Topic Modeling con K-NN

Jordi Gironés

Universitat Oberta
de Catalunya

Índice

- Sistemas de recomendación
- Caso de estudio
- Pre-Modelado
  - Acondicionado de datos
  - Matriz de Términos
- Nube de palabras
- Modelado
  - Data Frame para K-NN
  - Construcción del Modelo K-NN
  - Verificación del Modelo

Sistemas de recomendación

Topic Modeling

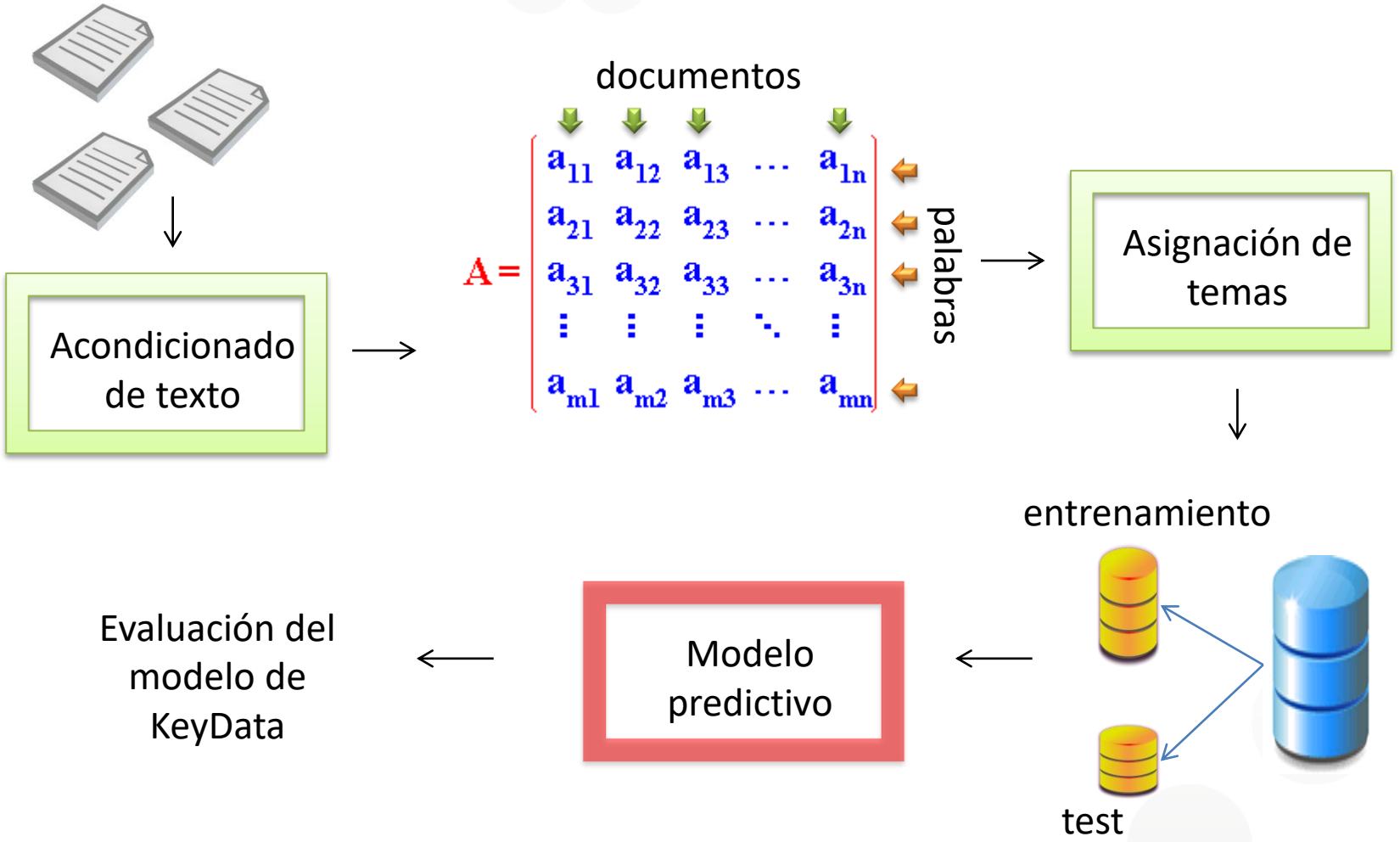
- Tiene por objetivo descubrir los conceptos más relevantes en un conjunto de documentos.

