Teach Machine To Teach

Igor Dziuba

/igorock/teach-machine-to-teach

About Me


KYIV

Career


Playtika


Store
Translate


epam


softheme


iPay.ua


LUXOFT


playtech


DEVELLAR

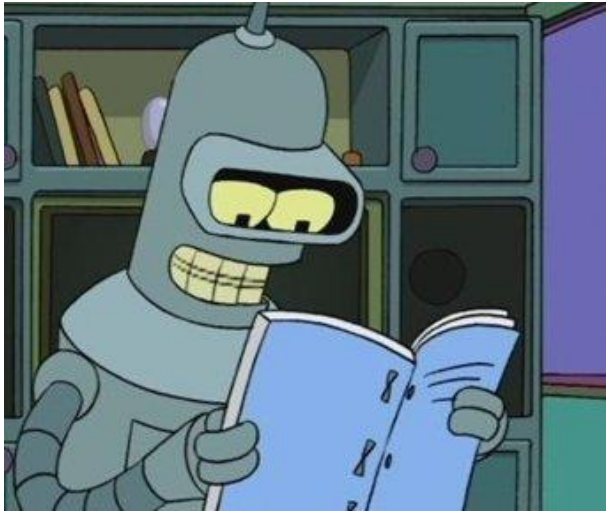
Technology Stack

Agenda

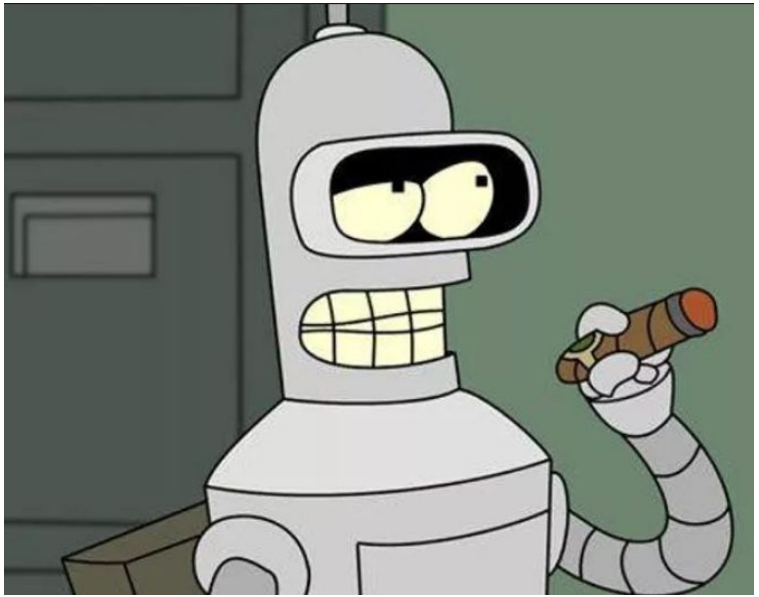
- What is Machine Learning
- What is Store Translate
- How it applies Machine Learning
- Text Classification with Logistic Regression
- Implementation with Spark ML
- Recommendations with Collaborative Filtering
- Implementation with Spark ML

Machine Learning

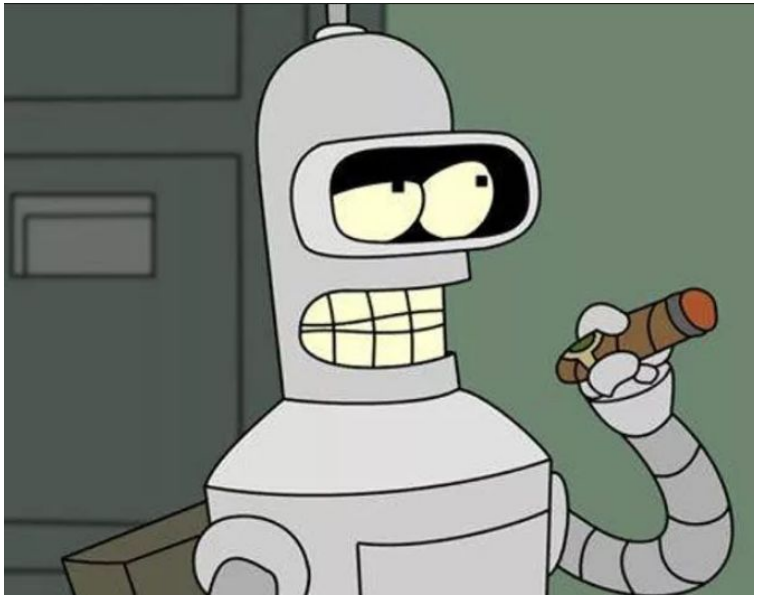
Science of getting computers to act without being explicitly programmed



