Блок

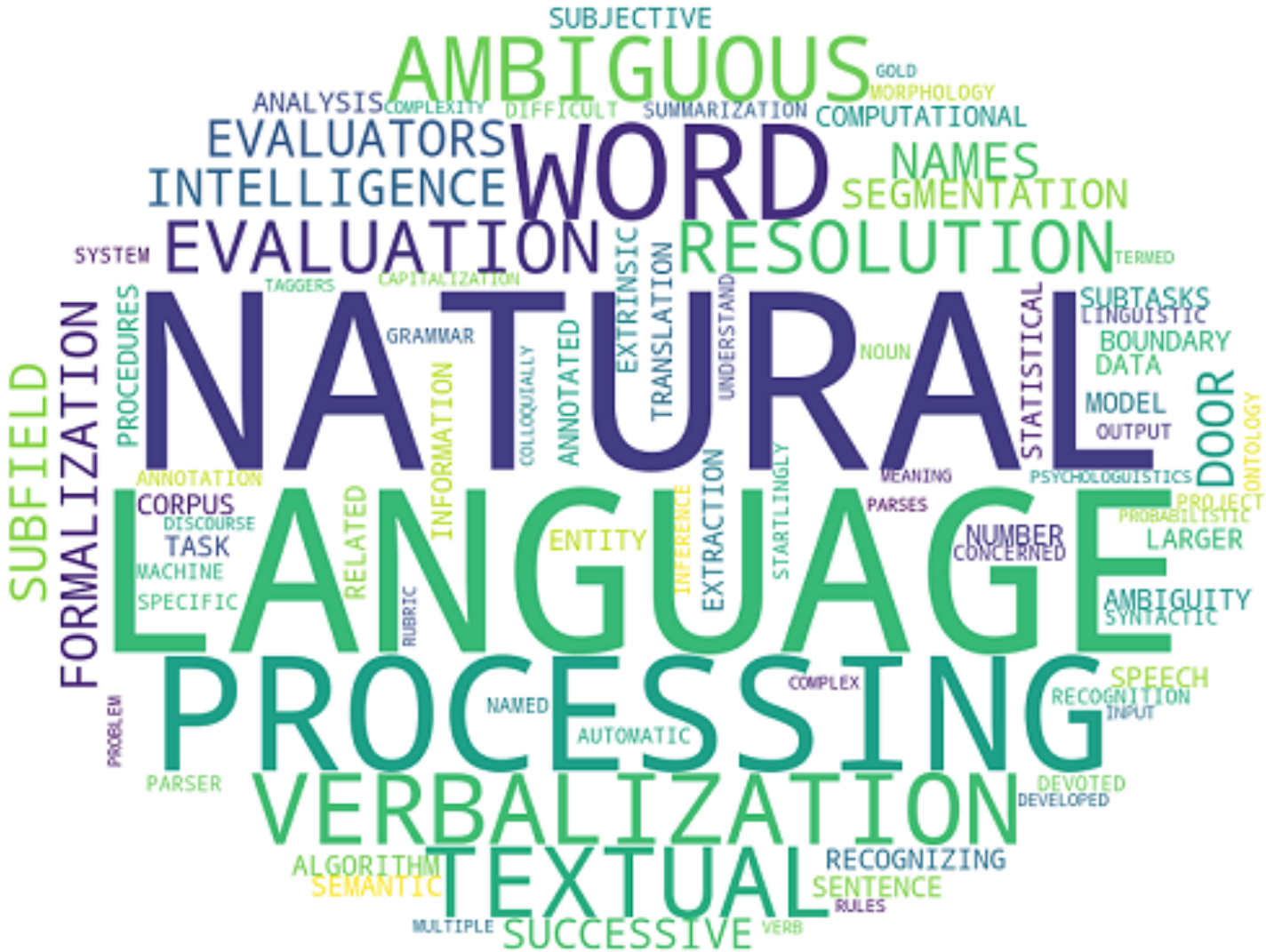
Feature Engineering

Занятие № 6

Работа с текстовыми данными

Цели занятия

- Посмотреть на текст с другой стороны
- Узнаем, как текст превращается в математику


SUBJECTIVE
AMBIGUOUS
GOLD
MORPHOLOGY
ANALYSIS
COMPLEXITY
DIFFICULT
SUMMARIZATION
COMPUTATIONAL
EVALUATORS
WORD
NAMES
INTELLIGENCE
SEGMENTATION
EVALUATION
RESOLUTION
SYSTEM
TERMED
TAGGERS
CAPITALIZATION
NATURAL
SUBTASKS
LINGUISTIC
GRAMMAR
BOUNDARY
PROCEDURES
EXTRINSIC
TRANSLATION
UNDERSTAND
NOUN
DATA
STATISTICAL
SUBFIELD
FORMALIZATION
INFORMATION
ANNOTATED
COLLOQUIALLY
MODEL
DOOR
OUTPUT
ANNOTATION
MEANING
PSYCHOLOGUISTICS
ONTOLOGY
LANGUAGE
CORPUS
PARSES
PROJECT
DISCOURSE
PROBABILISTIC
RELATED
ENTITY
NUMBER
LARGER
TASK
STARTLINGLY
CONCERNED
INFERENCE
EXTRACTION
MACHINE
AMBIGUITY
SPECIFIC
RUBRIC
SYNTACTIC
SPEECH
PROCESSING
COMPLEX
RECOGNITION
NAMED
INPUT
PROBLEM
AUTOMATIC
PARSER
DEVOTED
VERBALIZATION
DEVELOPED
RECOGNIZING
ALGORITHM
SEMANTIC
SENTENCE
TEXTUAL
RULES
MULTIPLE
VERB
SUCCESSIVE

Сколько слов в русском языке?

Сколько слов в русском языке?

Число лемм: ~500k

Число всех слов: >4.5 млн.

