Text Mining: Latent Dirichlet Allocation

Universidad de los Andes

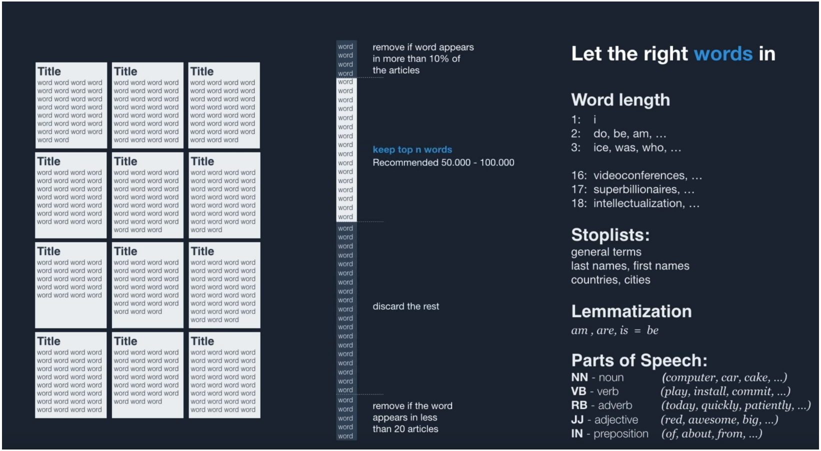
Juan Sebastián Moreno Pabón

3 de septiembre de 2019

Modelos de tópicos

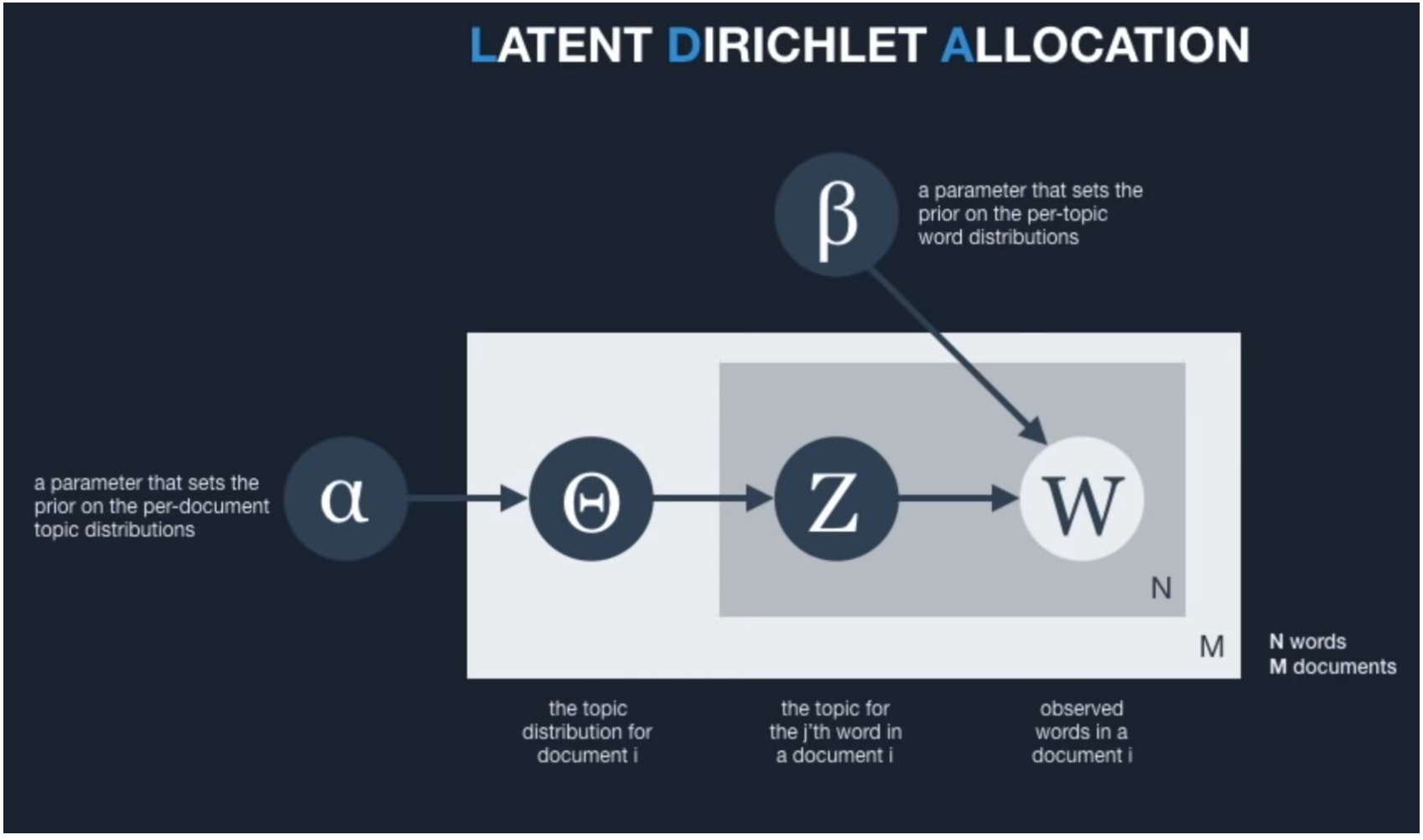
- El objetivo de un modelo de tópicos es descubrir las temáticas subyacentes en un corpus de texto.
- El objetivo es muy similar al análisis de clústeres pero para textos.
- Retos:
  - Un texto puede referirse a múltiples temáticas.
  - ¿Cómo tratar un texto como data estructurada?

