HUB IQ

Inteligencia Artificial & Machine Learning

Applicaciones en movilidad

Dr. Iván S. Razo Zapata

EngineeringX
Founded by the Royal Academy of Engineering and Lloyd's Register Foundation

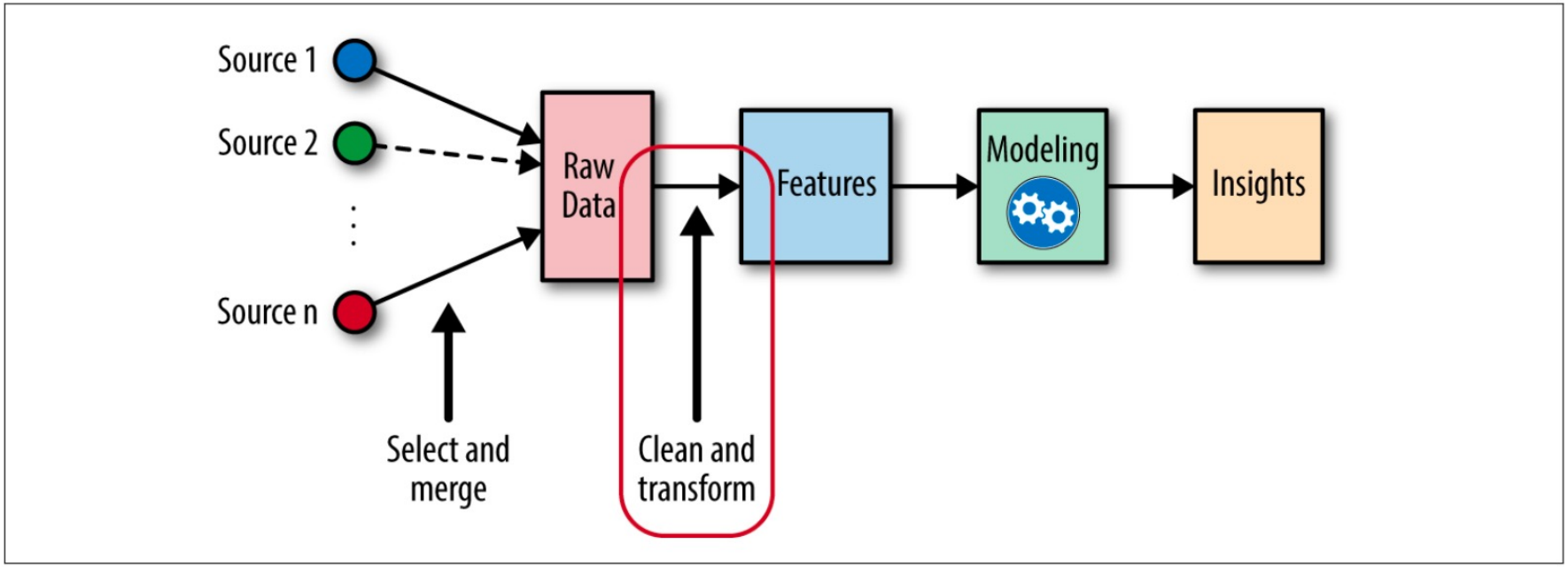
Aprendizaje Supervisado

Feature Engineering (Ingeniería de características)

