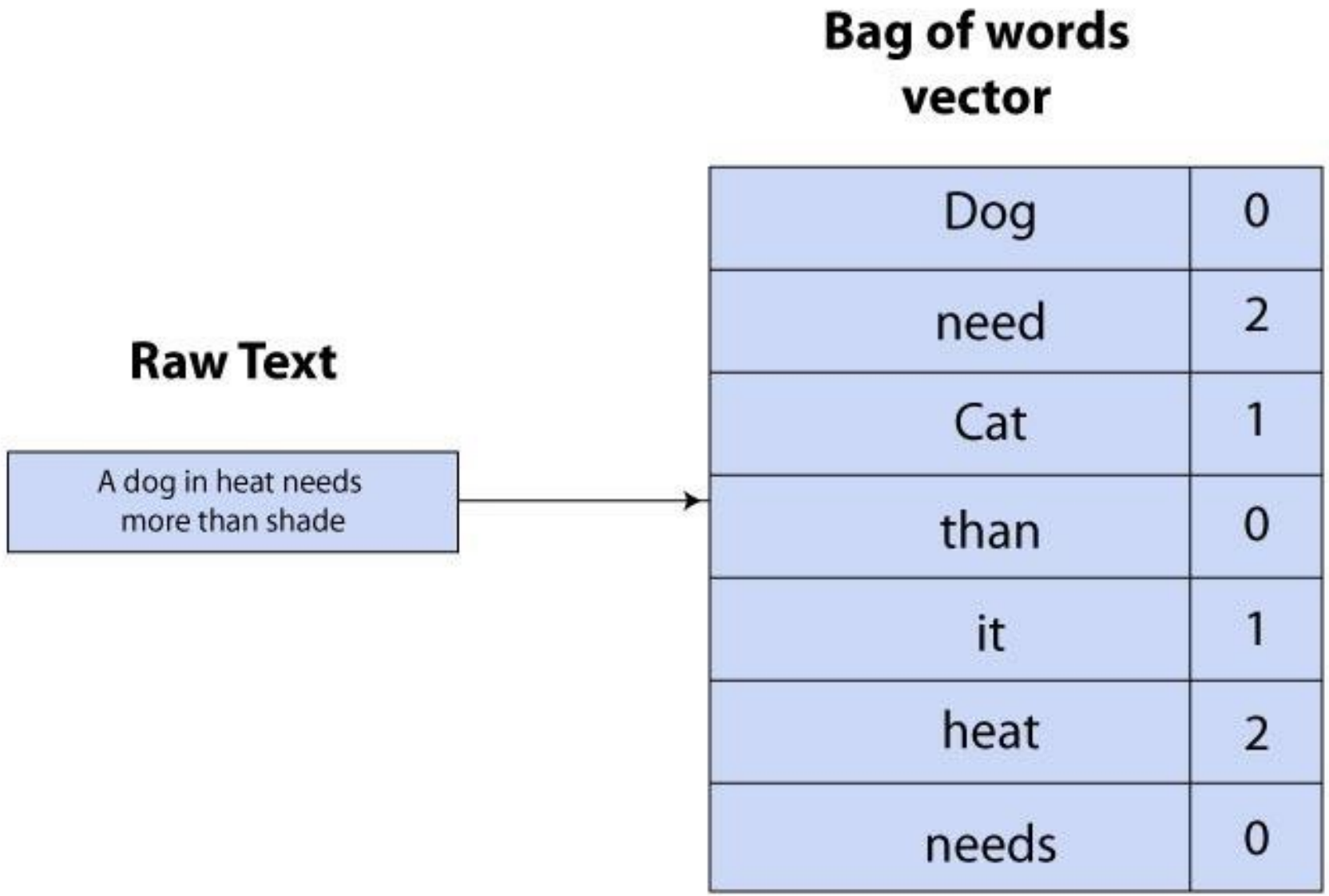
