Análisis de tópicos: minando tópicos en R

Salvador López Mendoza

Junio de 2018

Problema

Dada una colección de documentos y un valor de k, encontrar los k tópicos y clasificar los documentos de acuerdo a esos tópicos.

Se realizarán dos ejercicios:

1. Textos en inglés preparados para minado de tópicos.
   Son sobre temas diversos.
2. Noticias de un periódico.
   Seleccionadas de acuerdo a las secciones del periódico.
   Textos en **español**.

En ambos casos se usa *LDA (Latent Dirichlet Allocation)*.

PEUVI

LDA

Todos los documentos son una mezcla de varios tópicos.

Al examinar los documentos sólamente se ven frases o palabras.

Los tópicos están ocultos (están latentes) en la estructura de los documentos.

El objetivo es inferir la estructura latente de cada tópico dadas las palabras que conforman al documento.

LDA recrea los documentos del corpus ajustando la *importancia relativa* de los tópicos en los documentos y de las palabras en los tópicos. **Es un proceso iterativo.**

PEUVI

LDA. Funcionamiento

1. Procesa cada documento. Asigna aleatoriamente cada una de las palabras del documento a uno de los k tópicos.
   Esta asignación proporciona una representación de los tópicos de todos los documentos y de la distribución de las palabras en cada tópico.
2. Mejora de la representación.

Proceso de mejora

Para cada documento d:

1. Toma cada palabra w del documento d.
2. Para cada tópico t, cálcula:
   1. $p(t|d)$. La proporción de palabras en el documento d que están asignadas al tópico t.
   2. $p(w|t)$. La proporción de asignaciones al tópico t sobre todos los documentos que tienen la palabra w.
   3. Reasigna w a un nuevo tópico.
      Es el tópico t con probabilidad $p(t|d) * p(w|t)$.
      Es la probabilidad de que el tópico t haya generado a la palabra w.

Es un **proceso iterativo**. Después de una gran cantidad de iteraciones se llega a un estado de estabilidad.

Se usan estas asignaciones para estimar la mezcla de tópicos de cada documento del corpus.

PEUVI

El paquete *topicmodels* en R

Es un paquete con funciones para hacer minería de tópicos.

Es un complemento del paquete *tm*.

La función principal es *lda*.

También tiene otra función, *ctm* para el modelado de tópicos correlacionados.

La función *lda*

Invocación:

$$lda(x, k, method = \text{“}metodo\text{”}, control = lista, model = \text{“}modelo\text{”})$$

Parámetros:

- x. Matriz de documentos-términos.
- k. Cantidad de tópicos.
- *method*. Método de estimación usado. Puede ser *VEM* o *Gibbs*.
- *control*. Parámetros para el método seleccionado.
- *model*. Especificación de un modelo distinto a los conocidos.

PEUVI

Parámetros del método Gibbs

alpha. Valor de α. Por default es $50/k$.

estimate.beta. Valor booleano, indica si β debe ser estimado o toma un valor fijo. El default es *TRUE*.

verbose. Indica si debe imprimir durante la ejecución del algoritmo. Es valor entero, positivo indica cada cuántas iteraciones imprime.

prefix. Prefijo del nombre del archivo en el que guarda resultados intermedios.

save. Valor entero, 0 indica que no guarda resultados intermedios. Valor positivo, cada cuántas iteraciones guarda los resultados intermedios.

keep. Valor entero. Si es positivo indica cada cuántas iteraciones guarda los valores del logaritmo de la similitud.

PEUVI

Parámetros del método Gibbs (II)

seed. Semilla inicial para la generación de valores aleatorios. Puede ser un valor o una lista de valores.

nstart. Cantidad de veces que se realizará el proceso. La lista de valores de *seed* debe ser de tamaño igual a *nstart*.

best. Valor booleano. Si es *TRUE* sólo regresa la muestra con el valor más alto de *similitud posterior*.

delta. Parámetro para la distribución *prior* de la distribución de términos sobre los tópicos.Se recomienda un valor de 0.1.

iter. Indica cuántas muestras Gibbs se realizan.

burnin. Indica cuántas iteraciones inciales se desca[illegible].

thin. Cada cuántas iteraciones se regresa un resultado durante las *iter* iteraciones.

PEUVI

Funciones adicionales

logLike(). Obtiene el logaritmo de la similitud del modelo ajustado.

perplexity(). Perplejidad del modelo.

posterior(). Obtiene la distribución de tópicos para los documentos y la distribución de términos para los tópicos.

terms(). Obtiene los k términos más importantes para los tópicos.

topics(). Obtiene los k tópicos para los documentos.

PEUVI