Idea general

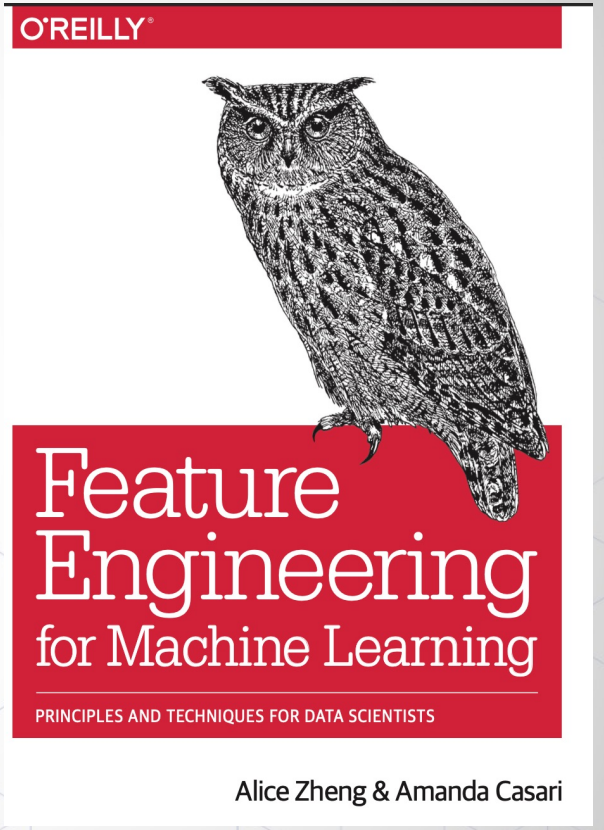
Ajustar y/o transformar las variables independientes (X) para permitir que los modelos descubran mejor las relaciones entre las variables independientes (X) y la respuesta esperada Y

Feature Engineering



Técnicas

- Normalización
- Imputación de valores faltantes
- Reducción de dimensionalidad
- One-hot encoding
- Dummy coding
- Effect coding
- Bag-of-Words
  - Bag-of-n-Grams
- Term frequency–Inverse document frequency



Técnicas

- Normalización
- Imputación de valores faltantes
- Reducción de dimensionalidad

Variables categóricas

- Una variable categórica se emplea para representar categorías o etiquetas
  - las principales ciudades del mundo, las cuatro estaciones del año, o la industria (petróleo, viajes, tecnología) de una empresa
- El número de valores de categoría suele ser finito en un conjunto de datos del mundo real
- En ocasiones los valores se pueden representar numéricamente
  - PERO … los valores de una variable categórica generalmente no se pueden ordenar entre sí. (El petróleo no es ni mayor ni menor que los viajes como tipo de industria)
  - Se llaman no ordinales / nominales

Variables categóricas extensas

- Las variables categóricas extensas son particularmente comunes en los registros transaccionales
  - ID de usuarios en plataformas web
  - La dirección IP de una transacción por Internet
- Son variables categóricas que pueden ser representadas numéricamente ... PERO su magnitud generalmente no es relevante para la tarea en cuestión
  - Por ejemplo, la dirección IP puede ser relevante al realizar la detección de fraudes en transacciones individuales; algunas direcciones IP o subredes pueden generar más transacciones fraudulentas que otras
  - Pero una subred de 164.203.x.x no es intrínsecamente más fraudulenta que 164.202.x.x; el valor numérico de la subred no importa

One-hot encoding

- Requiere la codificación empleando un grupo de bits (Z bits)
- Z = número de categorías … cada bit representa una categoría -> variable en el nuevo dataset
- Las observaciones son codificadas como vectores de longitud Z
- Ejemplo
  - New York, San Francisco, Seattle

	x_1	x_2	x_3
New York	1	0	0
San Francisco	0	1	0
Seattle	0	0	1

Dummy coding

- One-hot coding emplea un bit más de lo necesario para codificar las categorías, i.e. Z “grados de libertad” / dimensiones
- DC emplea el vector de cero para representar una de las categorías, i.e. la categoría de referencia

	x_1	x_2
San Francisco	1	0
Seattle	0	1
New York	0	0

Effect coding

	x_1	x_2
San Francisco	1	0
Seattle	0	1
New York	-1	-1

Código

Wrap up

	Ventajas	Desventajas
One-hot encoding	Cada variable corresponde a una categoría El vector de ceros se puede emplear para datos faltantes	Redundante
Dummy coding	Compacto	No puede manejar datos faltantes
Effect coding	Compacto Puede manejar datos faltantes	El vector de -1 es “pesado” computacionalmente