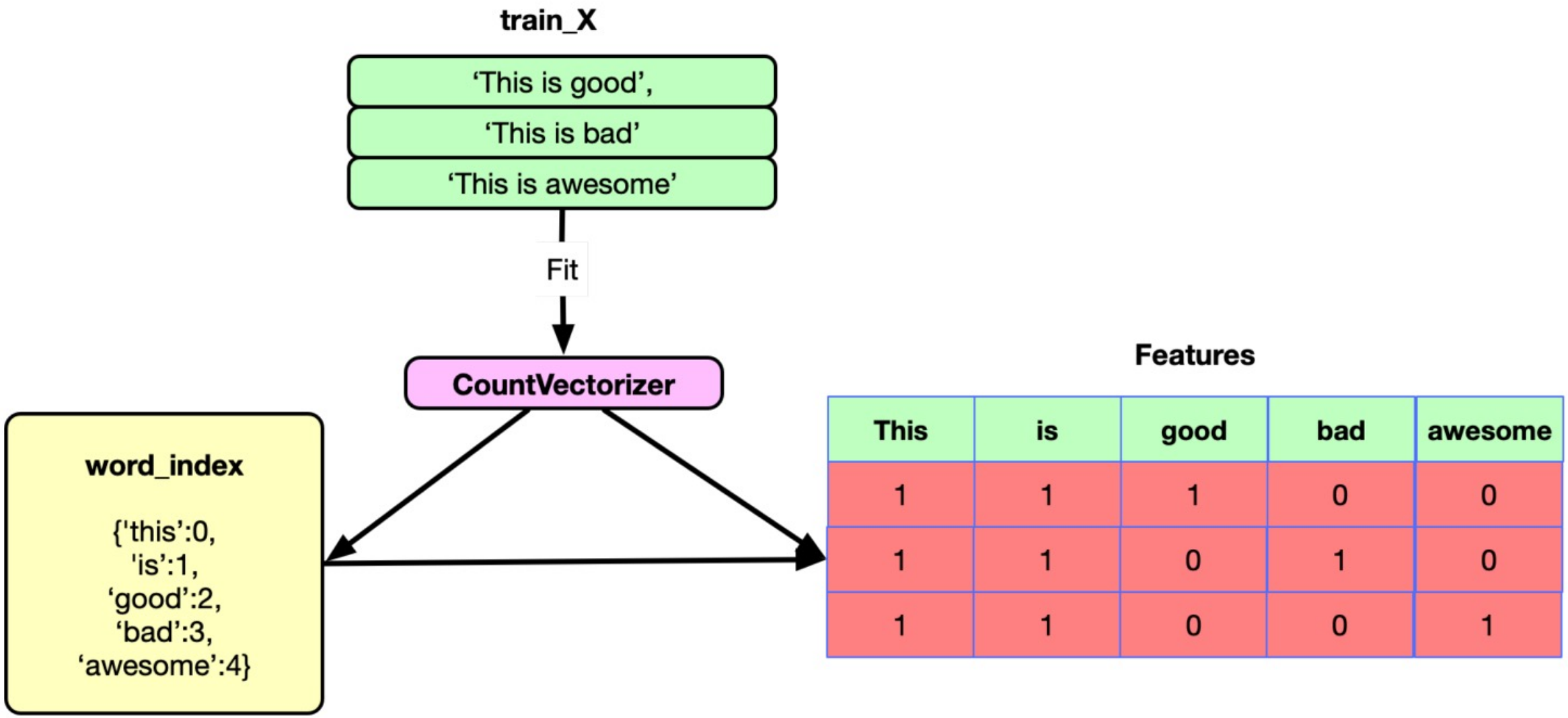
Работа с текстовыми данными