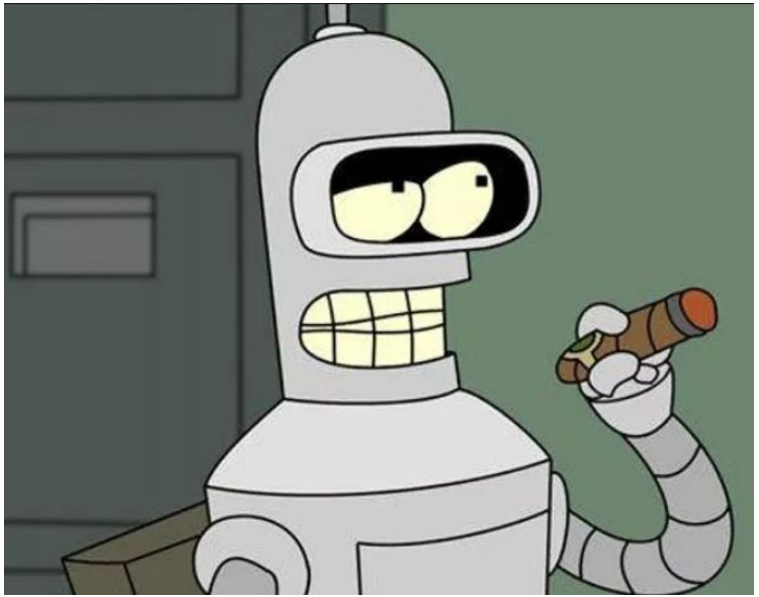
```java
public void sayGreetings(int currentTimeInHours) {

    if (currentTimeInHours > 4 && currentTimeInHours < 11) {
        System.out.println("Good morning!");
    } else if (currentTimeInHours > 17 && currentTimeInHours < 24) {
        System.out.println("Good evening!");
    } else {
        System.out.println("Hello!");
    }

}
```




Good morning!
10:08

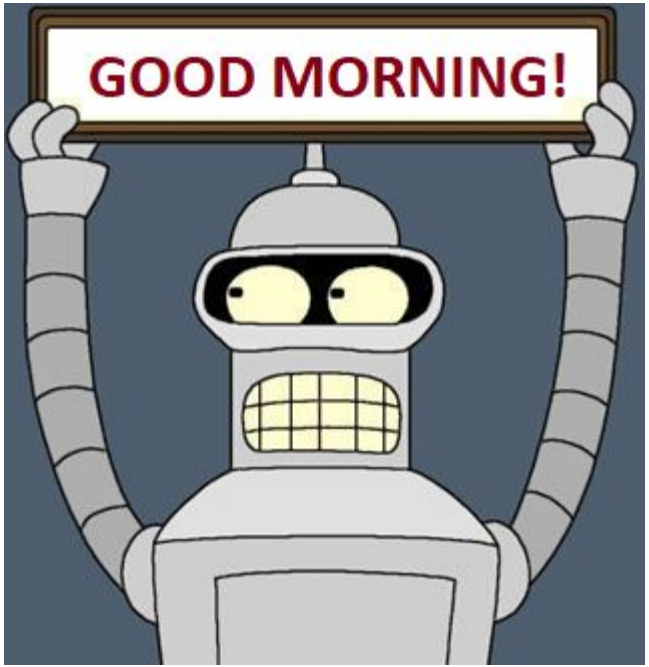



19:30
Good evening!




