

---

## Latente Dirichlet Allocation

### 1 Introducción

El análisis de grandes volúmenes de texto es una tarea recurrente en numerosos procesos actuales. A diferencia de los modelos clásicos de recuperación de información los modelos tópicos permite descubrir la estructura semántica oculta en una colección de documentos, basándose en la extracción de 'temas'. Uno de los algoritmos más utilizados es Latente Dirichlet Allocation (LDA).

### 2 Modelo LDA: Fundamentos

LDA es un modelo generativo probabilístico que utiliza distribuciones de Dirichlet para inferir temas a partir de documentos. Utiliza un enfoque bayesiano para analizar los datos, utilizando probabilidades condicionales para inferir temas basados en las palabras presentes en los documentos que analiza. El proceso generativo de LDA implica generar datos similares a los observados para comprender sus características subyacentes. Al terminar su entrenamiento LDA habrá calculado la proporción en la que ocurre cada uno de los  $K$  tópicos por documento. Es necesario aclarar que el número de tópicos a extraer es un hiperparámetro para este modelo. Elegir el número de tópicos en LDA es una decisión importante que influye directamente en los resultados obtenidos.

### 3 Línea de trabajo de LDA

#### 1. Inicialización:

Por cada documento de forma independiente se asigna un tópico a cada palabra presente en el texto en cuestión. Al finalizar se tiene por cada tópico una colección de palabras, donde cada palabra puede ocurrir en más de un tópico a la vez.

#### 2. Reasignación de tópicos:

Iterando por los documentos de la colección se reasigna un nuevo tópico a cada una de las palabras del documento pero esta vez basándose en:

- Cuántas veces el documento utiliza cada tema, medida por los recuentos de frecuencia calculados durante la inicialización y una distribución multinomial generada por Dirichlet sobre los temas para cada documento.

- Cuántas veces cada tema utiliza la palabra, medida por los recuentos de frecuencia calculados durante la inicialización y una distribución multinomial generada por Dirichlet sobre las palabras para cada tema.

Luego se calcula la probabilidad condicional de que la palabra pertenezca a ese tema, dado que el tema aparece en el documento. Una vez calculadas las probabilidades condicionales para todos los temas, la palabra se asigna al tema con la mayor probabilidad.

### 3. Aumento de la precisión:

La reasignación de tópicos (paso 2) se refina de manera iterativa hasta que alcanza un criterio de convergencia predefinido. Esto significa que el modelo ha llegado a un estado en el que los cambios en la asignación de tópicos son mínimos o insignificantes (otra alternativa a la condición de parada es definir el número de veces como hiperparámetro).

El cálculo de la probabilidad condicional de la existencia de una palabra en un tópico se efectúa aplicando la fórmula de probabilidad condicional sujeta a la estructura de la red bayesiana sobre la que se construyó el modelo (Anexos 1 y 2) .

## 4 Implementación y evaluación de LDA

Utilizando la implementación de la librería Gensim de Python, se verá la influencia de diferentes configuraciones de hiperparámetros en el rendimiento del modelo LDA. Los resultados serán evaluados a través de diversas métricas, cuya selección es fundamental debido a la complejidad presente en la tarea de evaluación de modelos de tópicos ya que la extracción de temas de un documento puede ser bastante subjetiva.

### 4.1 Dataset

Debido al bajo poder de cómputo disponible para el entrenamiento y posterior evaluación del modelo solo se tomarán como elementos del dataset los 20 primeros documentos del dataset Cranfield de la librería `ir_datasets` de Python.

## 5 Hiperparámetros

1. **K**: Número de temas.
2. **Hiperparámetro de Dirichlet  $\alpha$**  : Densidad de documentos y temas.
3. **Hiperparámetro de Dirichlet  $\beta$**  : Densidad de palabras y temas.

La distribución de Dirichlet, requiere como parámetro un vector  $\alpha$  cuya dimensión coincide con la cantidad de categorías que se desean modelar, en LDA la cantidad de tópicos y la cantidad de palabras. Las componentes de  $\alpha$  controlan la forma de la distribución y por ende la probabilidad de observación de un elemento u otro del espacio de posibles valores que puede tomar variable aleatoria. (Ver Anexo 3)

## 6 Métricas

### 6.1 Basadas en el juicio humano

Pueden ayudar a identificar problemas que las métricas automáticas pueden pasar por alto. Sin embargo, también tienen limitaciones, como la subjetividad del juicio humano y la dificultad de obtener evaluaciones consistentes de diferentes expertos.

1. **Word Cloud:** Permite visualización de las palabras más probales de un tema mediante la proporción directa entre la probabilidad de la palabra y el tamaño en el que se muestra en la imagen. Anexo

### 6.2 Métricas cuantitativas

Mientras que los métodos cualitativos se basan en el juicio humano, las métricas cuantitativas proporcionan una evaluación más objetiva y automatizada, permitiendo comparar modelos de manera más precisa.

1. **Coherence:**  
Las medidas de coherencia temática valoran la calidad de un tema al medir qué tan relacionadas semánticamente están las palabras más representativas de ese tema. Esto permite identificar temas con sentido y descartar aquellos que son resultado de patrones aleatorios.

## 7 Resultados


## 8 Conclusión

## 9 Anexos

Anexo 1: Red Bayesiana

Anexo 2: Formula de probabilidad condicional.

Anexo 3: Dirichlet variando alpha

Código fuente disponible en:  GitHub