Деепричастие
Прилагательное
Дополнение
Союз
Определение
Сказуемое
Подлежащее
Частица
Наречие
Местоимение
Обстоятельство
Союз
Глагол
Частица
Предлог
Существительное
Глагол
Союз
Причастие
Предлог
Сказуемое
Местоимение
Обстоятельство

Как связаны слово или предложения с алгеброй?

- Bag of words (мешок слов)
- Tf-iDF
- Эмбеддинги (например, word2vec)


Bag of words vector
Raw Text
A dog in heat needs more than shade
Dog 0
need 2
Cat 1
than 0
it 1
heat 2
needs 0


train_X
'This is good',
'This is bad'
'This is awesome'
Fit
CountVectorizer
word_index
{'this':0,
'is':1,
'good':2,
'bad':3,
'awesome':4}
Features
This
is
good
bad
awesome
1
1
1
0
0
1
1
0
1
0
1
1
0
0
1

TF*IDF

TF*IDF=TERM FREQUENCY * INVERSE DOCUMENT FREQUENCY

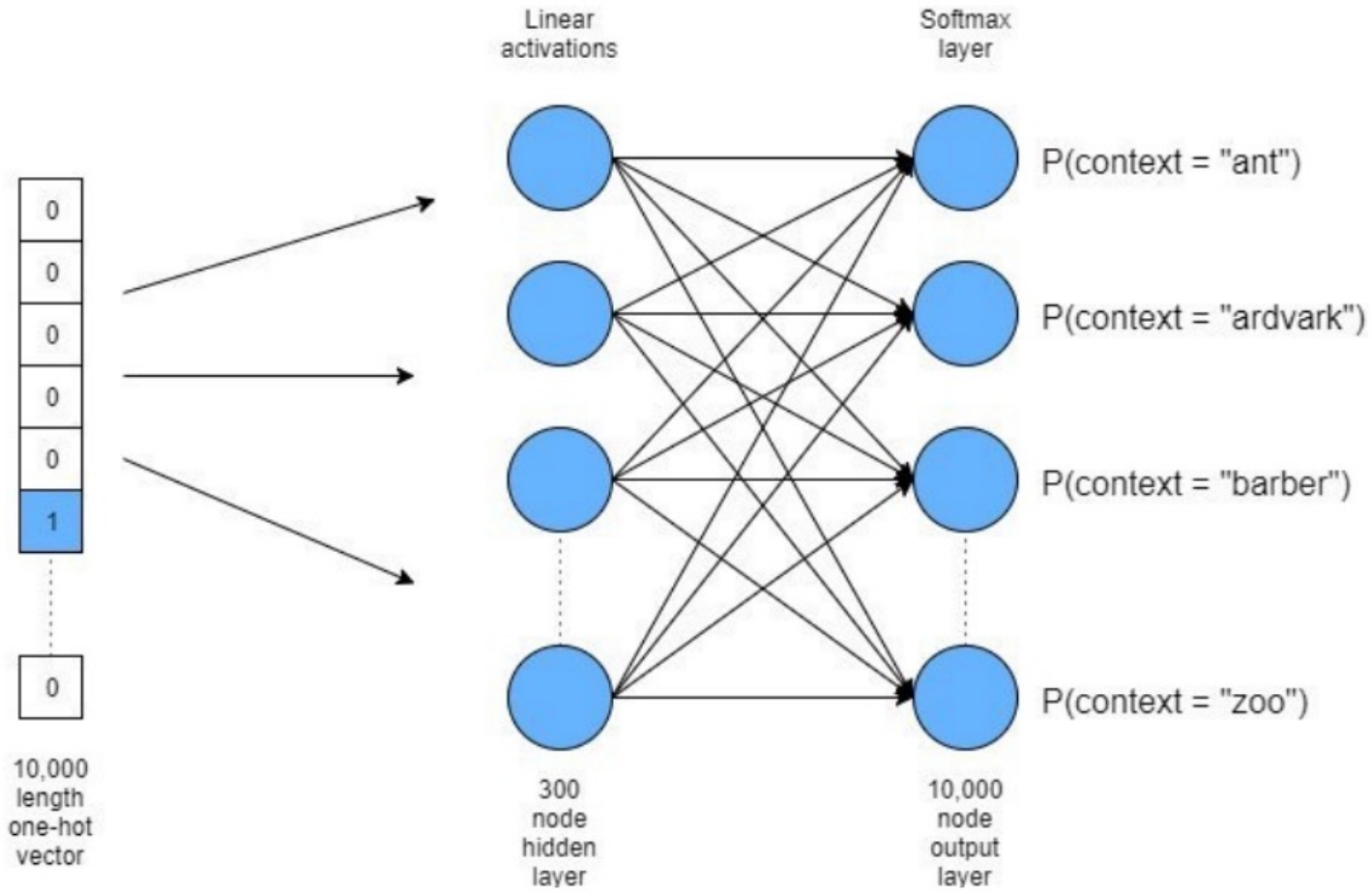
TERM FREQUENCY=

THE AMOUNT OF TIMES A TERM APPEARS IN A DOCUMENT

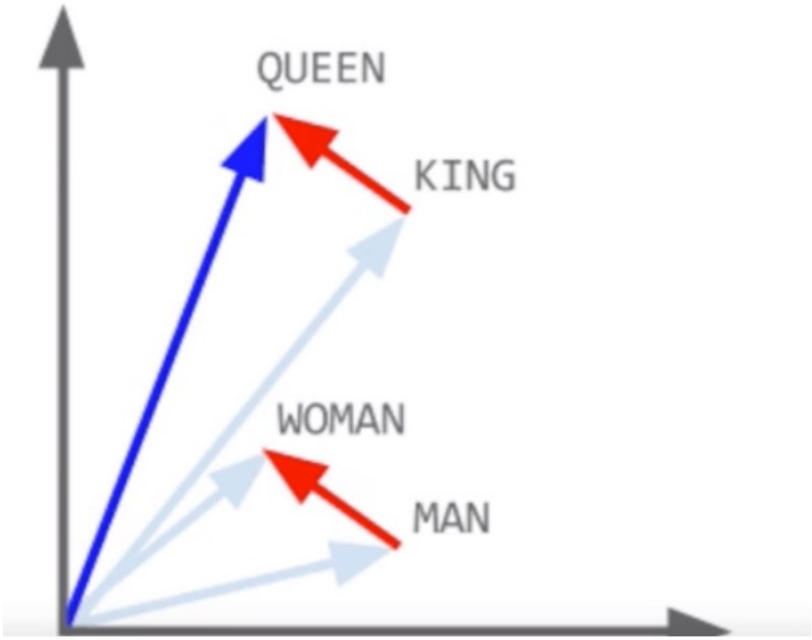
INVERSE DOCUMENT FREQUENCY=

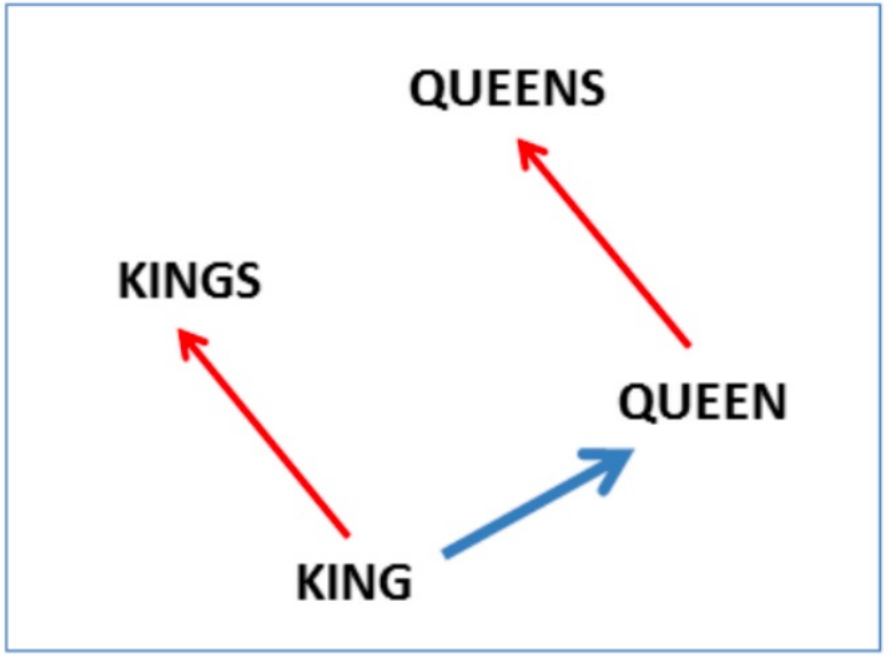
A MEASURE OF WHETHER A TERM IS RARE OR COMMON IN A COLLECTION OF DOCUMENTS.