Hello!
12:05

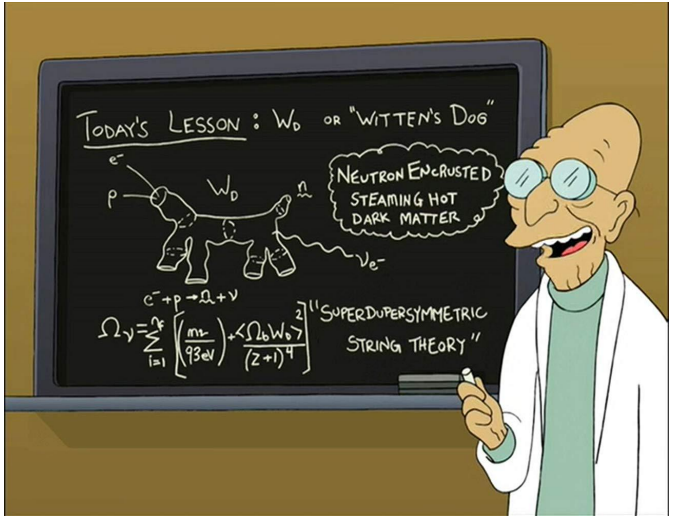

AM
PM
10:30
THU NOV 12


GOOD MORNING!

Machine Learning Tasks

- Superviced

- Unsupervised



Classification

- Identifies to which category a new observation belongs
- Uses a training set of data containing observations with known categories

Puppies

Kittens

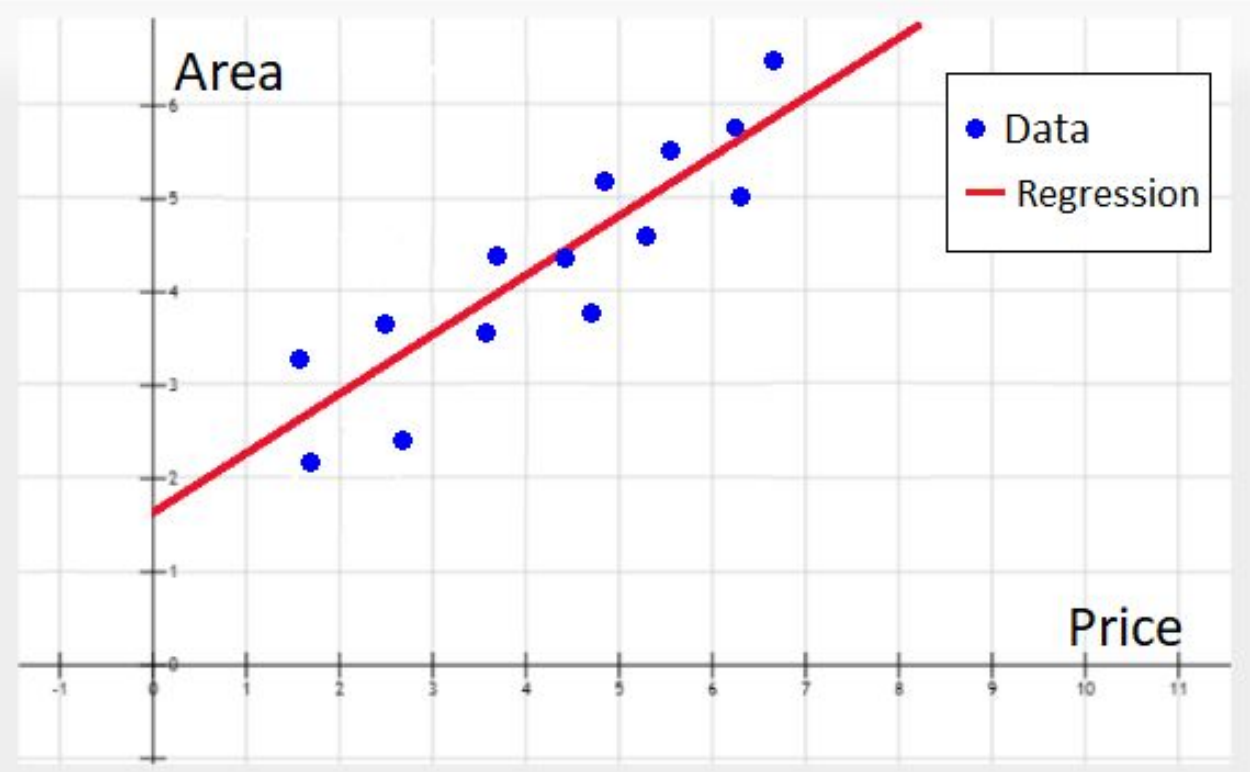
Baby raccoons

?