Latent Dirichlet Allocation (LDA)



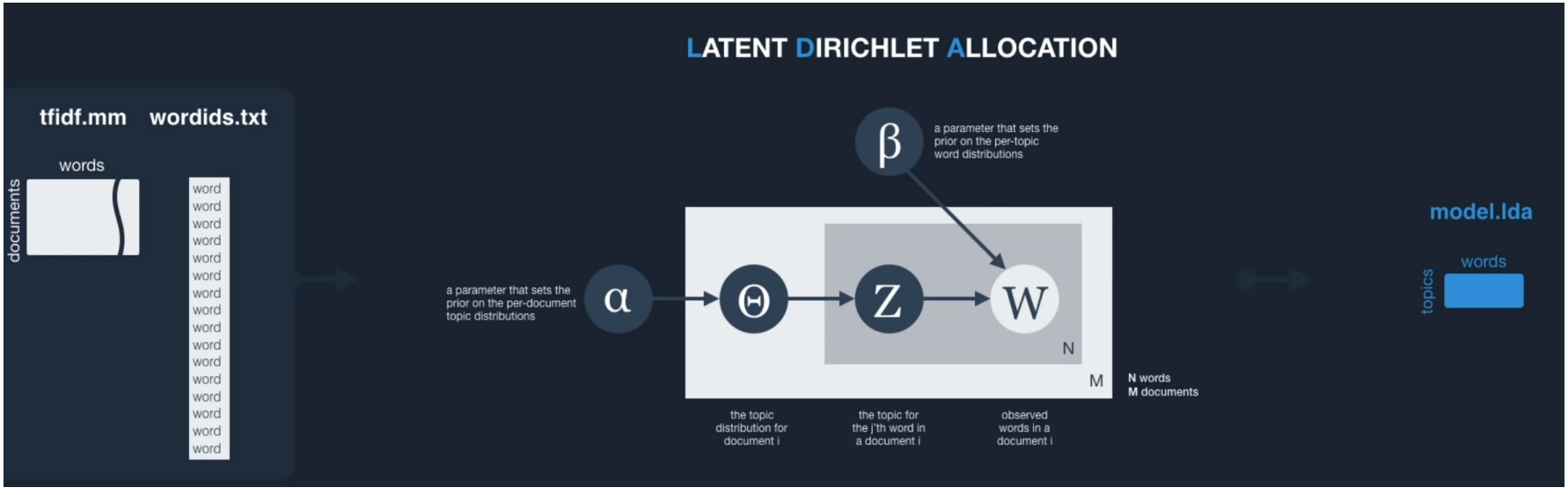
Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA)

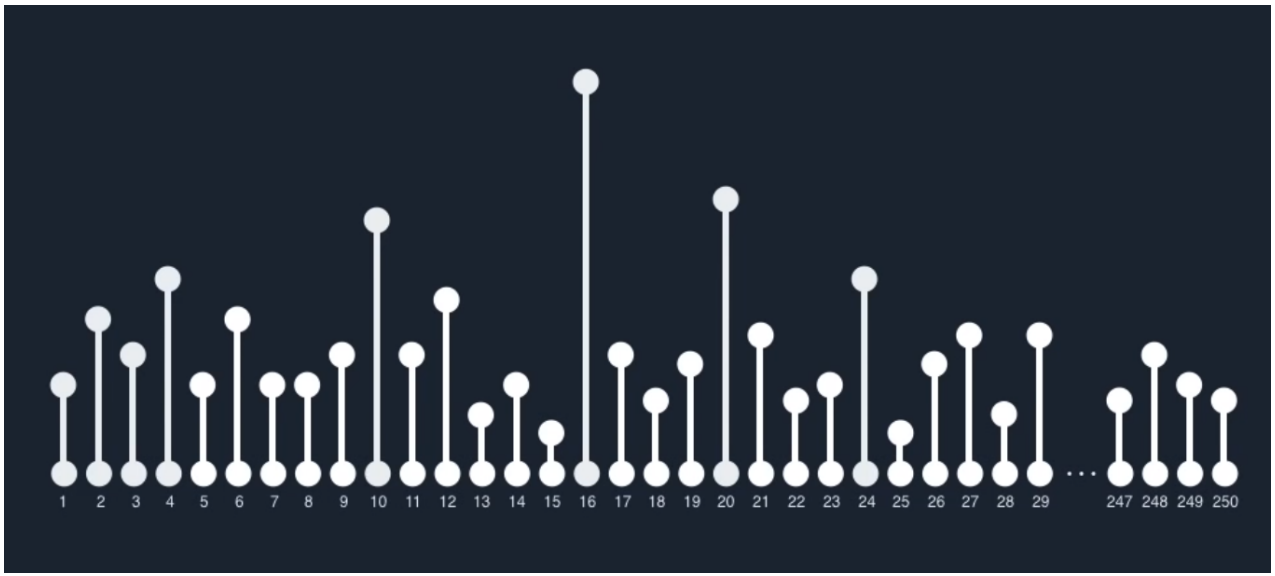


¿Qué hace Alfa y Beta?

