# Análisis de tópicos: términos

Salvador López Mendoza

Mayo de 2018

# Problema

- Entrada.
  - Una **colección** de $N$ documentos de texto.
  - Una cantidad $k$ de tópicos.
- Salida.
  - $k$ **tópicos**: $\{\theta_1, \ldots, \theta_k\}$
  - Porcentaje de cobertura de los tópicos en cada documento $d_i$: $\{\pi_{i1}, \ldots, \pi_{ik}\}$

Se cumple que:

$$\sum_{j=1}^{k} \pi_{ij} = 1$$

Entonces, $\pi_{ij}$ es la probabilidad de que el documento $d_i$ cubra el tópico $\theta_j$

**¿Cómo se define $\theta_i$?**

PEUVI

# Término = tópico

Se define cada tópico por un término que aparezca en los documentos.

Los términos deben ser **característicos**.

Ejemplo: En un periódico las noticias están agrupadas por secciones:

Política, Economía, Deportes, Cultura, etc.

**Para determinar a qué sección corresponde cada artículo, ¿qué palabra (o término) sirve para discriminar?**.

PEUVI

# Ejercicio

*El mediocampista mexicano Marco Fabián aseguró que tras pasar por momentos difíciles el año pasado se siente más fuerte y está listo para ayudar al Eintracht Frankfurt en la segunda vuelta de la Bundesliga y para pelear por un lugar para la Copa del Mundo de Rusia 2018. En entrevista a la revista del Eintracht, Fabián dijo que está ansioso de regresar a las canchas.*

PEUVI

# Ejercicio (II)

| **Tópico** | **Doc-1** | **Doc-2** | ... | **Doc-N** |
|---|---|---|---|---|
| $\theta_1$ (deportes) | $\theta_{11} = 30\,\%$ | $\theta_{21} = 0$ | ... | $\theta_{N1} = 0$ |
| $\theta_2$ (viajes) | $\theta_{12} = 12\,\%$ | $\theta_{22} = 25$ | ... | $\theta_{N2} = 10$ |
| $\vdots$ | $\vdots$ | $\vdots$ | ... | $\vdots$ |
| $\theta_k$ (ciencia) | $\theta_{1k} = 6\,\%$ | $\theta_{2k} = 15$ | ... | $\theta_{Nk} = 34$ |

**¿Cuáles son las palabras adecuadas para clasificar los documentos?**

PEUVI

# Minando $k$ tópicos

1. Analizar los textos en la colección de documentos ($C$) para obtener candidatos a términos (término = palabra).
2. Diseñar una función de calificación que permita *medir* qué tan bueno es cada término si se le usa para determinar un tópico.
   - Debe favorecer a los términos representativos (mayor frecuencia).
   - Evitar palabras demasiado frecuentes (artículos, etc.).
     Las funciones que asignan pesos en *recuperación de textos* (TF-IDF) son útiles.
   - Se pueden usar *heurísticas* específicas al dominio de trabajo (favorecer palabras del título, hashtags, etc.).
3. Tomar los $k$ términos con la calificación más alta.
   Evitar redundancia.

PEUVI

# Cobertura de los tópicos $\pi_{ij}$

¿Cómo se mide la cobertura de un tópico en un documento?

1. Contar las apariciones del término que define al tópico.
   Para el documento $d_i$:
   $count(deportes, d_i)$; $count(viajes, d_i)$; ...; $count(ciencia, d_i)$;
2. Considerar la proporción respecto a la cobertura de los otros tópicos.

$$\pi_{ij} = count(\theta_j, d_i) / \sum_{L=1}^{k} count(\theta_L, d_i)$$

PEUVI

# Efectividad

¿Qué tan bien funciona?

Sea $d_i$ el documento que se revisa:

> *... después de su derrota ante el Guadalajara, las águilas del América tuvieron que viajar a Los Ángeles para cumplir con su compromiso en ...*
>
> *... a pesar de que su estrella se lesionó ...*

¿Cuánto valen $count(deportes, d_i)$; $count(viajes, d_i)$; ...; $count(ciencia, d_i)$?

$\pi_{i1} = c(deportes, d_i) = 0$

$\pi_{i2} = c(viajar, d_i) = 1$

$\pi_{ik} = c(ciencia, d_i) = 0$

**¡Es un documento de la sección de viajes!**

PEUVI

# Problemas

1. $\pi_{i1} = c(deportes, d_i) = 0$
   Hay que considerar palabras relacionadas con *deportes*.
2. $\pi_{ik} = c(ciencia, d_i) = 0$
   Considerando la anterior, podría considerarse a *estrella* como una palabra relacionada con *ciencia*.
   *estrella* es una palabra ambigua. Puede aparecer en contextos diversos.
3. ¿Cómo minar tópicos complicados?

PEUVI

# Problemas (II)

- Falta poder de expresión.
  - Solamente se pueden representar tópicos muy sencillos o muy generales.
  - No es posible representar tópicos complicados.
- Cobertura incompleta del vocabulario.
  - No es posible capturar las variaciones en el vocabulario (palabras relacionadas, sinónimos).
- Ambigüedad en la identificación de las palabras.
  - El término que define los tópicos puede ser ambiguo. Por ejemplo *estrella*.
    *la estrella de cine*, *es una estrella del fútbol*.

PEUVI