

# ANÁLISIS DE TÓPICOS

Salvador López Mendoza

Mayo de 2018

# INTRODUCCIÓN

¿Qué es un tópico?

**Es el tema de una conversación o una discusión.**

¿Todos los tópicos son iguales?

Hay diferentes niveles de granularidad.

Libro, capítulo, sección.

# APLICACIONES

- ¿Cuáles son los tópicos de las conversaciones de twitter el día de hoy?
- ¿Cuáles son los tópicos del fútbol de esta semana?
- ¿Cuál fue el tópico principal en la entrega de los Globos de Oro?
- A nivel de investigación en minería de textos, ¿cuáles son los tópicos de investigación en este momento? ¿Qué tan distintos son de los tópicos de hace 10 años?
- ¿Cuáles son los tópicos que se están debatiendo en las campañas políticas?

# DEFINICIÓN (INFORMAL)

Hay dos cuestiones a considerar.

Dado un conjunto de documentos,

- 1 ¿Cuáles son los temas que se tratan en esos documentos?
- 2 ¿Qué tan extenso es el tratamiento de esos temas en cada documento?

# DEFINICIÓN

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

# PROPIEDADES

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  $\theta_j$ ?**