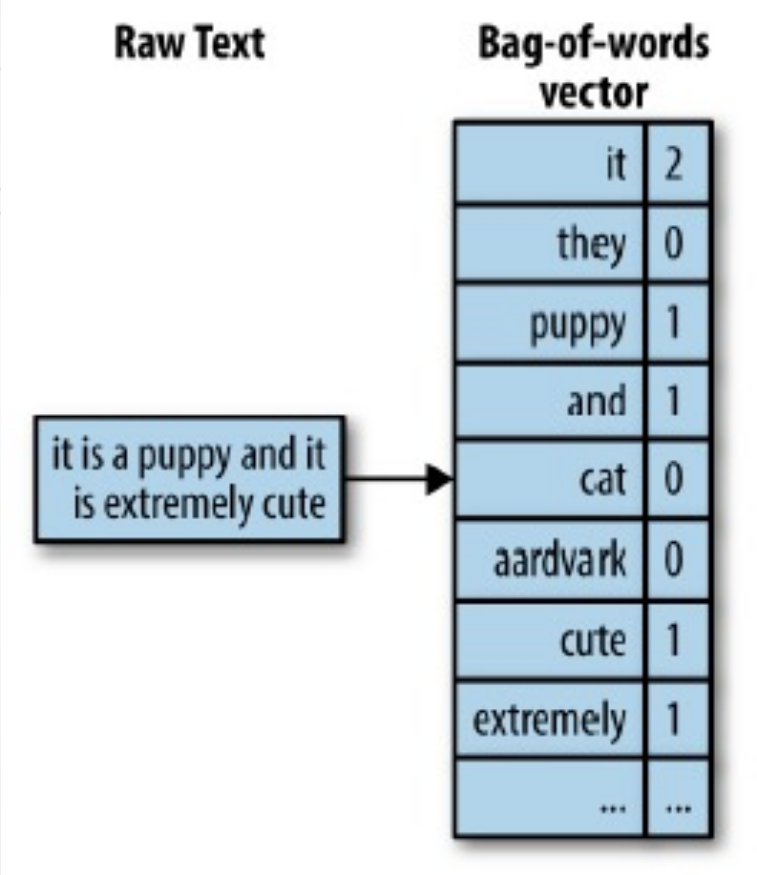
No emplearse con muchas categorías

Texto

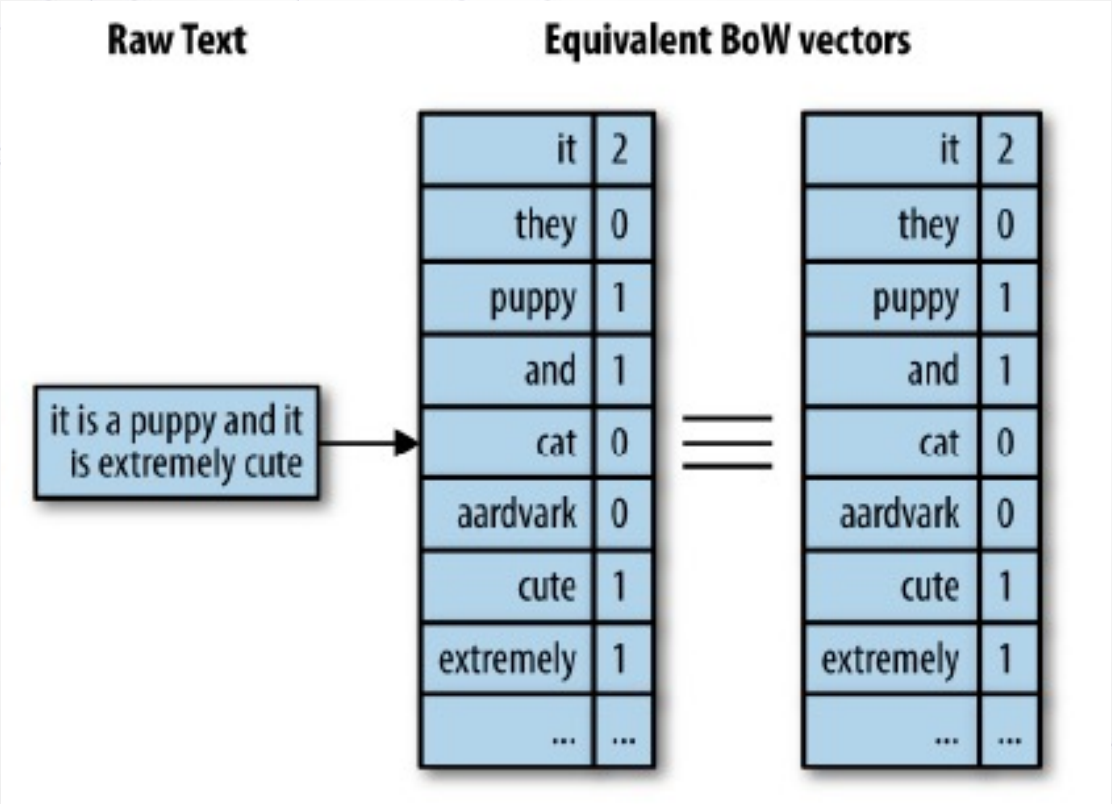
¿Qué harías si estuvieras diseñando un algoritmo para analizar el siguiente párrafo de texto?