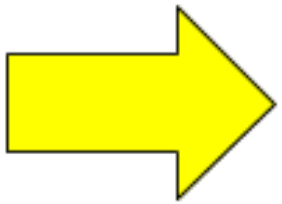
Слова и математика

Слова и математика

- Bag of words
- Tf-idf
- Word2vec/Glove

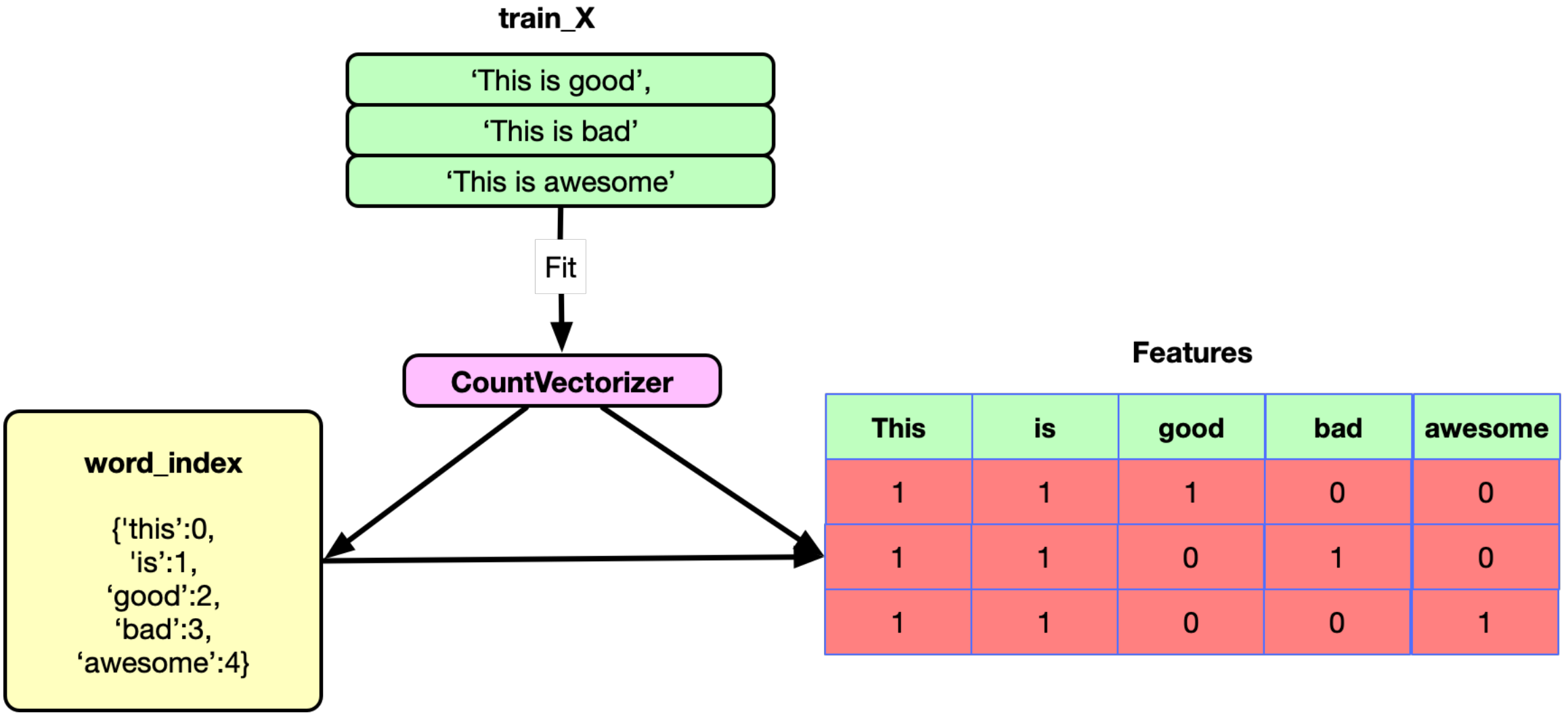
Bag of words

Color
Red
Red
Yellow
Green
Yellow



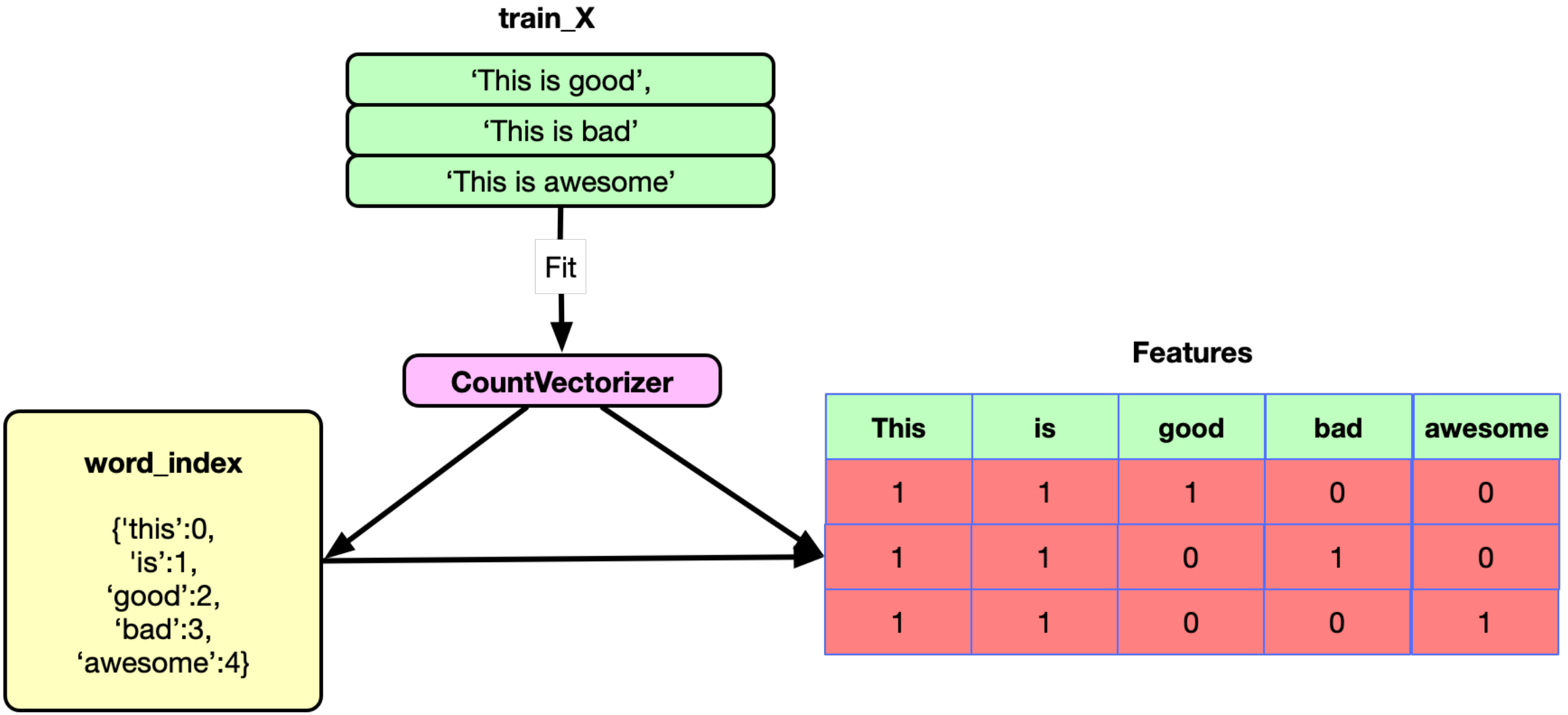
Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

