. Esto quiere decir que parte de definir una representación respecto de cómo se construye el corpus.

- La idea general es que cada documento se representa como una mezcla aleatoria sobre tópicos latentes, donde cada tópico se caracteriza por una distribución sobre las palabras del vocabulario. Asumimos que esta distribución de los tópicos se especifica antes que se haya generado cualquiera de los datos del corpus..

- Para cada documento del corpus. imaginamos que se genera las palabras en un proceso de dos etapas:
  - Elegimos aleatoriamente la distribución sobre los tópicos.
  - Para cada palabra en el documento:
    - Elegimos aleatoriamente el tópico de la distribución definida en el paso anterior.
    - Elegimos aleatoriamente una palabra de la distribución correspondiente a dicho tópico.

Podemos definir lo anterior formalmente, como el siguiente proceso:

1. Para cada tópico k en $k = 1, \ldots, K$,

   a. generar una distribución sobre las palabras $\beta \sim Dir_V(\eta)$,

   $\eta \in R_{>0}$ es un parámetro fijo

2. Para cada documento en $d = 1, \ldots, D$,

   a. generar un vector de proporciones de los tópicos $\theta_d \sim Dir_K(\alpha)$

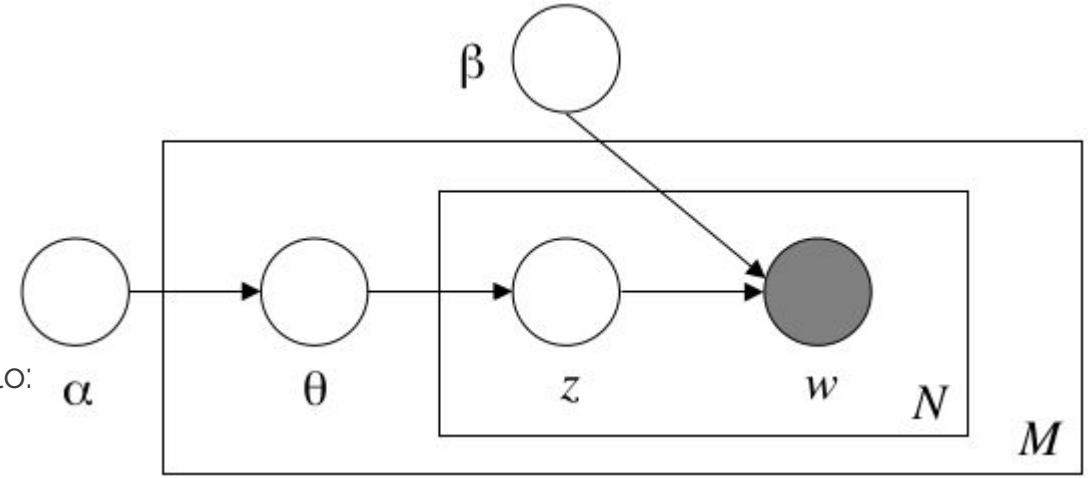
   $\alpha \in R^K_{>0}$ es un parámetro fijo

   b. para cada palabra en $n = 1, \ldots, N$,

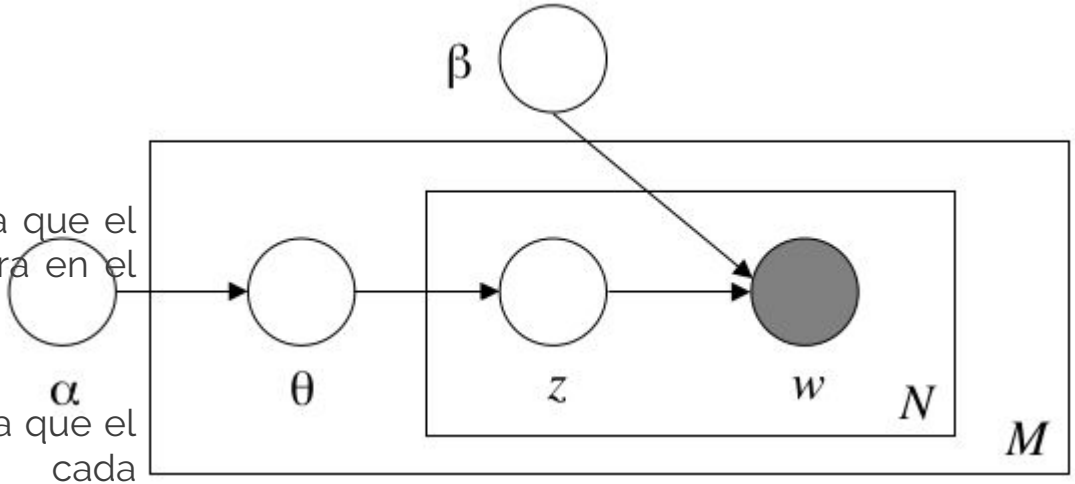
      i. generar una asignación del tópico $z_{dn} \sim Mult(\theta_d)$,
      ii. generar una palabra $w_{dn} \sim Mult(\beta_{zn})$

- Los modelos que definen este tipo de procesos se conocen como *Gaussian Mixture Models*. En criollo podríamos decir que hacen una verdadera mezcolanza de distribuciones.