Emma knocked on the door. No answer. She knocked again and waited. There was a large maple tree next to the house. Emma looked up the tree and saw a giant raven perched at the treetop. Under the afternoon sun, the raven gleamed magnificently. Its beak was hard and pointed, its claws sharp and strong. It looked regal and imposing. It reigned the tree it stood on. The raven was looking straight at Emma with its beady black eyes. Emma felt slightly intimidated. She took a step back from the door and tentatively said, “Hello?”

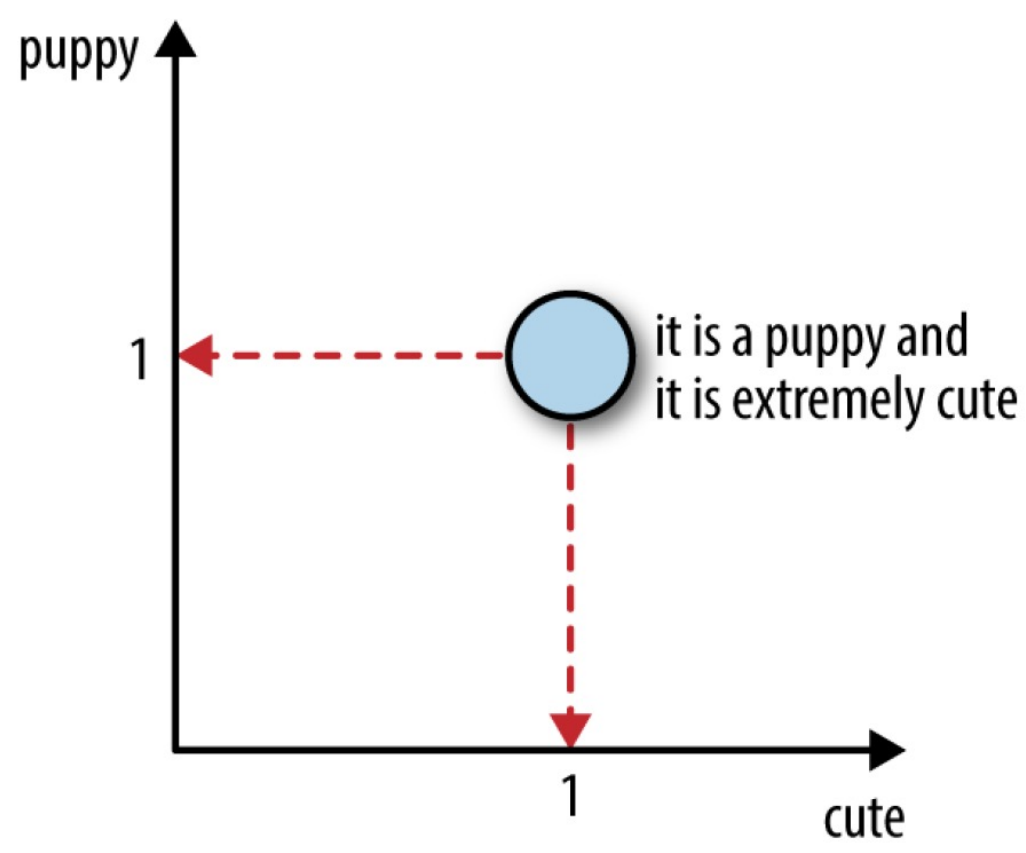
Bag-of-Words


Raw Text
Bag-of-words vector
it is a puppy and it is extremely cute
it 2
they 0
puppy 1
and 1
cat 0
aardvark 0
cute 1
extremely 1
... ...