Bag of words



This	is	good	bad	awesome
1	1	1	0	0
1	1	0	1	0
1	1	0	0	1

Bag of words



This	is	good	bad	awesome
1	1	1	0	0
1	1	0	1	0
1	1	0	0	1

Tf-iDF

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

Tf-iDF

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

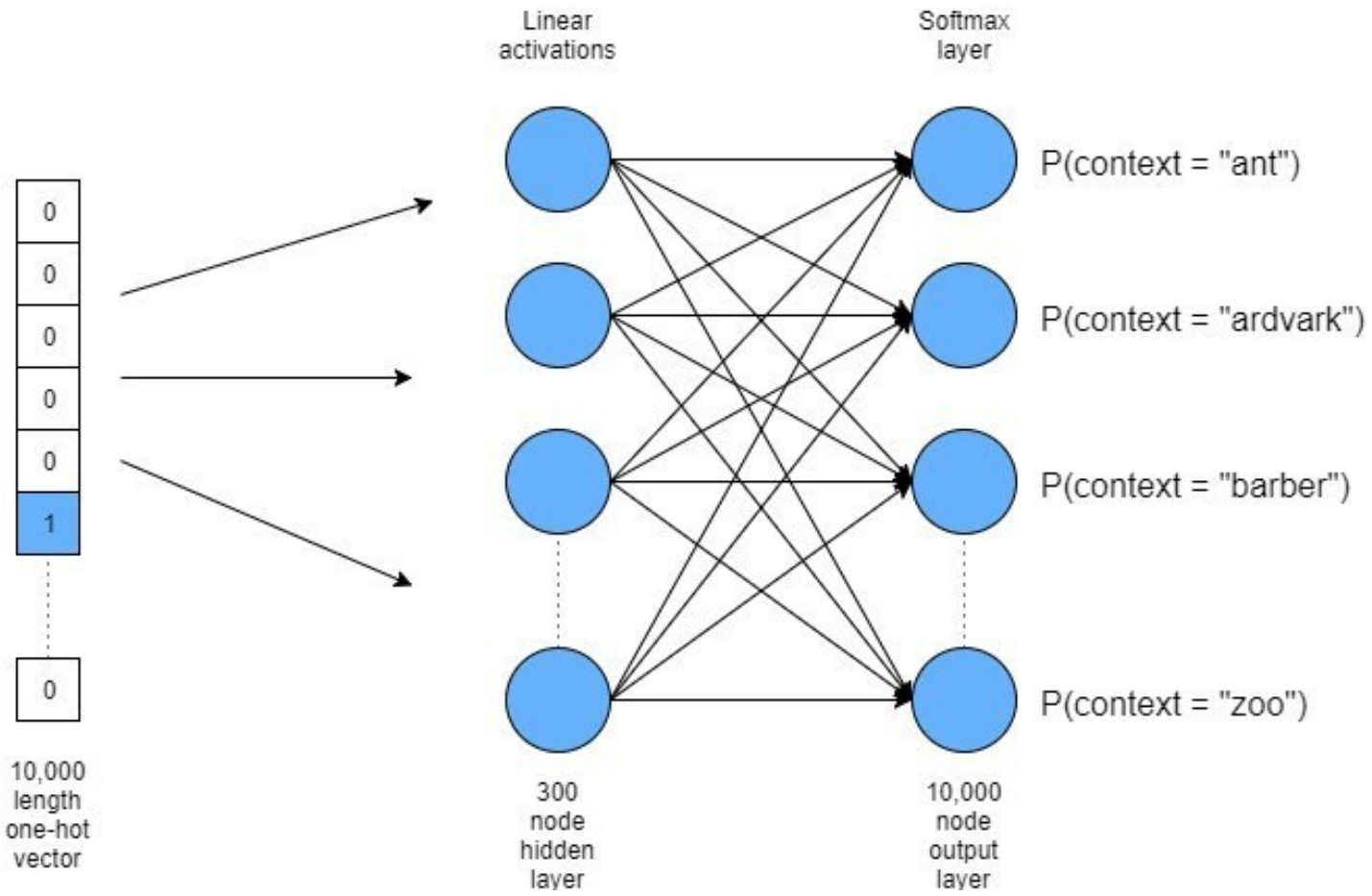
Term x within document y

$tf_{x,y}$ = frequency of x in y

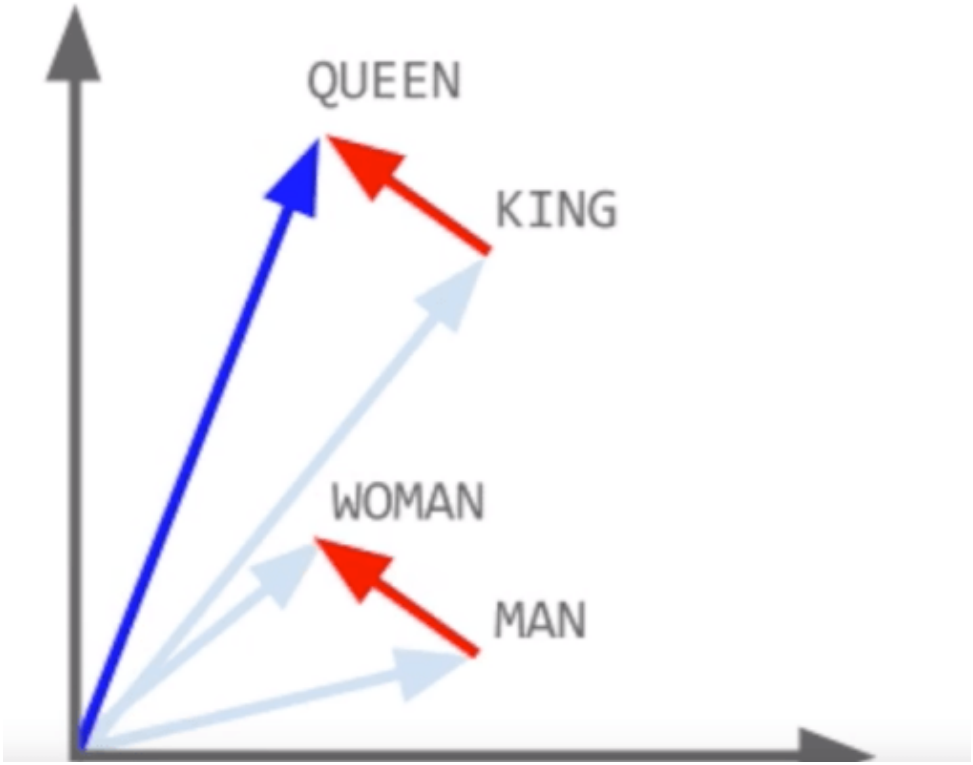