- Para entender un poco mejor, podemos ayudarnos de una representación gráfica del modelo:



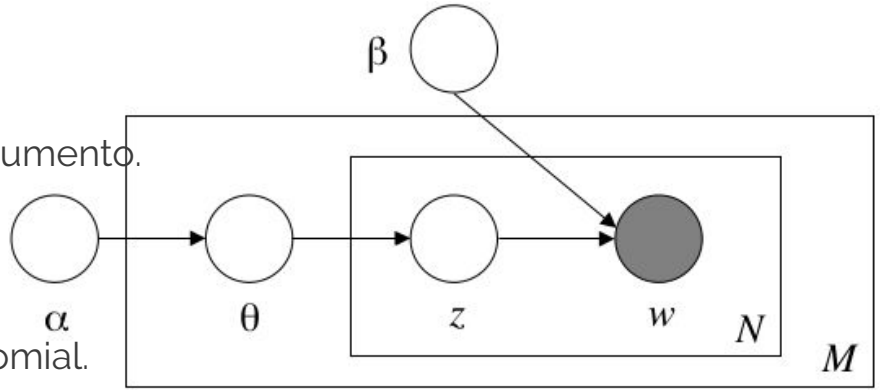
- Cada nodo representa una distribución de probabilidad.
- La arista significa que la distribución de salida define los parámetros de la distribución de entrada



- Los recuadros significan *replicación*:
  - El recuadro interior representa que el proceso se realiza para palabra en el documento.
  - El recuadro exterior representa que el proceso se realiza para cada documento en el corpus

Recordemos que

- $\beta \sim Dir_v(\eta)$: Palabras por tópico
- $\theta_d \sim Dir_K(\alpha)$: Proporciones por tópico de un documento.
- $z_{dn} \sim Mult(\theta_d)$: Asignación del tópico .
- $w_{dn} \sim Mult(\beta_{zn})$: Distribución de palabras.



~ **Mult()** significa que tiene una distribución multinomial.

~ **Dir()** significa que tiene una *distribución de Dirichlet*. Que definiremos a continuación.

DISTRIBUCIÓN DE DIRICHLET

- Un Proceso de Dirichlet es una familia de procesos estocásticos donde **las realizaciones son ellas mismas distribuciones de probabilidad**.

- Es decir, el rango de esta distribución (así como en una normal son los reales) son distribuciones de probabilidad.

- Se define a partir de una distribución de base, H, y un parámetro de escala, α.
  - La distribución H es el valor esperado, y las realizaciones son distribuciones en torno a H.
  - $\alpha \rightarrow 0$ implica que todas las realizaciones dan un mismo valor
  - $\alpha \rightarrow \infty$ implica que las distribuciones resultantes son continuas
  - Los valores en el medio dan distribuciones discretas.

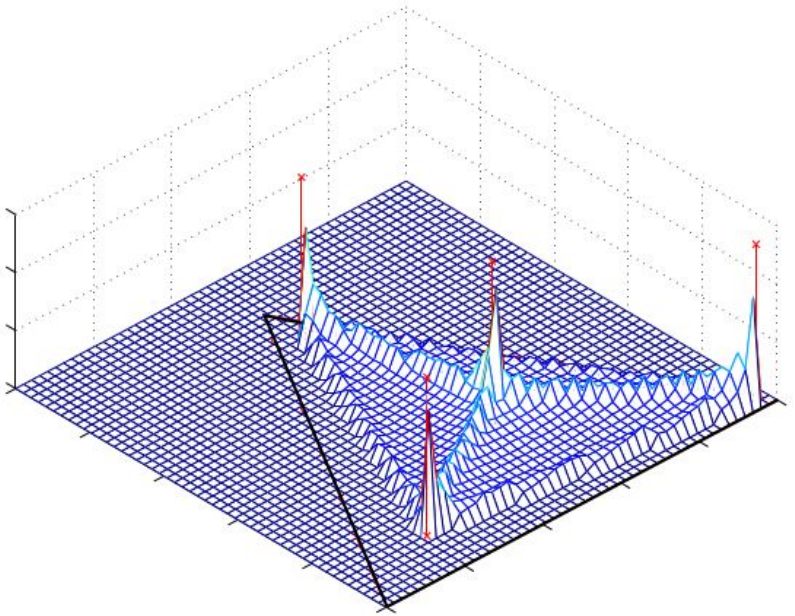
Distribución de Dirichlet

Para interpretarlo geométricamente, podemos pensar un ejemplo de la distribución de densidad para tres palabras y 4 tópicos.

El triángulo del plano x-y se dice que es un *simplex 2D*, y representa todas las distribuciones (multinomiales) posibles sobre las tres palabras

cada uno de los vértices del triángulo es una distribución de probabilidad que asigna una probabilidad de 1 a una de las palabras.

El punto medio de cada lado, es una distribución con probabilidad 0.5 a dos palabras.



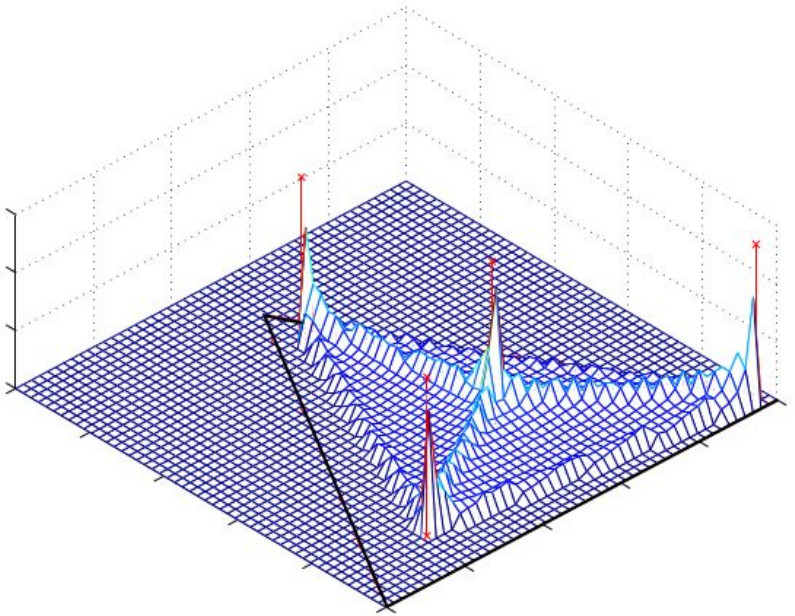
Distribución de Dirichlet

Para interpretarlo geométricamente, podemos pensar un ejemplo de la distribución de densidad para tres palabras y 4 tópicos.