Word2Vec

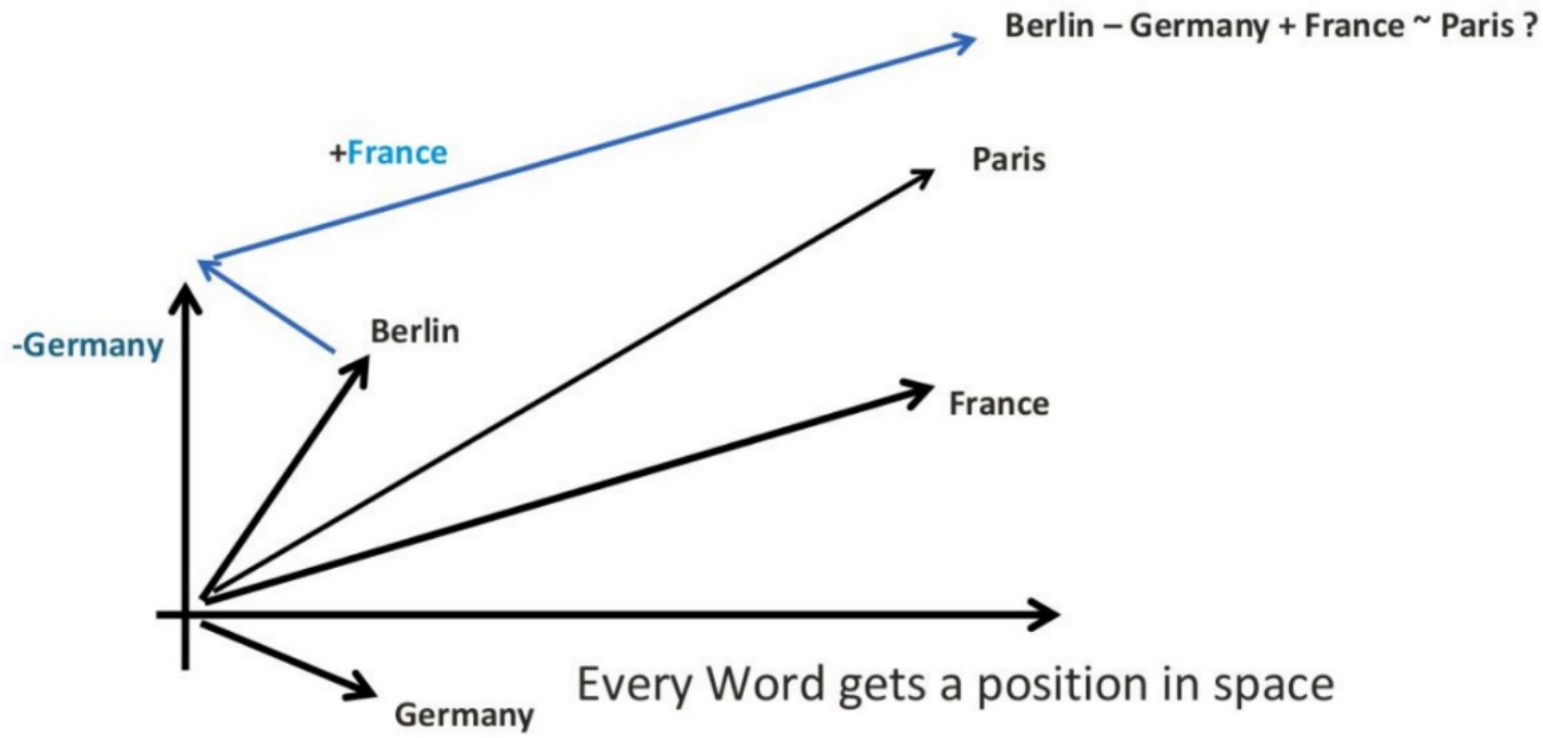

Linear activations
Softmax layer
0
0
0
0
0
1
0
P(context = "ant")
P(context = "ardvark")
P(context = "barber")
P(context = "zoo")
10,000 length one-hot vector
300 node hidden layer
10,000 node output layer