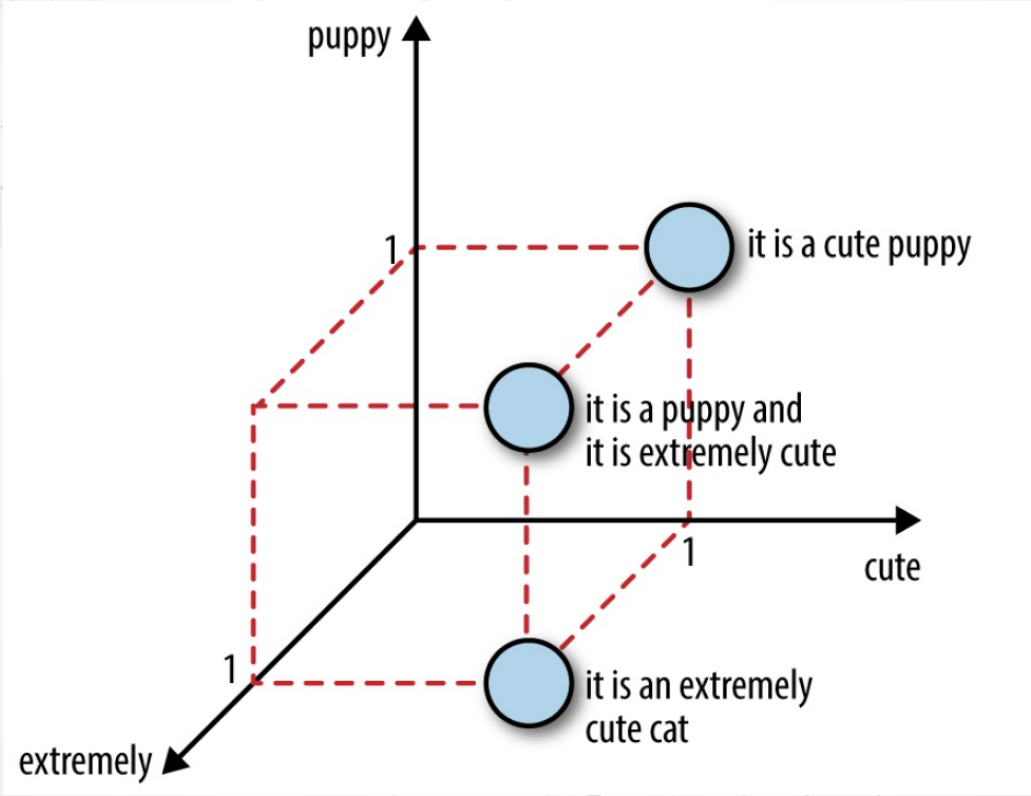
Bag-of-Words


Raw Text
Equivalent BoW vectors
it is a puppy and it is extremely cute
it 2
they 0
puppy 1
and 1
cat 0
aardvark 0
cute 1
extremely 1
... ...
it 2
they 0
puppy 1
and 1
cat 0
aardvark 0
cute 1
extremely 1
... ...

Bag-of-Words



Bag-of-Words


puppy
1
it is a cute puppy
it is a puppy and
it is extremely cute
1
cute
1
it is an extremely
cute cat
extremely

Bag-of-n-Grams

- "Bag-of-n-Grams" o "bolsa de n-gramos" es una extensión natural de la "bolsa de palabras" (bag-of-words).
- Un n-gramo es una secuencia de n tokens.
- Una palabra es esencialmente un 1-gramo, también conocido como un unigrama.
- Ejemplo:
  - “Emma knocked on the door” generates the n-grams:
    - “Emma knocked”, “knocked on”, “on the” and “the door”

Pregunta

- ¿Qué espacio es mayor?
- ¿Bag-of-Words o Bag-of-n-Grams?
- ¿Porqué?