El centroide del triángulo, asigna una probabilidad ⅓ a cada palabra.

Los cuatro puntos marcado con x son las distribuciones multinomiales de *p(w|z)* para cada uno de los cuatro tópicos.

La altura en el eje z es una posible distribución de densidad sobre el *simplex*, es decir, sobre las distribuciones de densidad multinomiales, dada por LDA



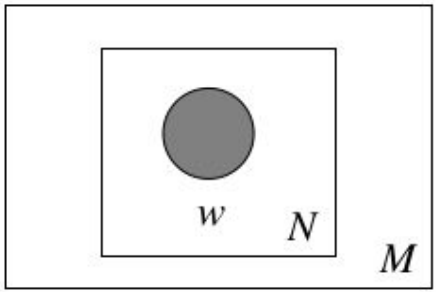
- En el caso de LDA, es importante señalar que no se trata de un simple modelo Dirichlet-multinomial.

- Un modelo de este tipo implicaría un proceso de dos etapas, en el cual el proceso de Dirichlet es sampleado una vez para todo el corpus, y cada realización de la multinomial resultante es seleccionado para cada documento.
  - Esto implicaría que cada documento tiene un único tópico

- El modelo propuesto tiene tres etapas:
  - El nodo del tópico se selecciona repetidas veces en cada documento (una vez por palabra). Es por esto que LDA permite que haya múltiples tópicos por documento

COMPARACIÓN CON OTROS MODELOS DE TOPIC MODELING

El modelo más sencillo que se puede pensar es el unigram:



(a) unigram

- Hay una única distribución de probabilidad de la palabras para todas las palabras de todos los documentos.

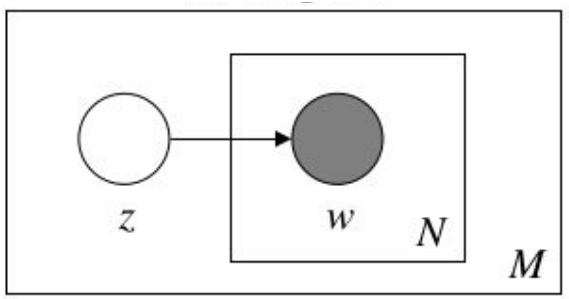
- La probabilidad de un documento, **w**, se describe como la probabilidad conjunta de sus N palabras:

$$p(\mathbf{w}) = \prod_{n=1}^{N} p(w_n)$$

- El siguiente nivel de complejidad es definir los tópicos como una variable aleatoria para cada documento.

- Así, cada documento es generado primero eligiendo un tópico *z*, y luego generando N palabras de forma independiente, mediante una distribución multinomial (definida por el tópico)

- De esta forma, cada documento habla de un único tópico



(b) mixture of unigrams

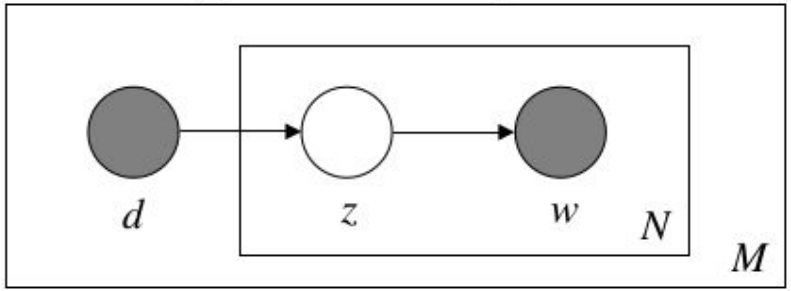
$$p(\mathbf{w}) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n|z)$$

Probabilistic latent semantic indexing fue el estado del arte hasta LDA:

- Cada etiqueta, *d*, y cada palabra, *w*, son generadas de forma condicionalmente independiente, dado un tópico *z*.

- Las etiquetas sirven como un ponderador sobre los tópicos, permitiendo que un documento hable de más de un tópico.

- pLSI no tiene una forma de asignar una probabilidad a un documento no visto en el training, y es propenso al overfitting.

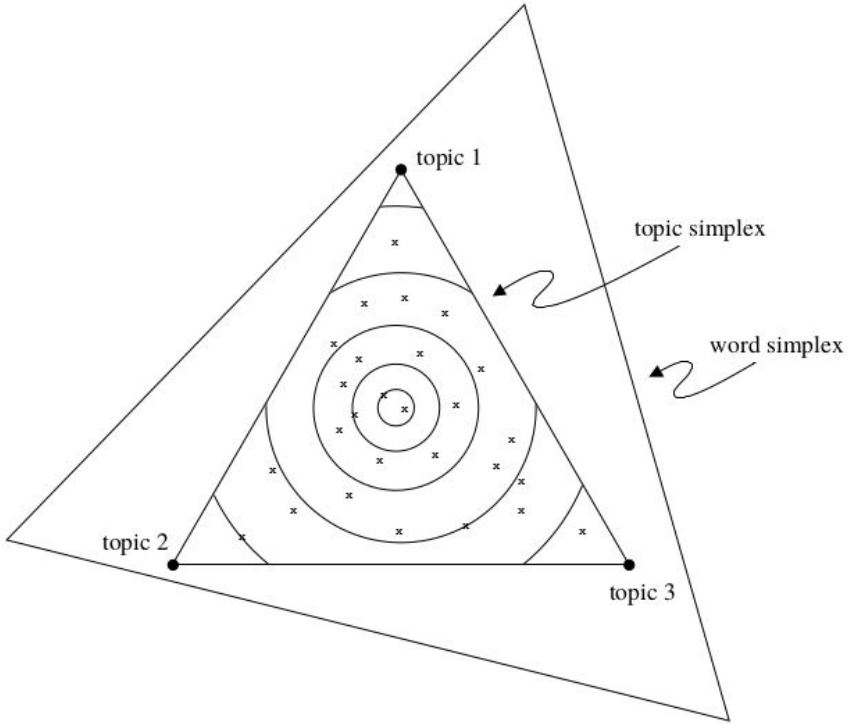


(c) pLSI/aspect model

$$p(d, w_n) = p(d) \sum_z p(w_n|z)p(z|d)$$

Podemos realizar la comparación geométrica de los modelos, analizando el caso de un simplex de 3 tópicos, dentro de un simplex de 3 palabras.