Word2Vec

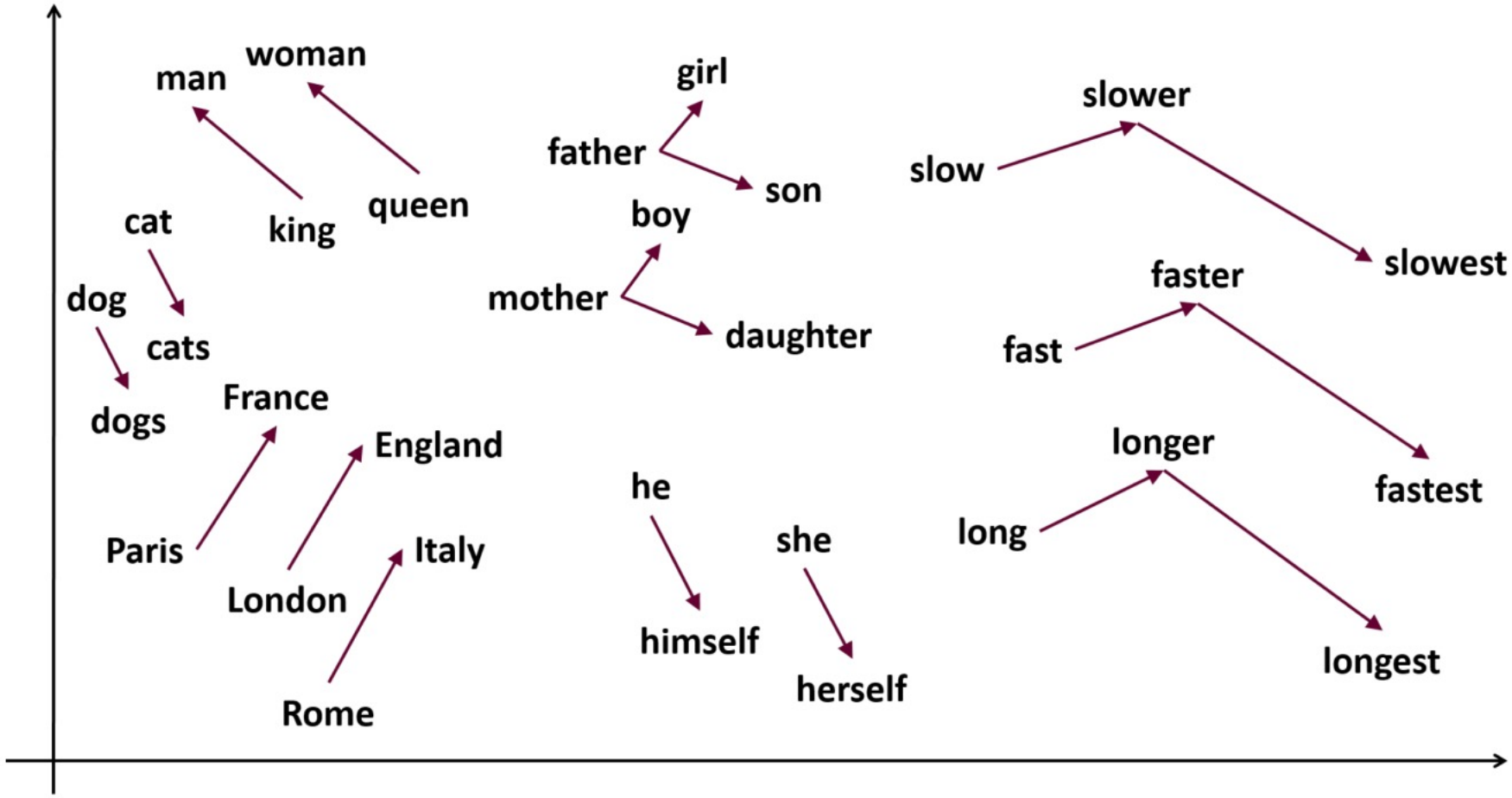




Word2Vec


Berlin – Germany + France ~ Paris ?
+France
Paris
-Germany
Berlin
France
Germany
Every Word gets a position in space

