Latent Dirichlet Allocation (LDA)

Salvador López Mendoza

Enero de 2018

Extensiones a PLSA

PLSA es un modelo básico. Se le puede mejorar.

- PLSA con conocimientos previos.
  PLSA controlado por el usuario.
  El usuario puede esperar que aparezcan ciertos tópicos.
  En opiniones sobre laptops se espera que aparezcan *batería* y *memoria*.
  El usuario puede saber qué tópicos están presentes (o no) en los documentos.
  Son *conocimiento previo (prior)*.
- PLSA como un modelo generador.
  Latent Dirichlet Allocation (LDA).

PEUVI

Deficiencias de PLSA

- No es un modelo generador.
  No se puede calcular la probabilidad asociada a un nuevo documento.
  Hay heurísticas para lograr cierto nivel de funcionamiento.
- Tiene muchos parámetros.
  Los modelos son muy complejos.

PEUVI

LDA

Convertir PLSA en un modelo generador.

Se le impone un conocimiento previo en el modelo de los parámetros.

Se usa una *distribución Dirichlet*.

- LDA es una versión bayesiana de PLSA.
- Los parámetros están regularizados.

La cobertura de los tópicos y las distribuciones de las palabras que definen los tópicos se pueden inferir usando *inferencia bayesiana*.

PEUVI

LDA informal

Cada documento de la colección es una mezcla de varios tópicos.

Un usuario solo puede observar los documentos y las palabras que los conforman.

Los tópicos son parte de la estructura *oculta (o latente)* de los documentos.

LDA infiere la estructura latente de los tópicos a partir de las palabras y los documentos.

LDA recrea los documentos en el corpus ajustando la importancia relativa de los tópicos en los documentos y de las palabras en los tópicos.

Es un proceso *iterativo*.

PEUVI

Funcionamiento

Para cada documento, asigna aleatoriamente cada palabra del documento a uno de los k tópicos.

Se obtiene una representación de los tópicos de todos los documentos y la distribución de todos los tópicos.

No es una buena representación.

PEUVI

Funcionamiento (II)

Para mejorar las asignaciones:

- Se toma cada documento d, y para cada palabra w del documento d.
- Para cada tópico θ, calcular $p(\theta|d)$.
  Es la proporción de palabras en el documento d que están asignadas en este momento al tópico θ.
- Calcular $p(w|\theta)$.
  Es la proporción de asignaciones al tópico θ sobre todos los documentos que tienen la palabra w.
- Reasignar w a un nuevo tópico.
  Se toma el tópico θ cuya probabilidad $p(\theta|d)p(w|\theta)$ sea la mayor.
  Se asume que todas las asignaciones a los tópicos son correctas, sólo falta asigar la palabra w.

El proceso se repite muchas veces.

En algún momento se estabiliza.

PEUVI