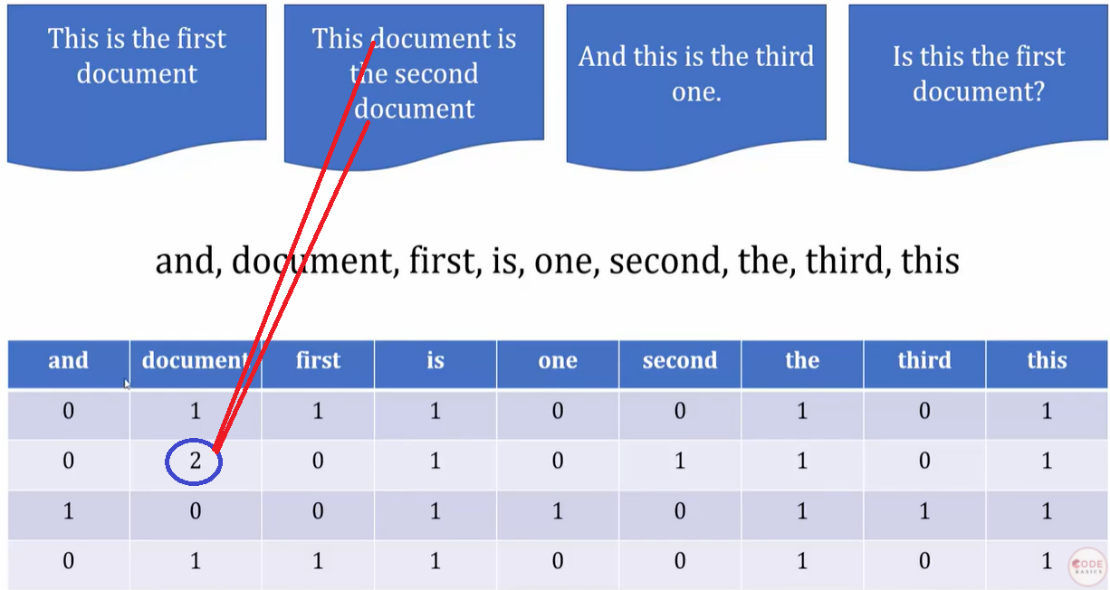
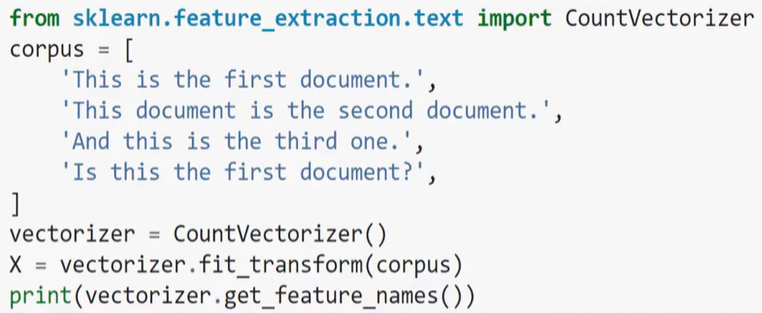
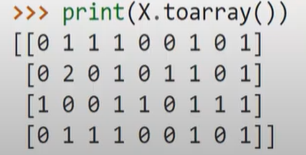
Different emails and with similar words



The CountVectorize separa las palabras y crea features donde presenta cuantas veces aparece palabras mas comunes







Se puede aliminar después features ‘and’, ‘is’, ‘the’ o previamente se eliminan, se limpia los textos

**from** sklearn.feature\_extraction.text **import** CountVectorizer

v **=** CountVectorizer()

X\_train\_count **=** v**.**fit\_transform(X\_train**.**values)

X\_train\_count**.**toarray()[:2]

array([[0, 0, 0, ..., 0, 0, 0],

[0, 0, 0, ..., 0, 0, 0]], dtype=int64