LDA

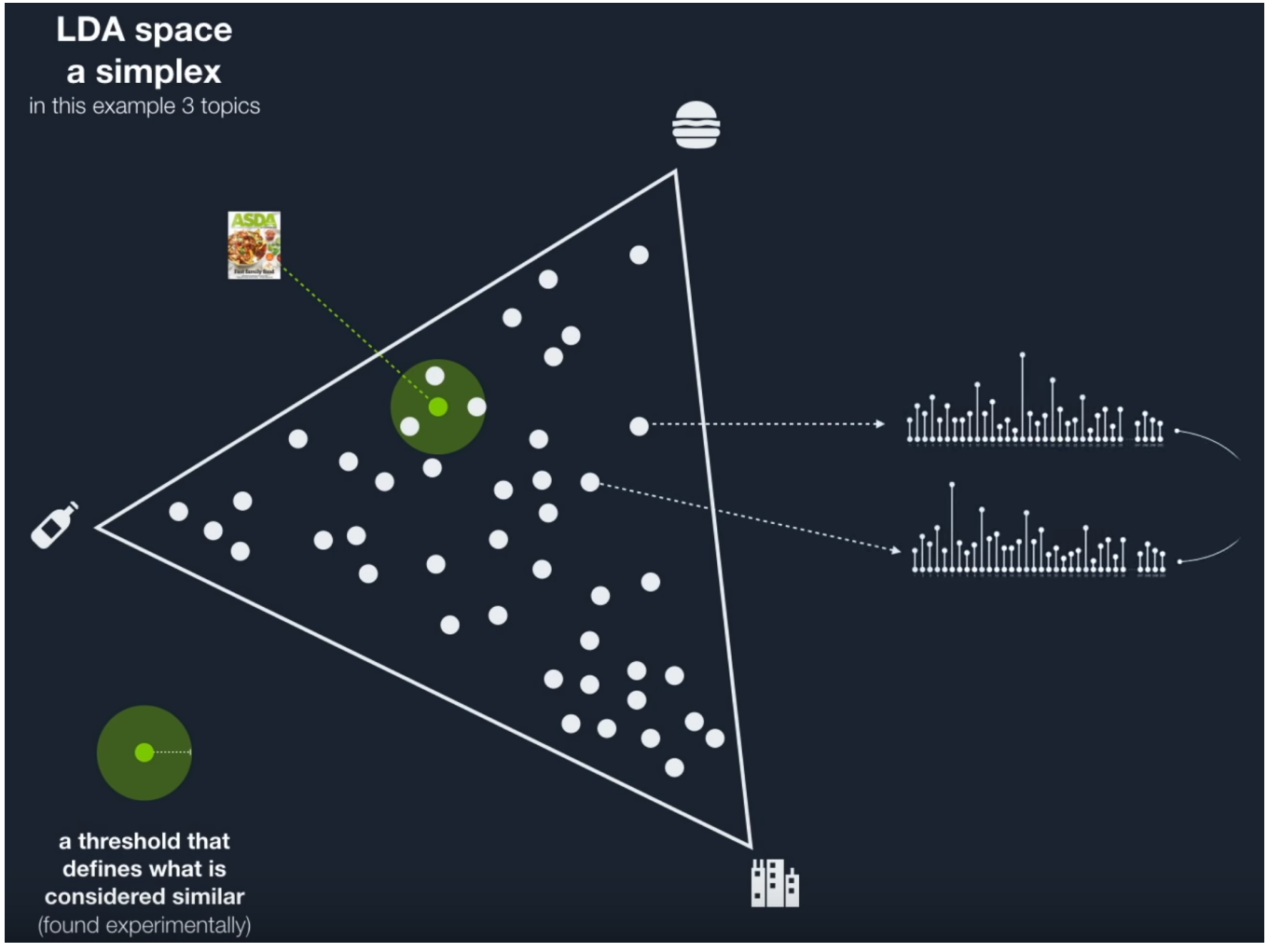

LATENT DIRICHLET ALLOCATION
tfidf.mm
wordids.txt
words
documents
word
word
word
word
word
word
word
word
word
word
word
word
word
word
word
word
word
word
a parameter that sets the prior on the per-topic word distributions
a parameter that sets the prior on the per-document topic distributions
α
Θ
Z
W
N
M
N words
M documents
the topic distribution for document i
the topic for the j'th word in a document i
observed words in a document i
model.lda
topics
words

LDA


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 ... 247 248 249 250


DNA

LDA


LDA space
a simplex
in this example 3 topics
a threshold that
defines what is
considered similar
(found experimentally)