- Para ello se basa en la frecuencia de aparición de palabras y expresiones.

- La principal aplicación del Topic Modeling son los **motores de recomendación** orientados a la predicción de la relevancia de palabras o expresiones.

- El rastreo en sistemas documentales y el filtrado de contenidos son sus principales actividades.

Caso de estudio



Índice

- Sistemas de recomendación
- Caso de estudio
- Pre-Modelado
  - Acondicionado de datos
  - Matriz de Términos
- Nube de palabras
- Modelado
  - Data Frame para K-NN
  - Construcción del Modelo K-NN
  - Verificación del Modelo

Acondicionado de datos

Se realizan tareas de limpieza de texto como:

Eliminación de signos de puntuación

Eliminación de signos de espacios en blanco innecesarios

Conversión a minúsculas

Reducción de palabras a su raíz

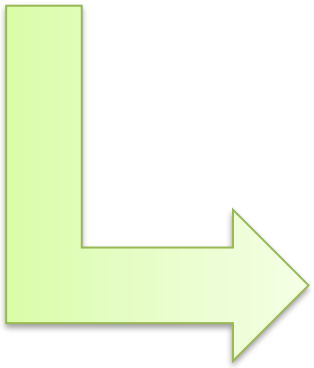
Eliminación de palabras sin significado propio

```
```{r,eval=TRUE,echo=TRUE}
acondicionaCorpus <- function(corpus) {
  corpus.tmp <- tm_map(corpus, removePunctuation)
  corpus.tmp <- tm_map(corpus.tmp, stripWhitespace)
  corpus.tmp <- tm_map(corpus.tmp, content_transformer(tolower))
  v_stopwords <- c(stopwords("english"),c("dont","didnt","arent","cant","one","also","said"))
  corpus.tmp <- tm_map(corpus.tmp, removeWords, v_stopwords)
  corpus.tmp <- tm_map(corpus.tmp, removeNumbers)
  corpus.tmp <- tm_map(corpus.tmp, stemDocument, language="english")
  return(corpus.tmp)
  }
```
```

Matriz de términos

```
# Temas que distinguiremos
temas <- c("Adq", "Crudo")
# Ruta principal de los documentos de noticias Reuters
nombreruta <- paste(getwd(),"/txt", sep = "")
```

```
tdm <- lapply(temas, generaTDM, ruta = nombreruta)
```



tdm[[1]]$name	tdm[[1]]$tdm
tdm[[2]]$name	tdm[[2]]$tdm

Lista de (carácter, matriz)

Adquisiciones	$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{pmatrix}$
Crudo	$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{pmatrix}$

Nube de palabras

```
d_adq <- data.frame(word=names(v_adq), freq=v_adq)
```

Matriz de palabras y frecuencias

```
> head(v_adq)
 dlrs share compani   pct   mln   inc
  100    86      80    70    65    55
```

```
wordcloud(d_crude$word, d_crude$freq, min.freq=3, random.color=TRUE, colors=rainbow(7))
```