Código

Term frequency–Inverse document frequency

Idea general

- Tf-idf utiliza un recuento normalizado donde cada recuento de palabras se divide por el número de documentos en los que aparece esta palabra
  - $bowN(w,d) =$ número de veces que la palabra w aparece en el documento d (**frecuencia de término**)
  - $tfidf(w,d) = bowN(w,d) \times \frac{N}{nd}$
- N es el número total de documentos en el conjunto de datos
- nd es el número de documentos en los que aparece la palabra w
- $\frac{N}{nd}$ se conoce como **frecuencia inversa de documento**

Idea general

$$tfidf(w,d) = tf(w,d) \times idf(w,d)$$

- Si una palabra aparece en muchos documentos, entonces su frecuencia inversa de documento es cercana a 1

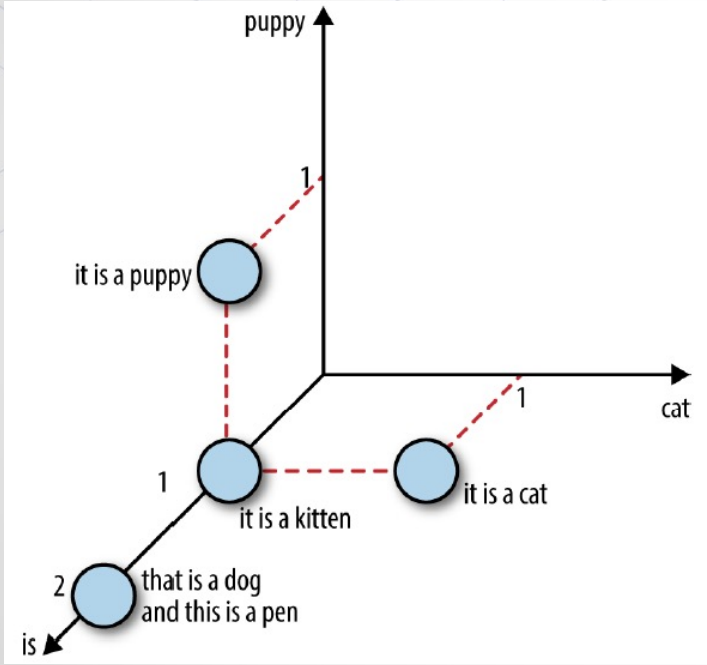
- Si una palabra aparece en solo unos pocos documentos, entonces la frecuencia inversa del documento es mucho más alta