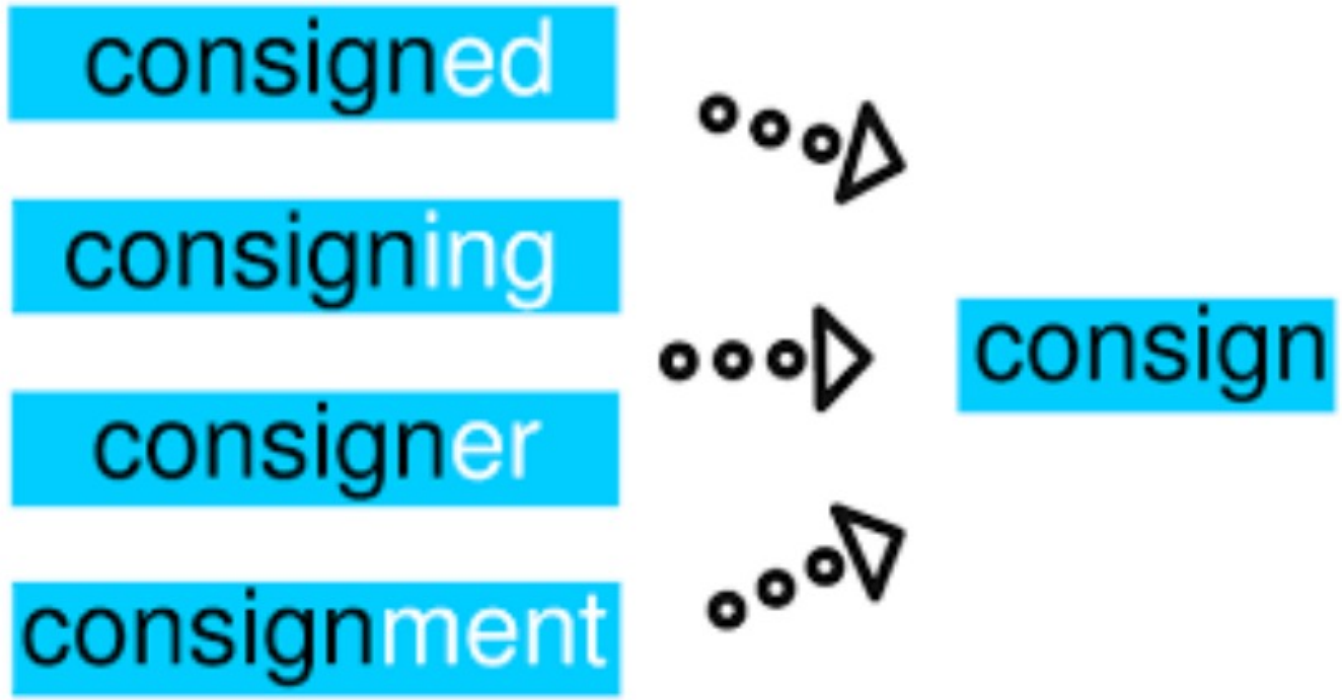
Word2Vec


man
woman
king
queen
girl
father
son
boy
mother
daughter
slow
slower
slowest
cat
cats
dog
dogs
fast
faster
fastest
France
England
Paris
London
Italy
Rome
he
himself
she
herself
long
longer
longest

Стемминг

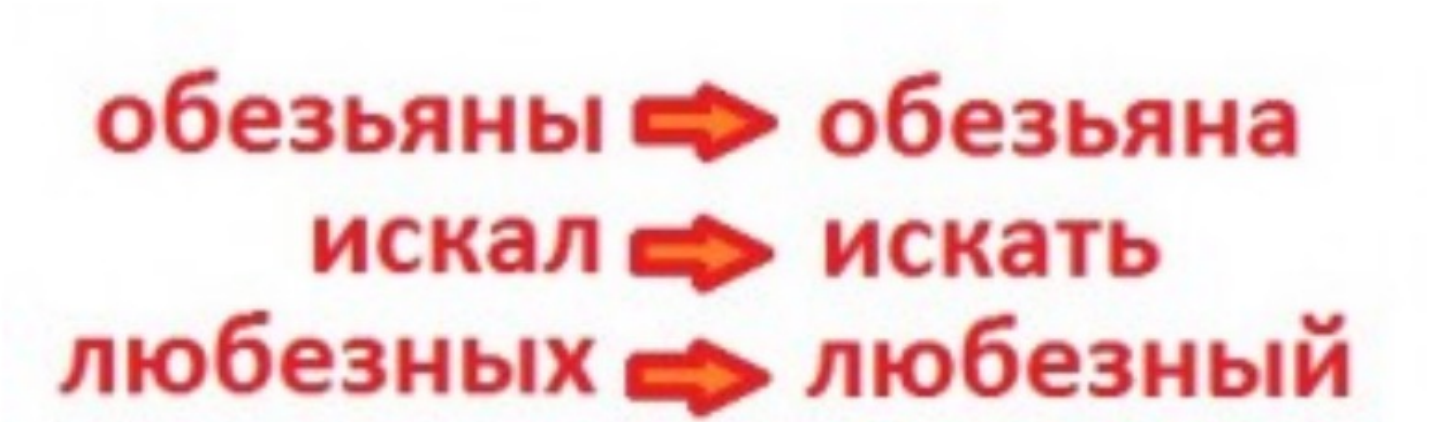

поддержк
основ
рынк
интернет
сфер
отношен
собствен
местн
приня
жил
интерес
льгот
нов
независим
образован
зарплат
кажд
наш
власт
форм
мест
сист
сем
ограничен
содействова
реформ
гражда