df_x = number of documents containing x

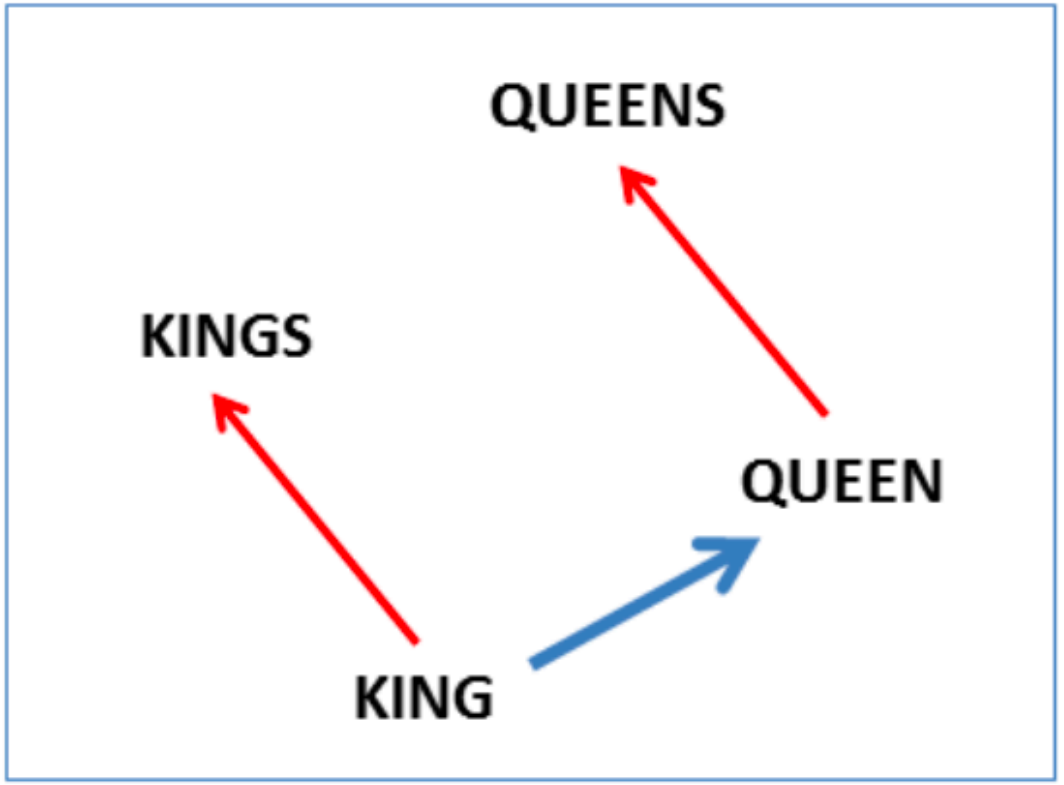
N = total number of documents

Word2vec

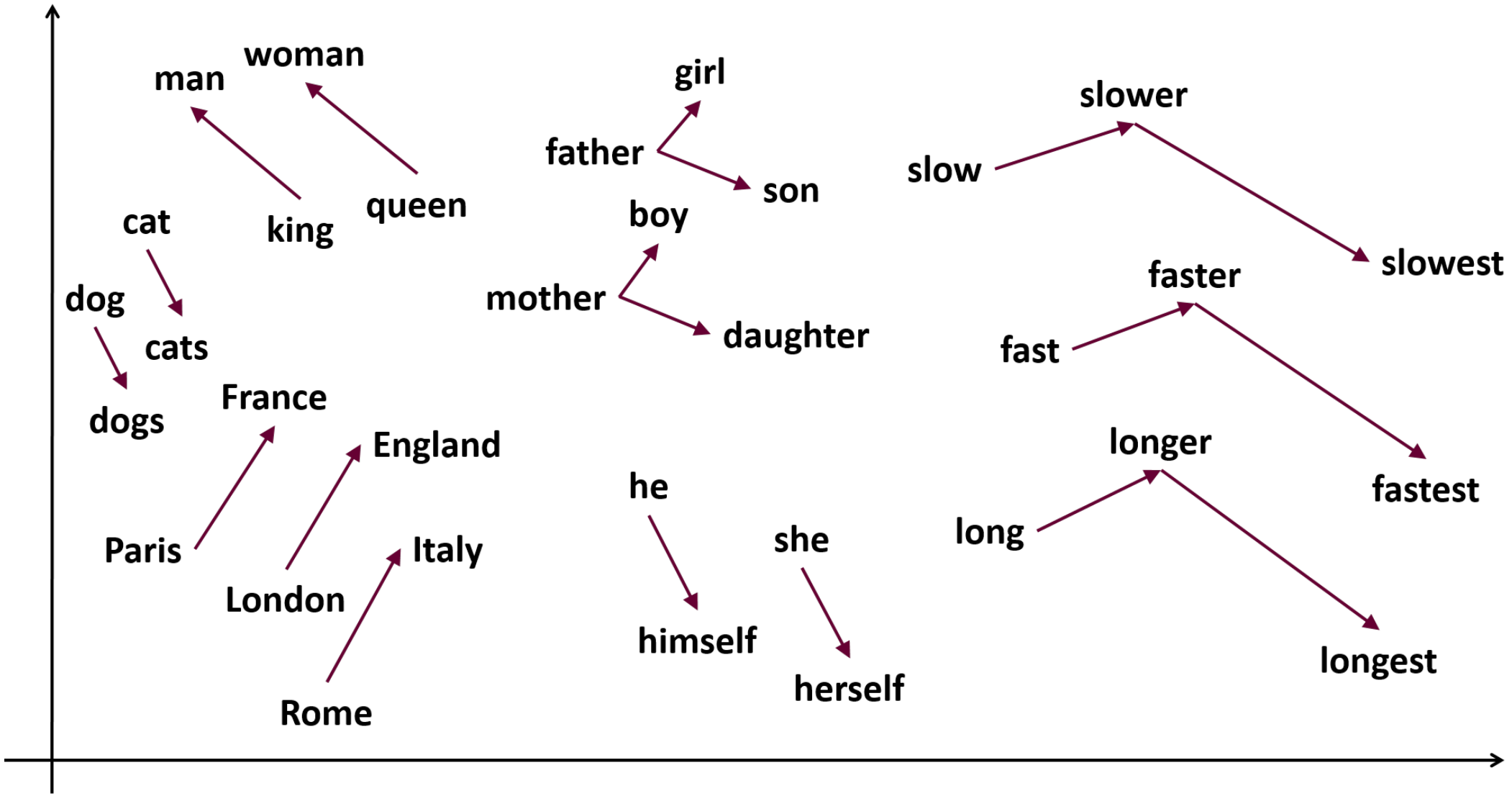

Linear activations
Softmax layer
0
0
0
0
0
1
0
P(context = "ant")
P(context = "ardvark")
P(context = "barber")
P(context = "zoo")
10,000 length one-hot vector
300 node hidden layer
10,000 node output layer