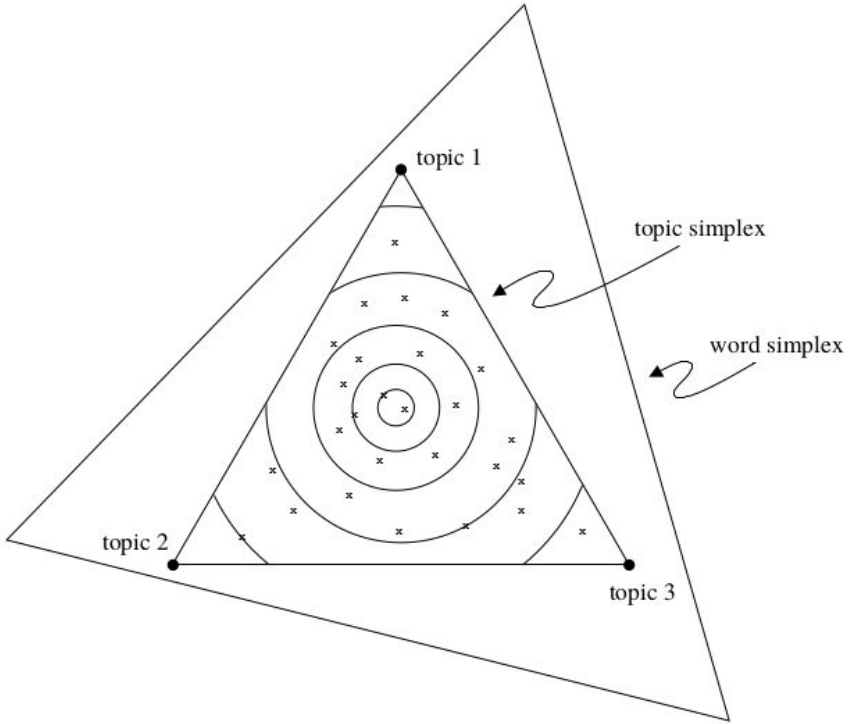
- Los vértices del simplex de palabras representan la p=1 de cada palabra
- Los vértices del simplex de tópicos representan los tres las distribuciones puras sobre el espacio de palabras de cada uno de los tópicos
- El modelo **unigram** es tan sólo un punto dentro del espacio de palabras (el simplex de palabras)



Interpretación geométrica

Podemos realizar la comparación geométrica de los modelos, analizando el caso de un simplex de 3 tópicos, dentro de un simplex de 3 palabras.

- El mixture of unigrams, cada documento en encuentra en uno de los vértices del topic simplex.
- pLSI induce, para cada ejemplo de training un punto dentro del topic-simples. (cruces).
- LDA define una distribución suave sobre el topic-simplex. (Líneas de contorno).



INFERENCIA

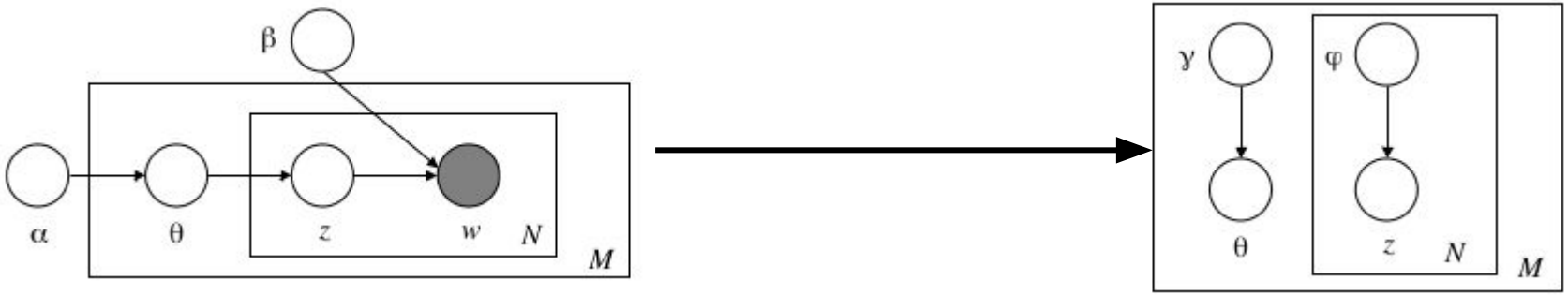
Cuando encaramos el problema de LDA, no contamos con los tópicos, ni la distribución de los mismos, sino con las palabras y documentos.

- La clave es descubrir, a partir de las variables observadas, las variables latentes u ocultas, que propone el modelo. Bayesian style

$$p(\theta, \mathbf{z} \,|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} \,|\alpha, \beta)}{p(\mathbf{w} \,|\alpha, \beta)}$$

- El problema es que esta ecuación es intratable, por la interacción entre θ y β.