Regression

- Predicts quantity
- Looks for dependencies between values

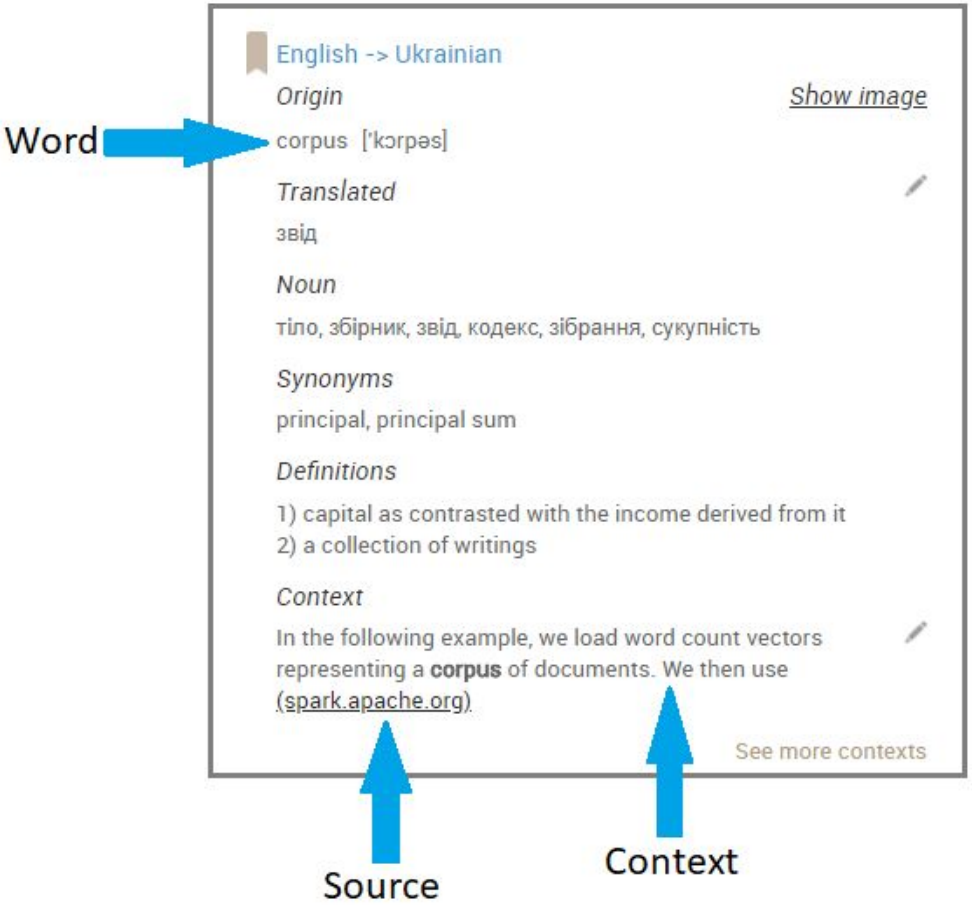
House Price Prediction



Clustering

- Unsupervised algorithm
- Detects dependencies without training data

Cloud notebook for translations



https://storetranslate.com

Logistic Regression

Recommend articles based on translation history

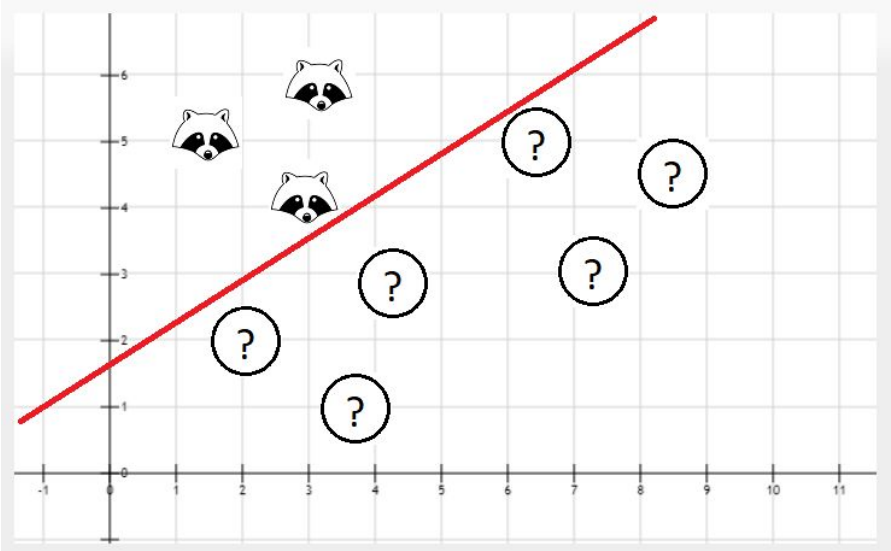
Recommend new words

Collaborative Filtering

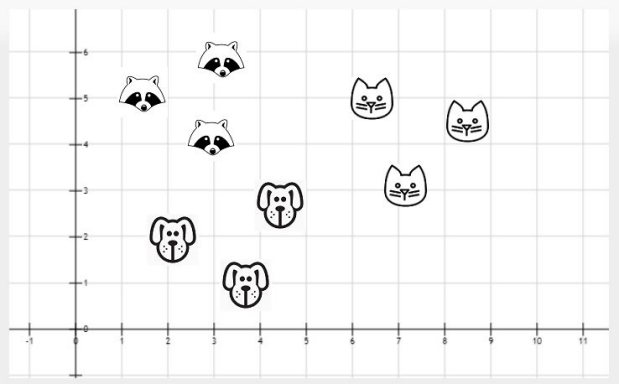
Logistic Regression (binary)

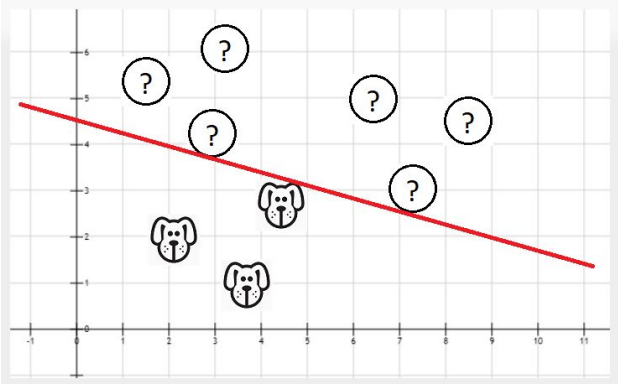
- Supervised **classification** algorithm

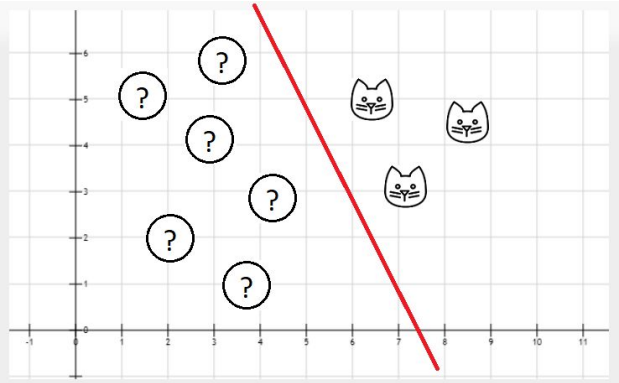
- Determines if observations belongs to some category - **yes/no**

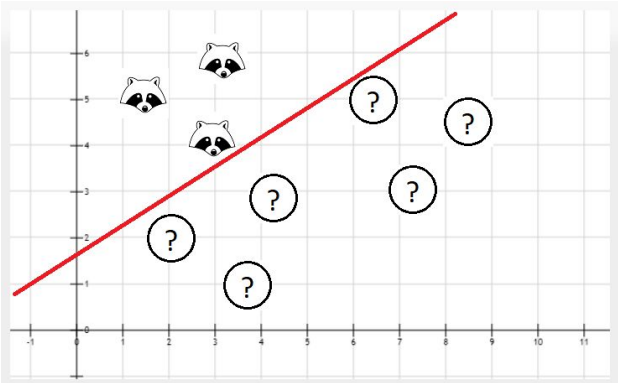


Logistic Regression (Multiclass)









Logistic Regression

Recommend articles based on translation history

Detect translation topic

↓

Find sources of similar words of other users

↓

Find the most relevant to detected topic

Detect translation topic

emphasize	http://source11	context11
probe	http://source12	context12
exploration	http://source11	context13
overfitting	http://source13	context14
explicitly	http://source13	context15

achieve	http://source21	context21
emphasize	http://source22	context22
recognition	http://source23	context23
frequently	http://source22	context24
occur	http://source23	context25

rhythm	http://source31	context31
display	http://source32	context32
overdrive	http://source31	context33
significance	http://source33	context34
emphasize	http://source33	context35

emphasize	http://source11	context11
probe	http://source12	context12
exploration	http://source11	context13
overfitting	http://source13	context14
explicitly	http://source13	context15

achieve	http://source21	context21
emphasize	http://source22	context22
recognition	http://source23	context23
frequently	http://source22	context24
occur	http://source23	context25

rhythm	http://source31	context31
display	http://source32	context32
overdrive	http://source31	context33
significance	http://source33	context34
emphasize	http://source33	context35