Palabras

Frecuencias

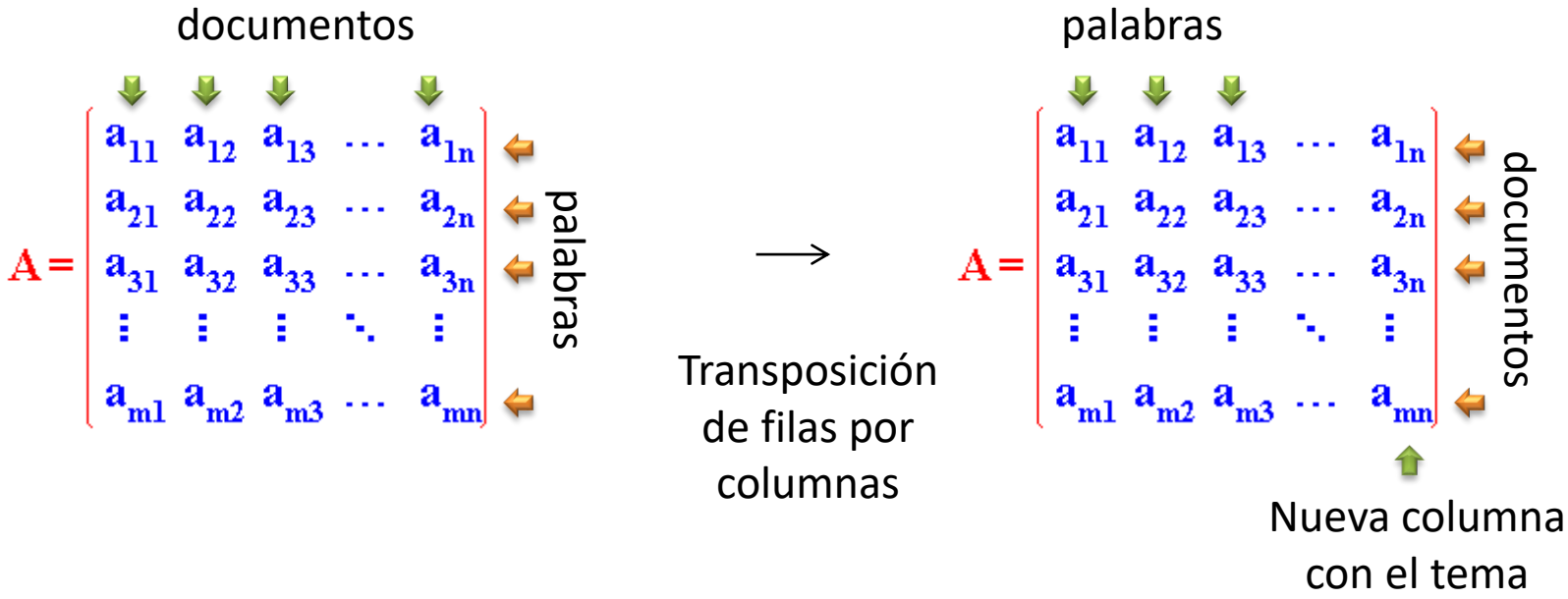
mínimo

Índice

- Sistemas de recomendación
- Caso de estudio
- Pre-Modelado
  - Acondicionado de datos
  - Matriz de Términos
- Nube de palabras
- Modelado
  - Data Frame para K-NN
  - Construcción del Modelo K-NN
  - Verificación del Modelo

Data-Frame para K-NN (1)

Función unirTemaTDM()



Lista de (carácter,data.frame)

Lista de (data.frame)

Data-Frame para K-NN (2)

Apilado de Data-Frames

temaTDM

$$\left(\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}, \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix} \right)$$

Fusión de la lista de dos Data-Frames en un solo Data-Frame

$\longrightarrow$

tdm.pila

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}$$

Lista de (data.frame)

Data - Frame

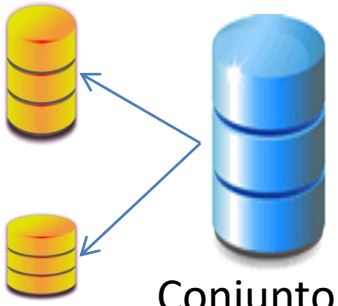
Construcción del modelo K-NN

```
entrena.idx <- sample(nrow(tdm.pila), ceiling(nrow
(tdm.pila) * 0.7))
```

entrenamiento

```
test.idx <- (1:nrow(tdm.pila))[-entrena.idx]
```

test



Conjunto de documentos

Modelo K-NN

```
knn.pred <- knn(tdm.pila.nl[entrena.idx, ], tdm.pila.nl[test.idx, ], tdm.tema[entrena.idx])
```

Matriz de frecuencias de los documentos de entrenamiento

Matriz de frecuencias de los documentos de prueba

Temas de los datos de entrenamiento

Verificación del modelo

Matriz de confusión

		Predicted	
		Negative (**N**) -	Positive (**P**) +
Actual	Negative -	True Negatives (**TN**)	False Positives (**FP**) **Type I error**
	Positive +	False Negatives (**FN**) **Type II error**	True Positives (**TP**)

Observaciones reales

Precisión (Accuracy)= (TP + TN) / (P + N)

Exactitud (Precision)= TP / (TP + FP)

Sensibilidad (Sensitivity)= TP / P

Especificidad (Specificity) = TN / N