- Para resolverlo, se recurre a la *inferencia variacional*:
  - La intuición es que se busca una familia de modelos que se sabe que son una cota inferior de probabilidad, y son tratables.
  - Estos modelos tienen *parámetros variacionales*, que se ajustan para obtener el modelo que más ajusta la cota inferior.
  - La forma de obtener una familia de modelos tratables es considerar algunas modificaciones sobre el modelo gráfico original, removiendo nodos y aristas.



- Para entrenar el modelo, se utiliza como método de estimación de parámetros del tipo de *Empirical Bayes*:
- Dado un Corpus D = {$\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M$}, buscamos los parámetros de **α** y **β** que maximizan el loglikelihood de los datos:

$$\ell(\alpha, \beta) = \sum_{d=1}^{M} \log p(w_d \mid \alpha, \beta)$$

- Como vimos, este problema no tiene solución analítica, y para resolverlo utilizamos inferencia variacional, que nos provee de una cota inferior.
- Esta cota se puede maximizar respecto de **α** y **β**

- Podemos aproximar los estimadores de *Empirical Bayes* alternando el proceso de *variational Expectation Maximization (EM)*. Este es un algoritmo iterativo, que alterna entre dos momento

  - (paso E): Optimizamos los parámetros variacionales γ y φ.

  - (paso M): Para valores fijo de γ y φ, maximizamos la cota inferior respecto a los parámetros del modelo, α y β.

- Estos dos pasos se alternan hasta converger.

- Por último, se realiza un suavizado sobre las probabilidades que un tópico asigna a cada palabra del vocabulario, para que siempre sean mayores a 0. Esto evita los problemas de los ítems de vocabulario no observados en los ejemplos de entrenamiento.

DYNAMIC TOPIC MODEL

- En lo visto hasta ahora, se asume que cada palabra de cada documento es una realización independiente de una mezcla de distribuciones multinomiales.

- En ciertos problemas, asumir esa independencia no es posible

- Por ejemplo, si los documentos surgen en una secuencia temporal, ésta puede afectar las distribuciones multinomiales de los tópicos.

- Un caso de este tipo de problemas lo constituyen las sucesivas ediciones de una revista a lo largo del tiempo.
  - Resulta interesante ver cómo varía la composición de los tópicos a lo largo del tiempo.
  - Para ello, se realiza algunas modificaciones sobre topic modeling, para poder captar el análisis de serie de tiempo

- En el modelo original, asumimos que los documentos son realizaciones independientes, y por lo tanto intercambiables entre sí
- en el modelo dinámico se supone que los datos están divididos por porciones de tiempo, por ejemplo anuales.
- Modelamos los documentos para cada momento con un modelo de K componentes
- Los tópicos se asocian al momento *t*, que evoluciona de los tópicos del momento *t-1*

La distribución de Dirchlet del modelo tradicional no permite realizar este tipo de modelados secuenciales.

Para un modelo con K tópicos y V términos, definimos $\beta_{t,k}$ como un vector V-dimensional de parámetros para el tópico *k* en el momento *t*.

Este parámetro evoluciona para cada tópico en un espacio con ruido gaussiano. Es un encadenamiento de variables aleatorias mediante una distribución gaussiana, que luego se mapea al simplex del modelo original

$$\beta_{t,k} | \beta_{t-1,k} \sim \mathcal{N}(\beta_{t-1,k}, \sigma^2 I)$$

La distribución de Dirchlet del modelo tradicional no permite realizar este tipo de modelados secuenciales.

En el modelo dinámico, los parámetros que evolucionan en el tiempo son α y β, que en el modelo clásico tenían distribución de Dirichlet.

β evoluciona para cada tópico en un espacio con ruido gaussiano.

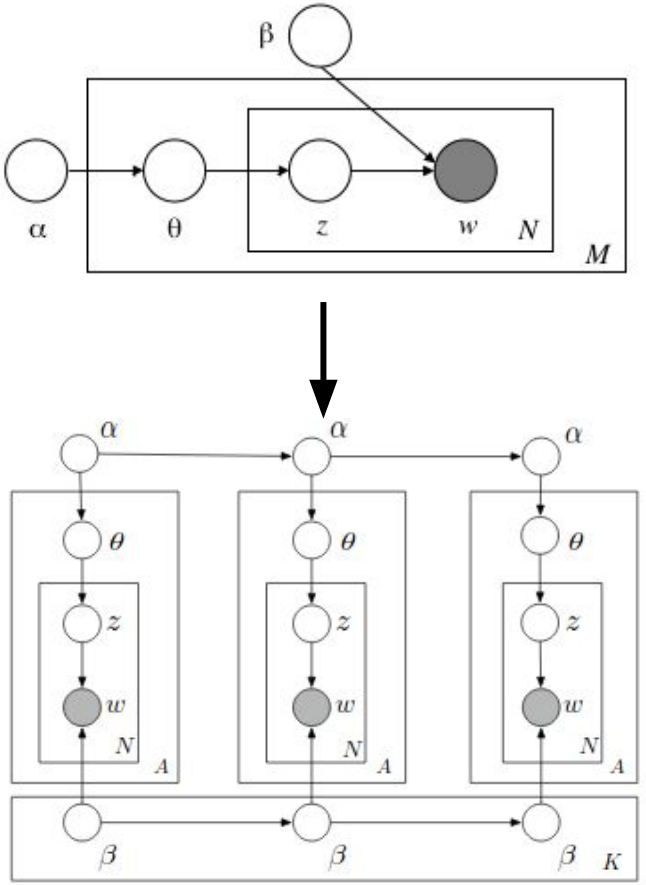
α evoluciona como una normal logística.