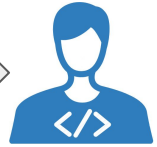
emphasize	http://source11	context11
emphasize	http://source22	context22
emphasize	http://source33	context35

Classify contexts

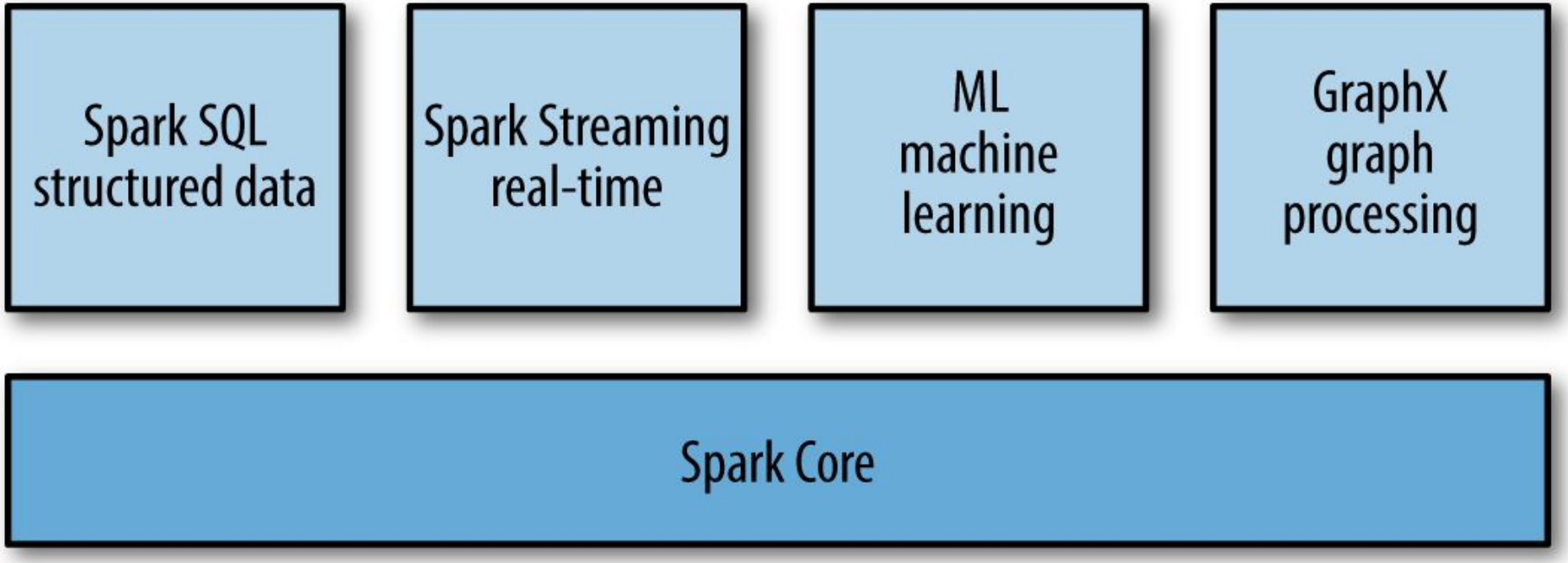
http://source11	science
http://source22	**science**
http://source33	music