Word2vec

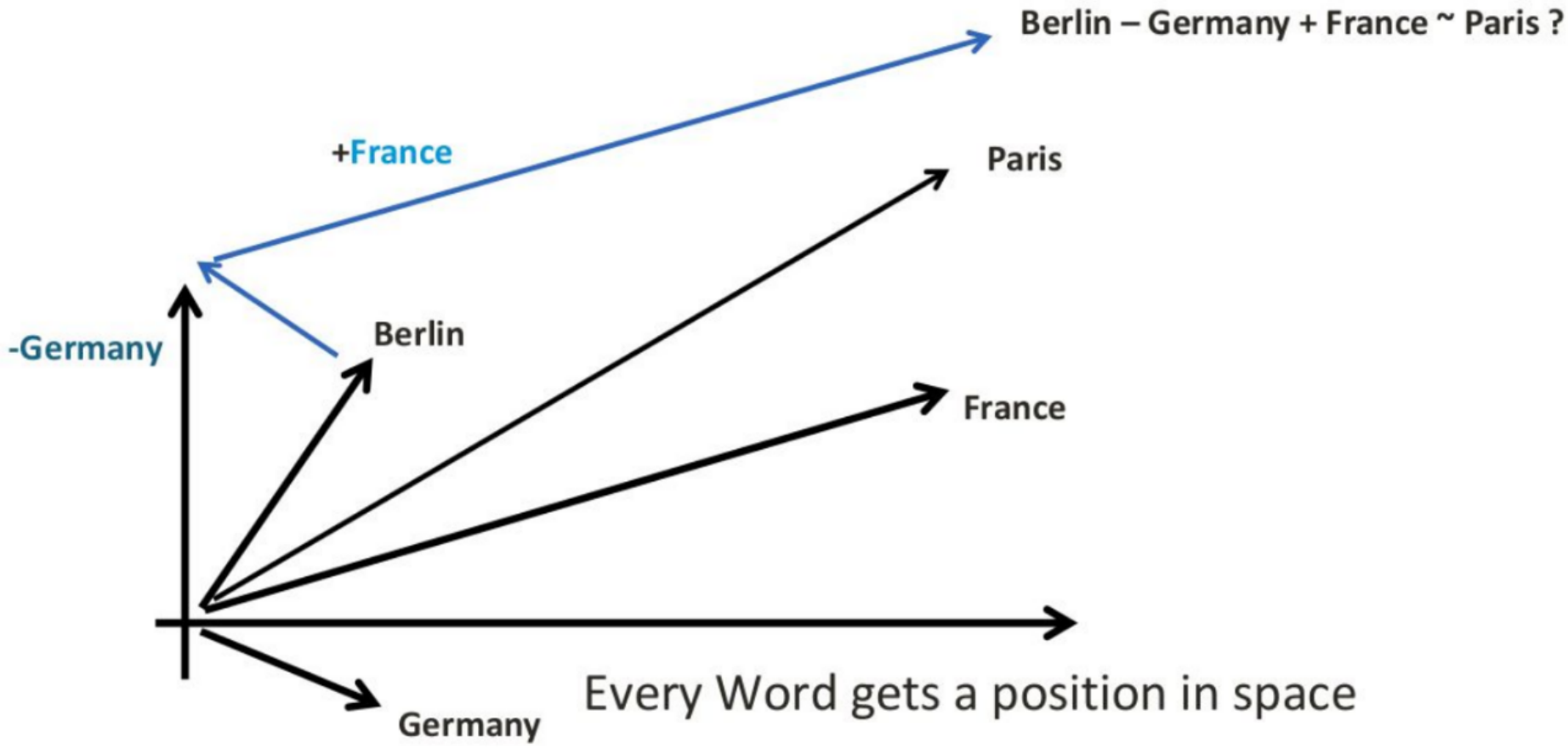




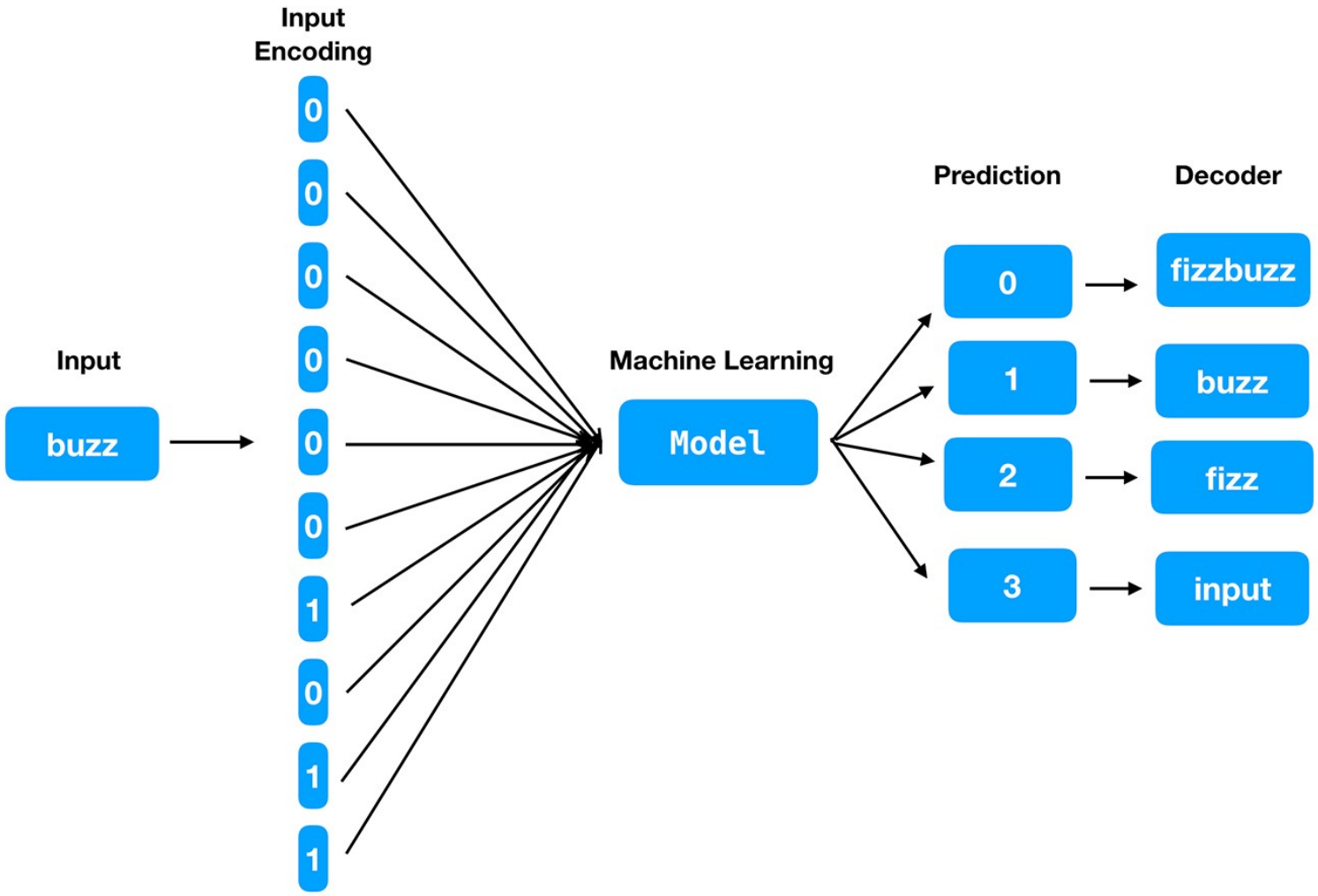
Word2vec


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

Word2vec


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