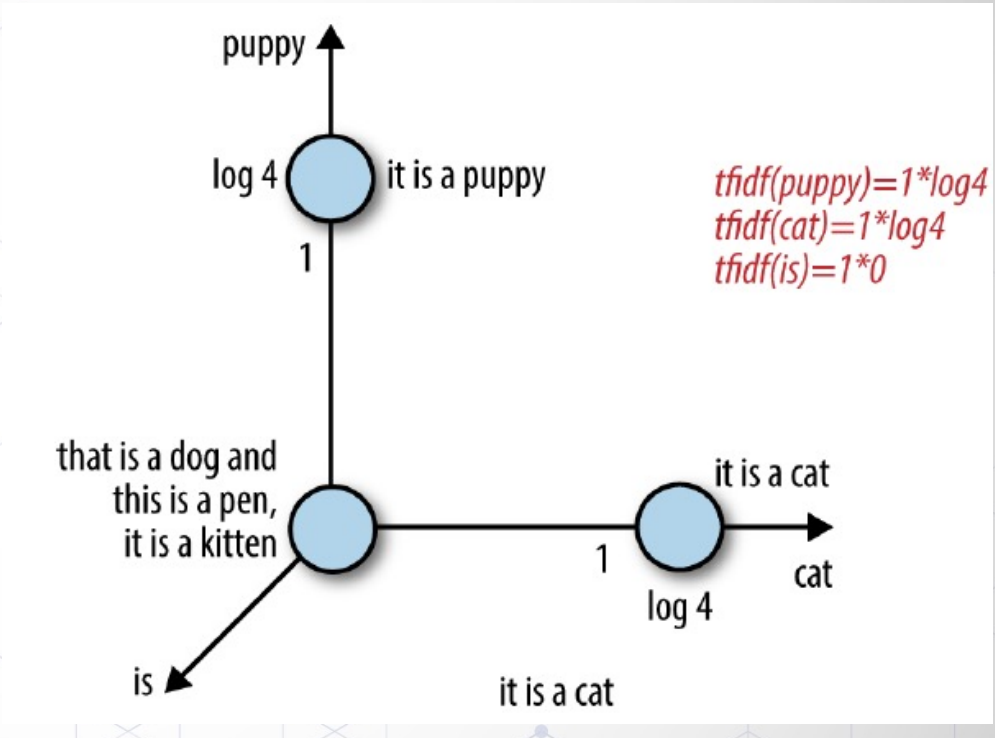
Twist

- Alternativamente, podemos tomar una transformación logarítmica en lugar de usar la frecuencia inversa de documentos cruda.
- El logaritmo convierte 1 en 0 y hace que los números grandes (aquellos mucho mayores que 1) sean más pequeños.
  - $tfidf(w,d) = tf(w,d) \times \log(idf(w,d))$
- Entonces, una palabra que aparece en todos los documentos se anulará, y una palabra que aparece en muy pocos documentos tendrá una cuenta aún más grande que antes

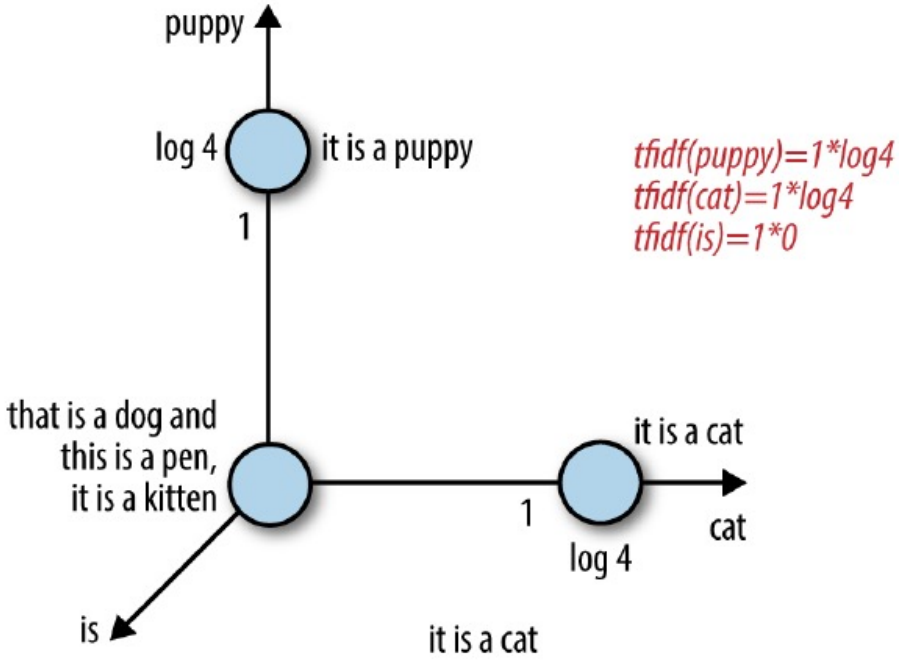
Twist

Cuatro oraciones sobre perros y gatos.





Twist



- hace que las palabras raras sean más prominentes
- efectivamente ignora las palabras comunes

Código

Things to try

- Visualizar matrices X de BoW
  - Entrenamiento
  - Prueba
- Visualizar matrices X de TF-IDF
  - Entrenamiento
  - Prueba

Wrap up

Tf-idf es una operación de columna en la matriz de datos

	it	is	puppy	cat	pen	a	this
it is a puppy	1	1	1	0	0	1	0
it is a kitten	1	1	0	0	0	1	0
it is a cat	1	1	0	1	0	1	0
that is a dog and this is a pen	0	2	0	0	1	2	1
it is a matrix	1	1	0	0	0	1	0

Código

EngineeringX

Founded by the Royal Academy of Engineering
and Lloyd's Register Foundation

GRACIAS

https://hubiq.mx/

HUBIQRO HUBIQ HUBIQRO