Стемминг



Лемматизация

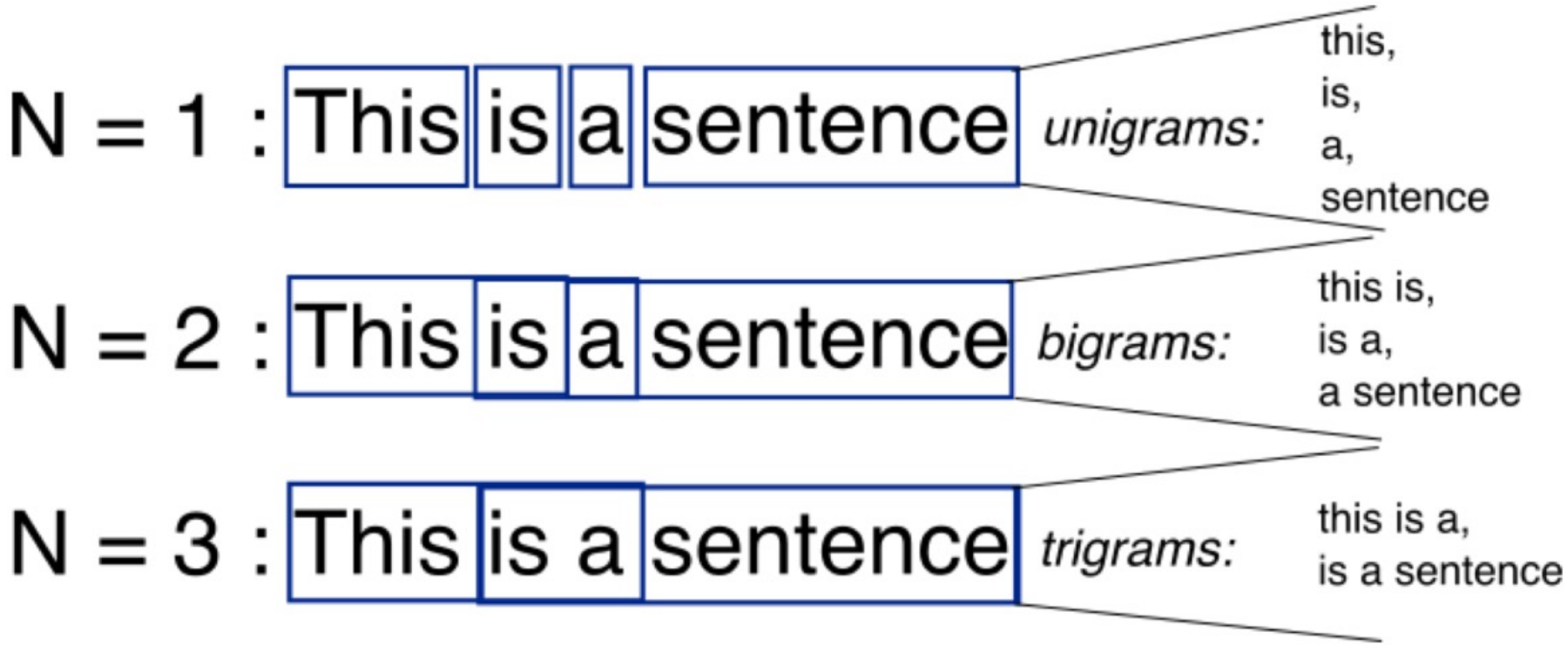
Лемматизация



СТОП-слова!

- Союзы
- Предлоги
- Частицы
- Высокочастотные слова
- И т.д. и т.п.

N-gramms


N = 1 : This is a sentence
unigrams:
this,
is,
a,
sentence
N = 2 : This is a sentence
bigrams:
this is,
is a,
a sentence
N = 3 : This is a sentence
trigrams:
this is a,
is a sentence