Char2vec

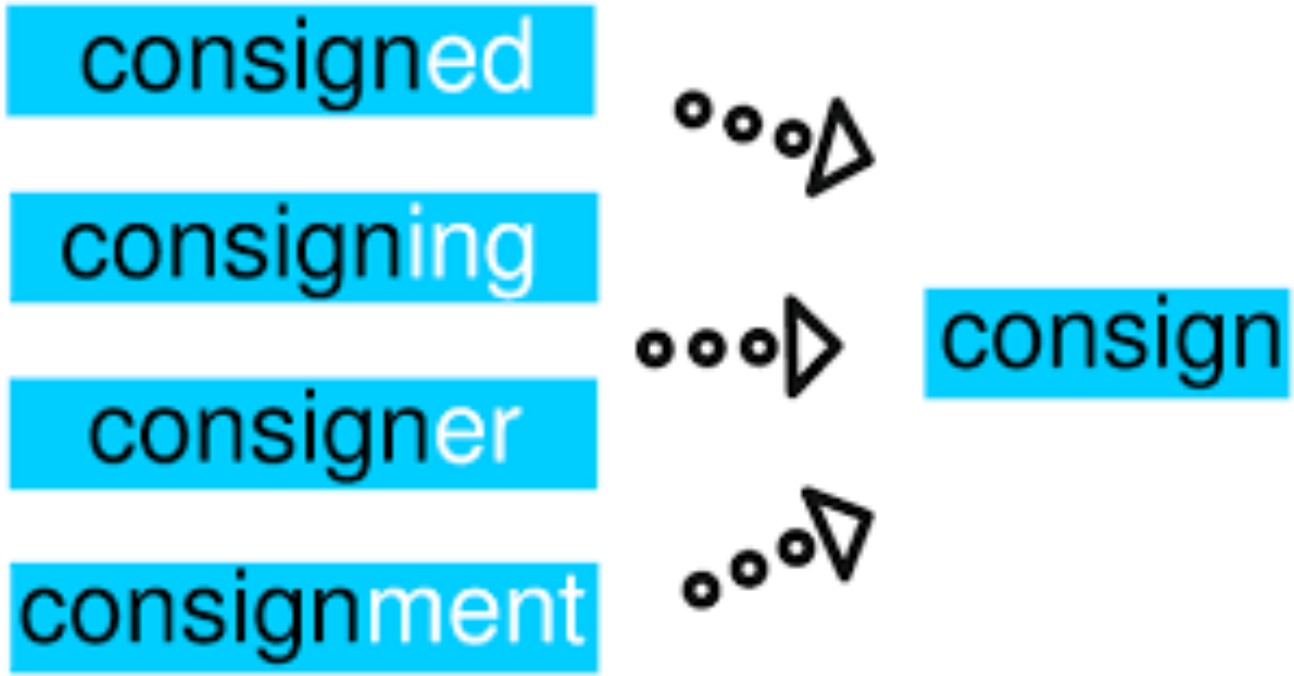

Input
Encoding
Input
buzz
0
0
0
0
0
1
0
1
1
Machine Learning
Model
Prediction
Decoder
0
fizzbuzz
1
buzz
2
fizz
3
input

Слов много, куда девать?