$$\beta_{t,k} | \beta_{t-1,k} \sim \mathcal{N}(\beta_{t-1,k}, \sigma^2 I)$$

$$\alpha_t | \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$$

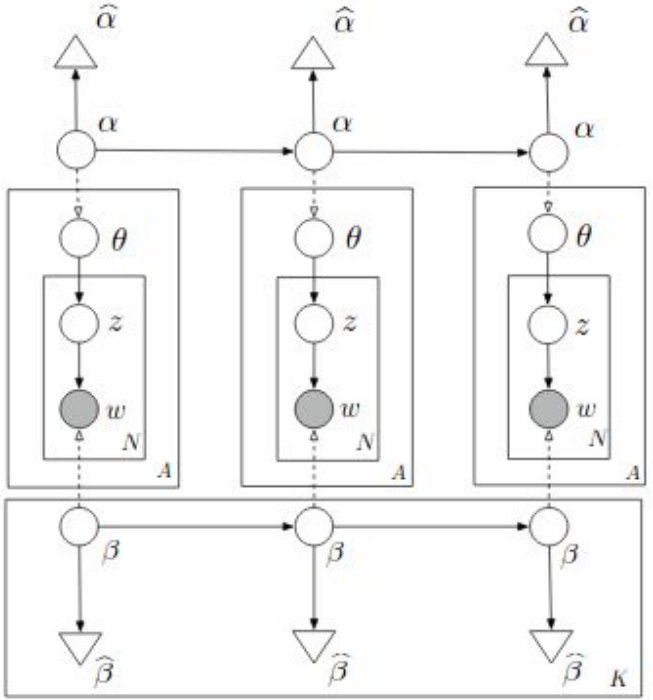
- α define la incertidumbre respecto de la distribución de los tópicos en los documentos
- β define la distribución de las palabras en los tópicos

Si bien el modelo es conceptualmente muy similar al clásico, el encadenamiento de los parámetros α y β implica una modificación de las funciones utilizadas para modelar la distribución estos parámetros.



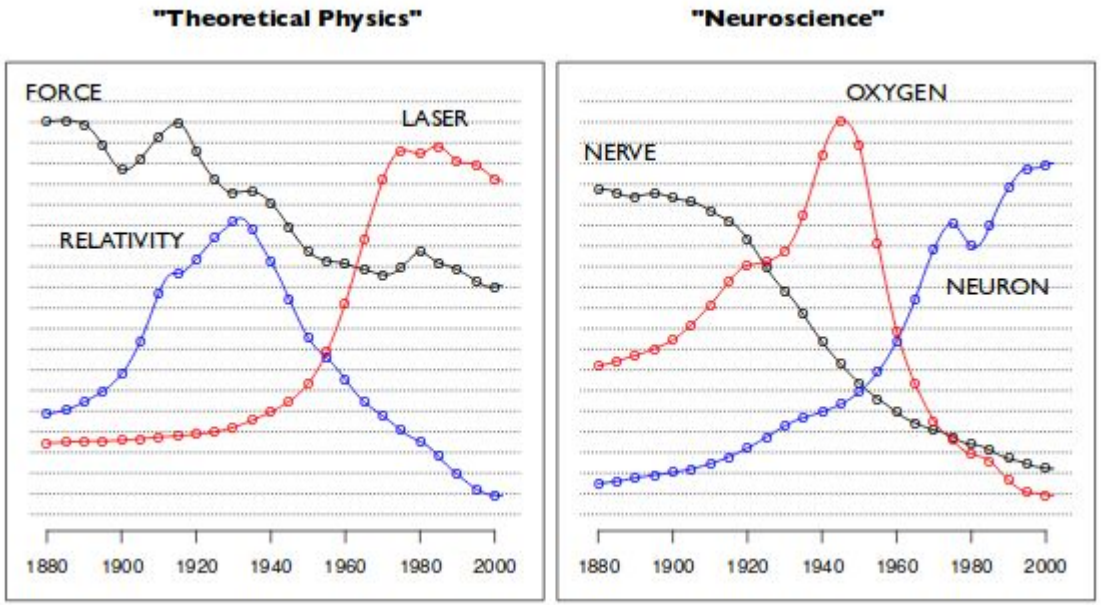
- Sin entrar en detalles, este modelo tampoco tiene una resolución computable, y también debe utilizar métodos de inferencia variacional.

- La diferencia con el modelo tradicional es que la distribución de Dirichlet tiene una solución mediante métodos variacionales que es más simple que en el caso de las distribuciones gaussianas.

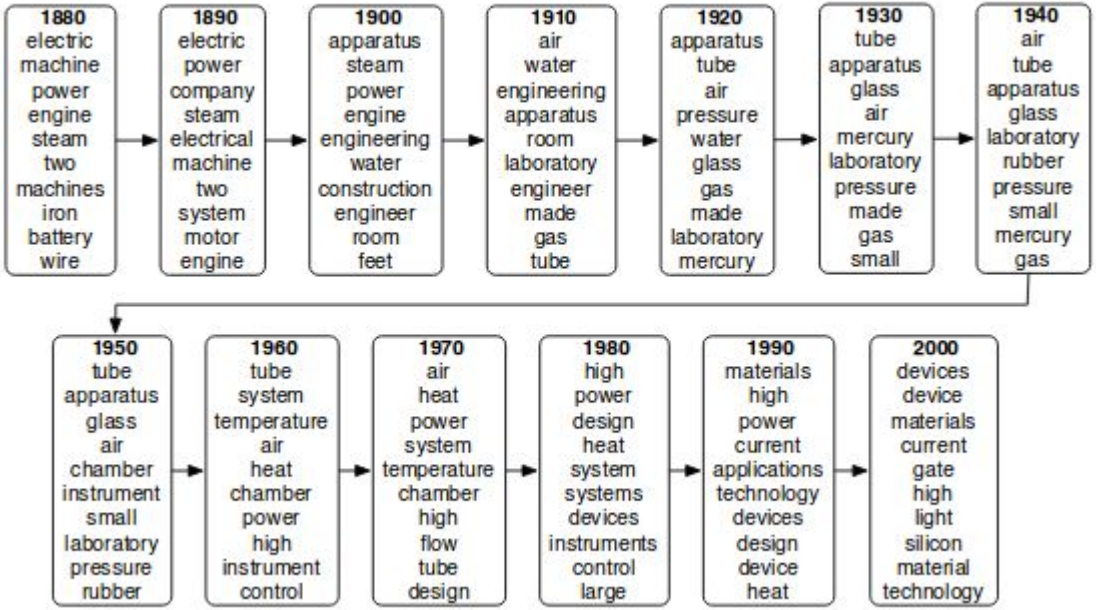


Los tópicos dinámicos permiten seguir la importancia a lo largo del tiempo de las palabras dentro de un tópico.