Whitelist and robots.txt filter

http://source22



Apache Spark

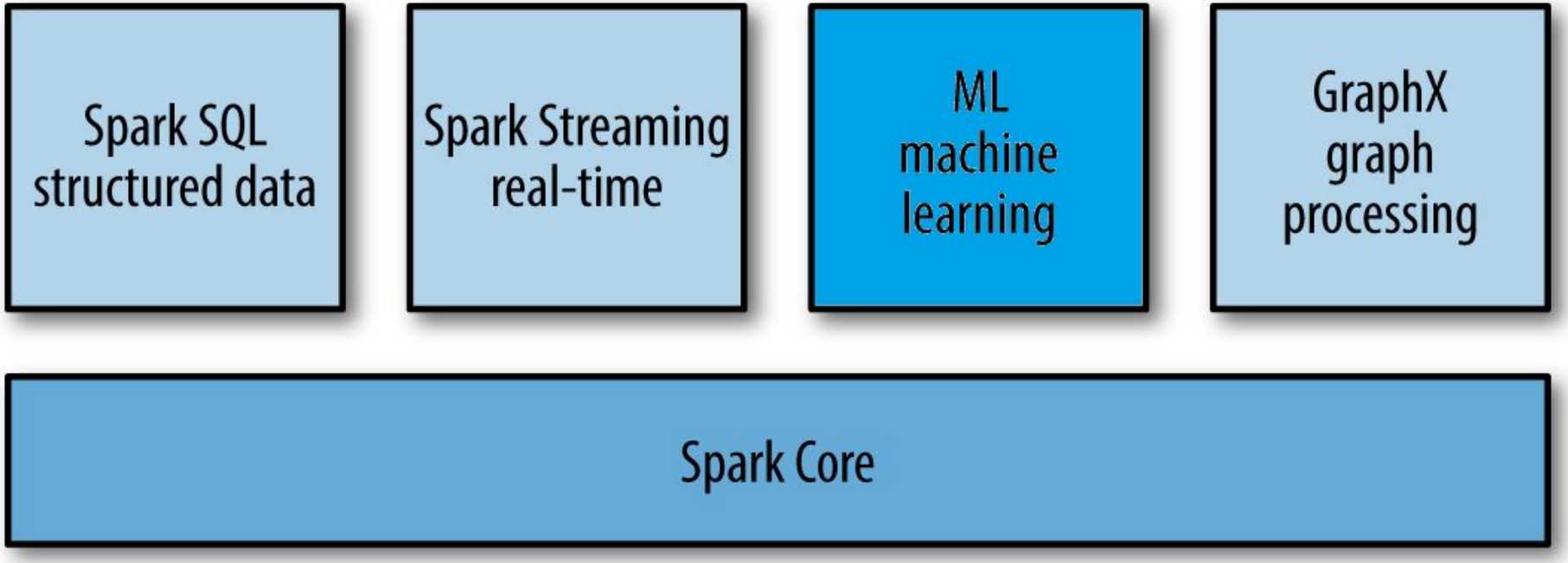


Spark Core

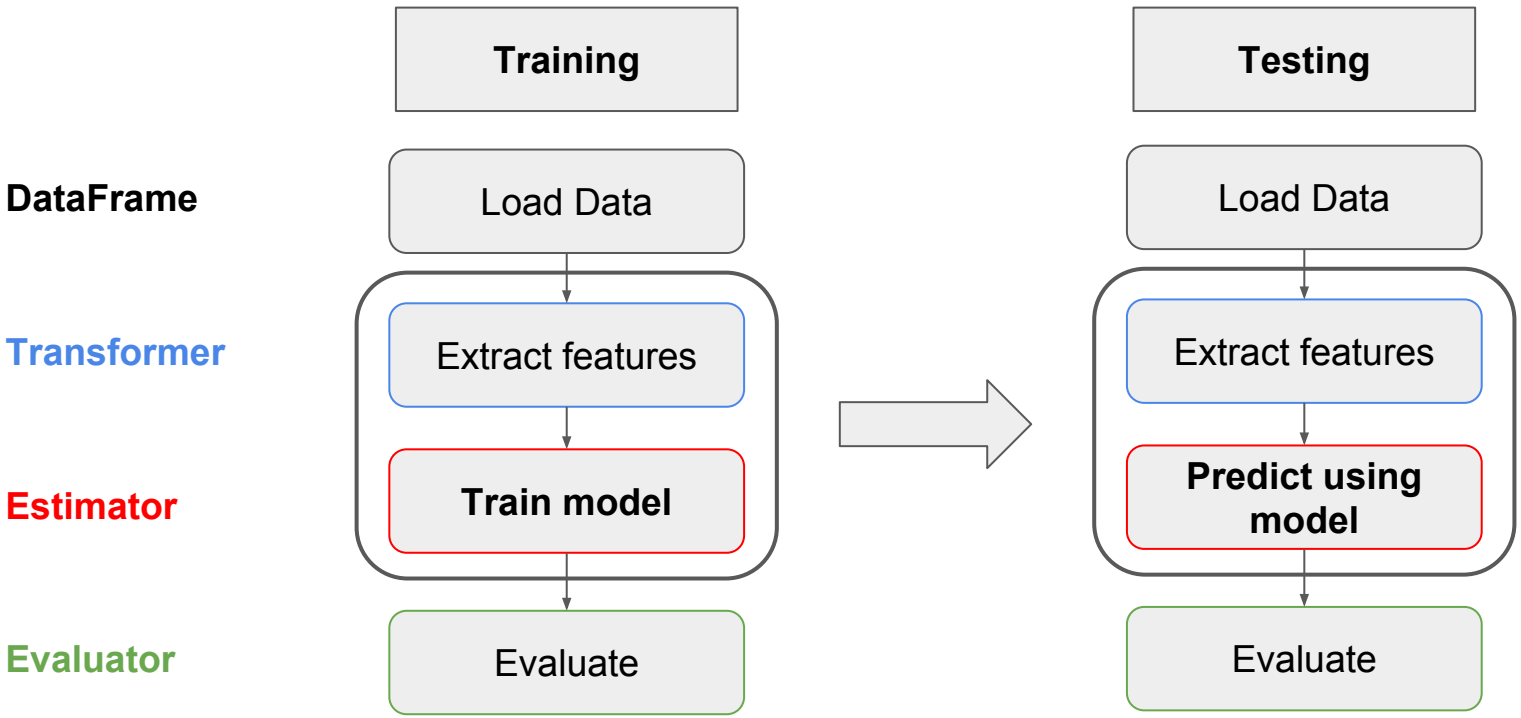
- **RDD** - Resilient Distributed Dataset
  - transformation - map, flatMap
  - action - count, collect

