Naomi Lahera Champagne
Universidad de La Habana

Tarea No.2

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 aprovecha las propiedades de la distribución de Dirichlet para alcanzar su objetivo. Como modelo tópico probabilista a partir de probabilidades condicionales es capaz de inferir temas basados en las palabras presentes en los documentos que analiza. Al terminar su entrenamiento por cada documento LDA habrá calculado la proporción en la que ocurren cada uno de los K tópicos que inicialmente debían ser conformados. 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 Linea de trabajo de LDA

1. **Inicialización:**
   Por cada documento de forma independiente es asigado 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 de la palabra 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 formúla de probabilidad condicional sujeta a la estructura de la red bayesiana sobre la que se construyó el modelo.

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 nivel de hardware disponible para el entrenamiento y posterior evaluación del modelo solo se tomarán como corpus los 700 primeros documentos del dataset Cranfield de la librería ir_datasets de Python.

5 Hiperparámetros

1. **K**: Número de temas.
2. **Hiperparámetro de Dirichlet α :** Densidad de documentos y temas.
3. **Hiperparámetro de Dirichlet β :** Densidad de palabras y temas.

La distribución de Dirichlet, requiere como parámetro un vector α cuya dimensión coincide con la cantidad de categorías que se desean modelar; en el caso LDA serían la cantidad de tópicos y la cantidad de palabras. Las componentes de α 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.

6 Métricas

6.1 Basadas en el juicio humano

Las métricas 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:** Esta métrica permite visualizar las palabras con courrencia más probales en un tema mediante la proporción directa entre la probabilidad de la palabra y el tamaño en el que se muestra en la imagen (Anexo 1).

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

Las pruebas realizadas barrieron los valores de los hiperparámetros en los siguientes rangos:

- k: $3 \leq k \leq 5$
- β: $0,1 \leq \beta \leq 1$ con un salto de 0.3. Se tuvo en cuenta el valor $\frac{1}{k}$ conocido en gensimm como *symetric*. En este caso el valor escogido es común a cada una de las componentes del vector.
- $\alpha = 0,1 \leq \beta \leq 1$ con un salto de 0.3. El valor seleccionado es común a cada una de las componentes del vector. Se tuvieron en cuenta también los valores $\frac{1}{k}$ y $\alpha =< a_1, a_2..., a_k >$ donde $a_i = \frac{1}{i+\sqrt{k}}$ conocidos en gensim en gensim como *symetric* y *asymetric* respectivamente.

El mejor resultado alcanzado fue con los valores:

- k = 5
- $\beta = 0.7$ (en cada una de las componentes del vector).
- $\alpha =< a_1, a_2..., a_k >$ donde $a_i = \frac{1}{i+\sqrt{k}}$ $(asymetric)$

Los valores de alpha y beta probados no muestran una influencia tan marcada en la coherencia como el número de tópicos. La combinación de 5 tópicos y alpha asimétrico resultó ser la más efectiva en este análisis preliminar.

8 Conclusiones

El número de tópicos parece ser el factor más influyente en la coherencia del modelo en este caso, pero es necesario realizar más experimentos para confirmar esta tendencia y optimizar los otros parámetros. Los resultados obtenidos son específicos para el conjunto de datos analizado. Es fundamental reconocer que la configuración óptima de los hiperparámetros puede variar significativamente entre diferentes corpus.

Es recomendable explorar un rango más amplio de valores para los hiperparámetros, especialmente para el número de tópicos (K) y (β). Se recomienda además ampliar el corpus utilizado para el análisis del modelo ya que en el presente estudio no fue posible emplear un datset más extenso debido al bajo nivel del hardware utilizado para realizar estas pruebas.

9 Anexos

Anexo 1

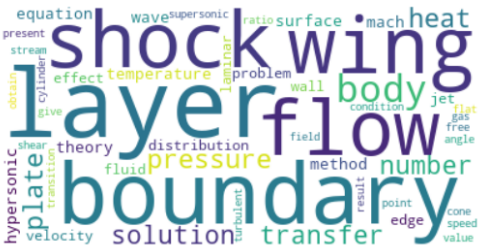

Figure 1: Métrica Word Cloud asociada a uno de los tópicos extraídos por el modelo final presentado en este informe.

Anexo 2
Código fuente disponible en: GitHub