En este ejemplo, tenemos la evolución de algunas palabras en los tópicos *Física Teórica* y *Neurociencia*, de la revista *science*



Otra forma de analizar los resultados, es visualizar la evolución de las palabras más frecuentes dentro del tópico.



- Otra extensión del modelo original es el denominado *correlated topic model*.

- Es natural esperar que haya tópicos que están muy relacionados entre sí. Por ejemplo, si pensamos en el ejemplo clásicos del corpus de la revista *science,* es natural esperar mayor co-ocurrencia entre los tópicos *química* y *biología* que entre estos y *teoría política*.

- El problema surge por la distribución de Dirichlet no permite establecer una matriz de correlación entre los tópicos.

- Para poder definir la correlación, se reemplaza la distribución de Dirichlet por una distribución logística-normal, que incorpora una estructura de **covarianzas** entre los componentes.

CORRELATED TOPIC MODELS

- Como se mencionó en los modelos dinámicos. Eliminar la distribución de Dirichlet implica un sacrificio en la facilidad para computar la distribución a posteriori.
- El proceso generativo se redefine de la siguiente manera:

  - Antes, la proporción de los tópicos, θ, surgía de una distribución de dirichlet.
  - Ahora, se define un *parámetro natural* para definir la distribución de los tópicos, $\boldsymbol{\eta}$:

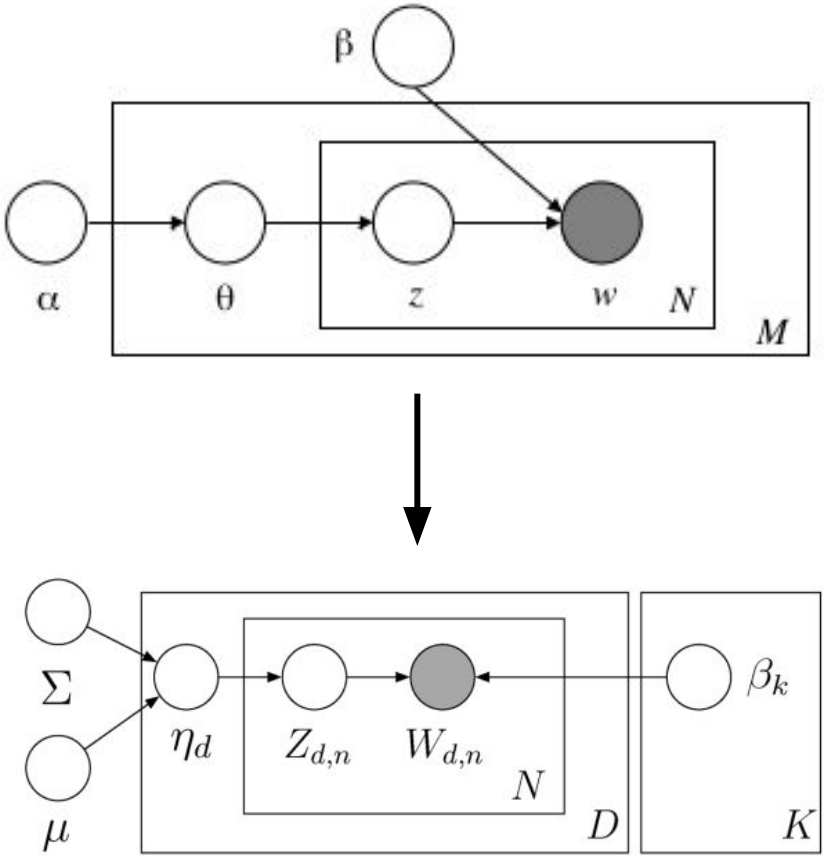
$$\boldsymbol{\eta}_d \sim \mathcal{N}(\mu, \Sigma)$$

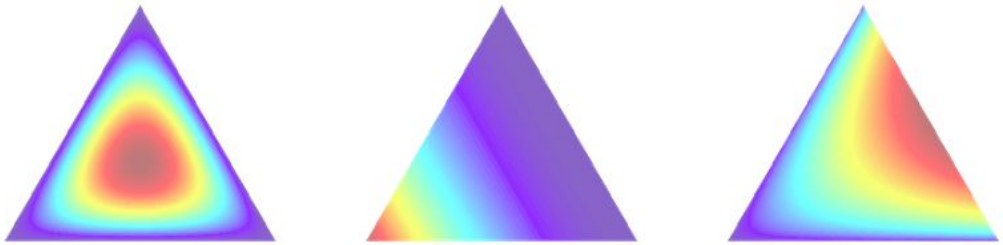
$$\theta_d = f(\boldsymbol{\eta}_d) = \frac{e^{\boldsymbol{\eta}_d}}{\sum_i e^{\boldsymbol{\eta}_i}}$$

- El parámetro $\boldsymbol{\eta}$, que define la proporción de los tópicos, θ, surge de una gaussiana en lugar de una Dirchlet