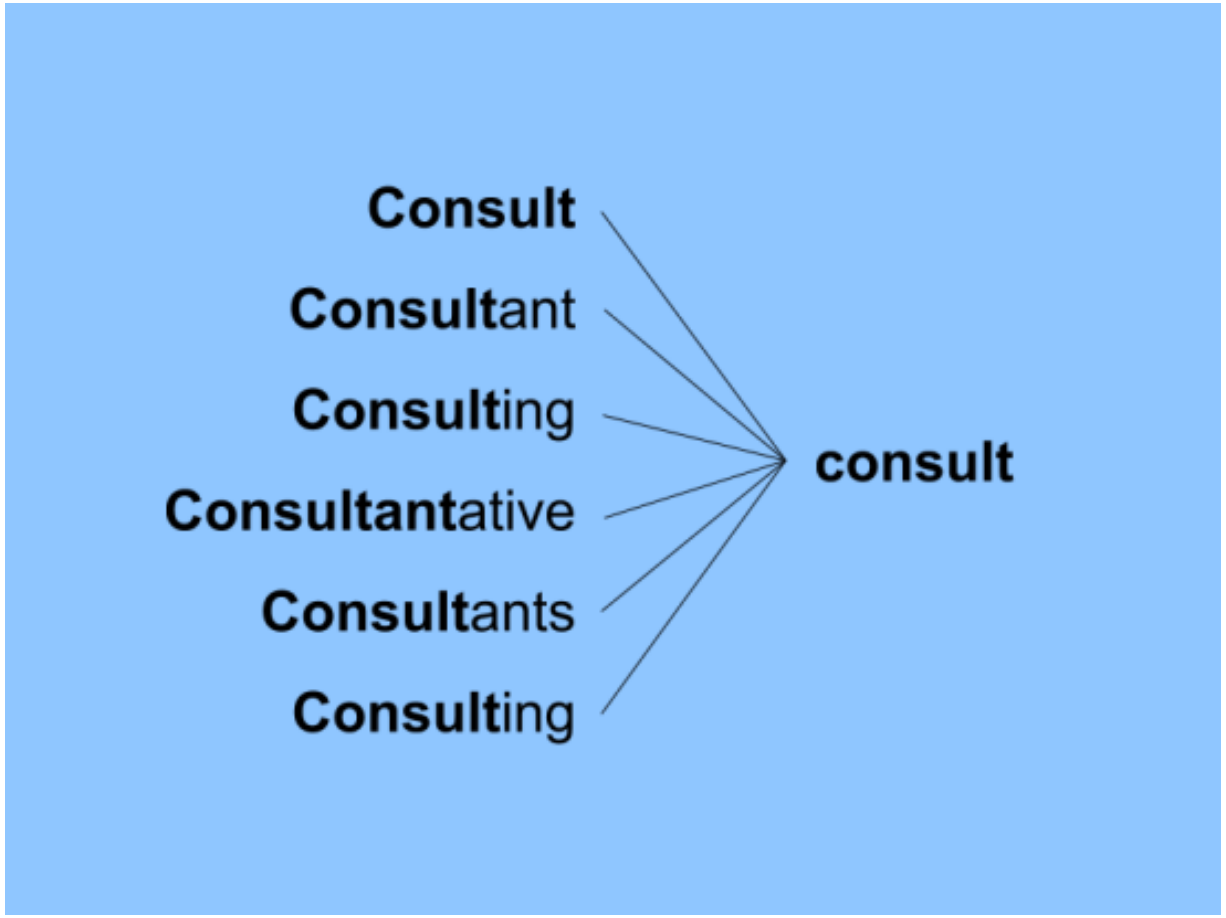
Стемминг


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

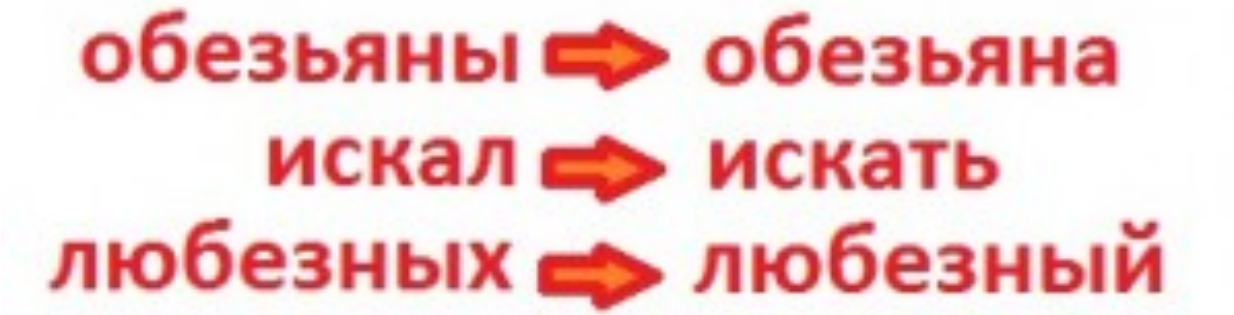
Стемминг



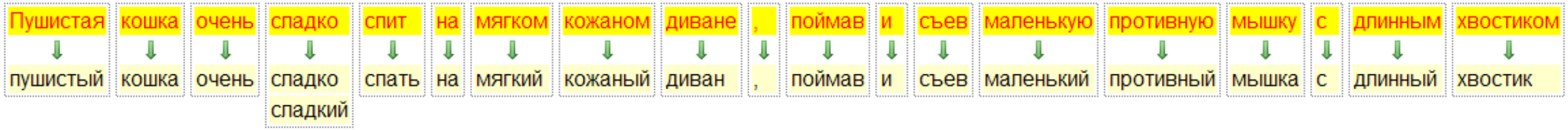
Стемминг


Consult
Consultant
Consulting
Consultantative
Consultants
Consulting
consult

Лемматизация



Лемматизация

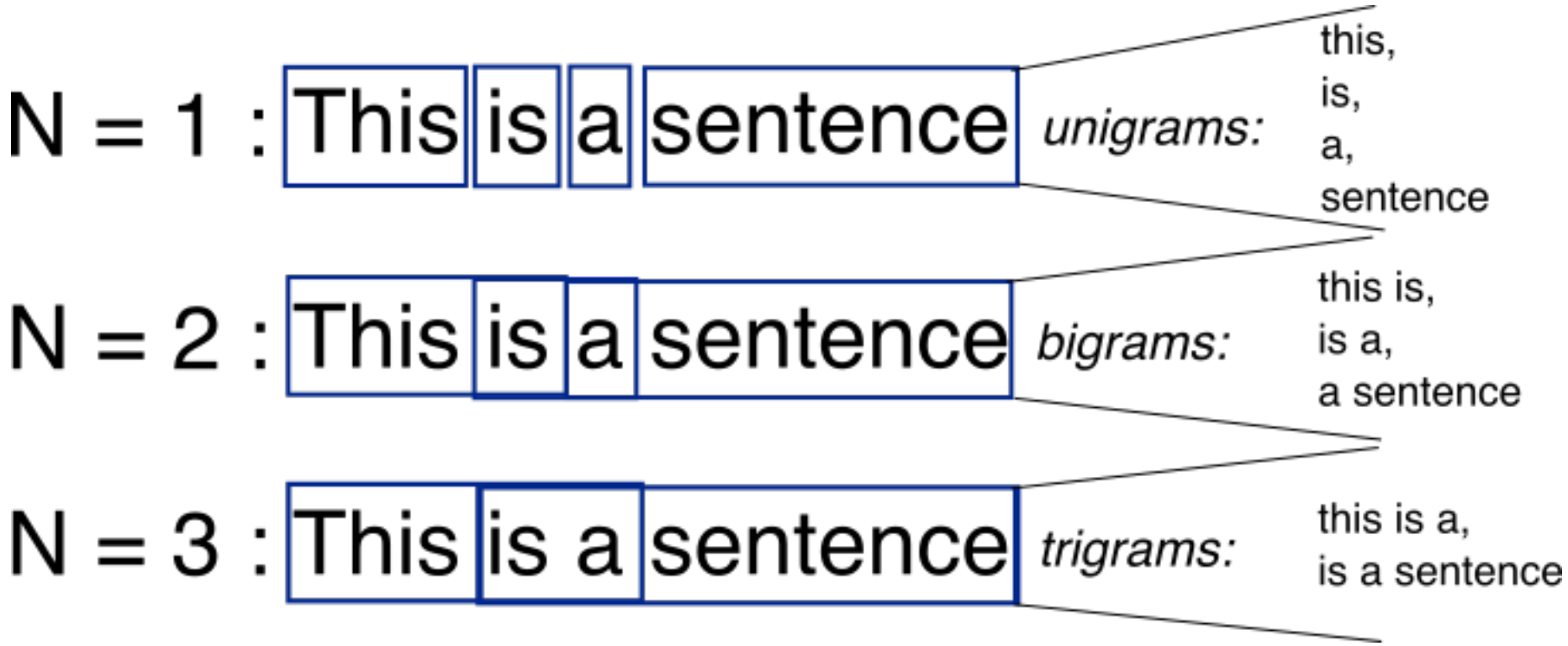


СТОП-слова

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

А как можно определить взаимодействие слов?

N-gram


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

N-gram


The quick brown fox jumped over the lazy dog .
The quick brown fox jumped over the lazy dog .
The quick brown fox jumped over the lazy dog .
The quick brown fox jumped over the lazy dog .
The quick brown fox jumped over the lazy dog .
The quick brown fox jumped over the lazy dog .

ВОПРОСЫ