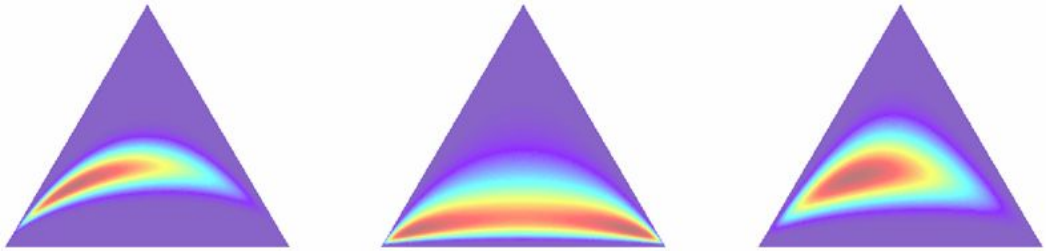
- Los β, que en el modelo original eran realizaciones de una dirichlet. Ahora se definen de forma repetida, como K vectores de distribución de las palabras en los tópicos

- θ se reemplaza por un parámetro natural de θ, que no surge de una dirichlet, sino de una normal-log





Dirichlet Simplex:
No hay correlación
(convexidad)

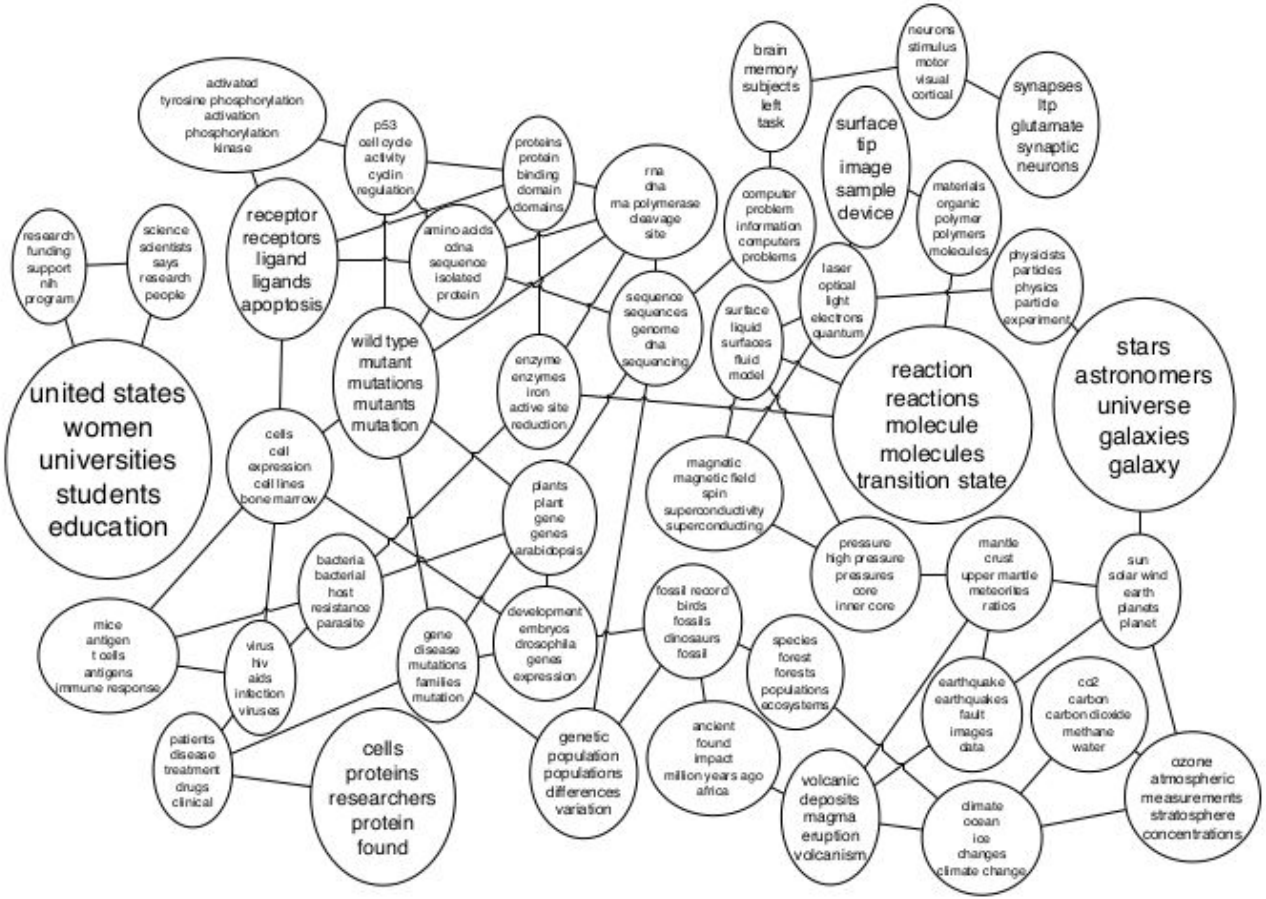


Logit-normal Simplex:
Correlación entre
tópicos

Correlated Topic Models

La matriz de correlación nos permite establecer las conexiones entre los tópicos.

En este ejemplo de la revista *science*, a partir de las correlaciones, se puede construir un grafo de los tópicos, ponderados por su importancia relativa en el corpus.



CONCLUSIONES

A MODO DE CIERRE

- En esta clase vimos el detalle de la construcción del modelo LDA.
- También vimos extensiones del mismo dentro del procesamiento de textos con modelos dinámicos y con correlación.
- Es importante tener en cuenta que es modelo no sirve solamente para análisis de texto.
- Podemos aprovechar las implementaciones de LDA para cualquier modelo en el cual tengamos distribuciones encadenadas de esta forma.
- A mencionar: ELMo, lda2vec