- **DataFrame** - Structured wrapper on RDD

Apache Spark



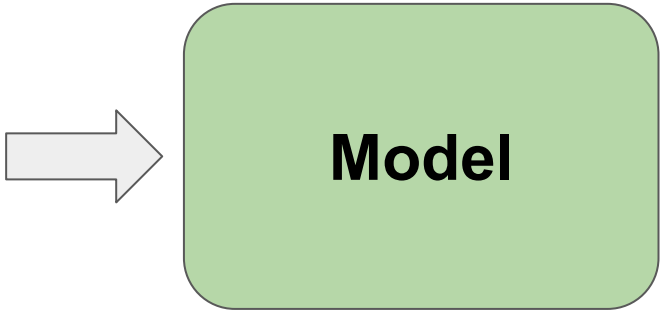
Spark ML Workflow



Features

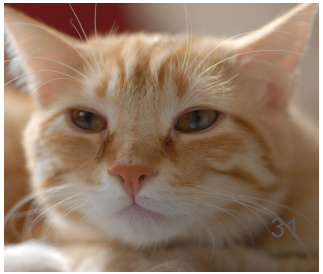
Elongated face
Black mask found around the eyes

Round face
No black mask found around the eyes

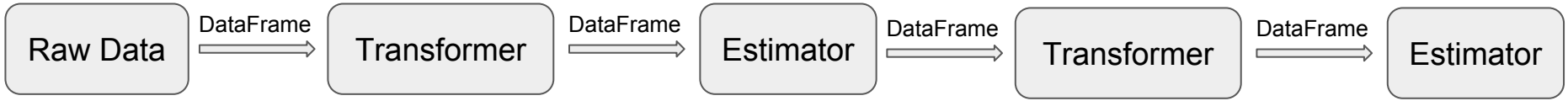


?

?



Spark ML Pipeline

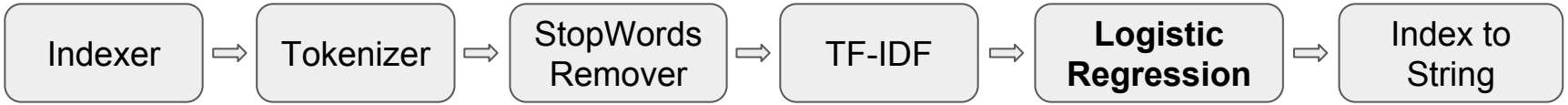


- **DataFrame** - ML dataset like SQL table
- **Transformer** - transforms one DataFrame to another
- **Estimator** - fits DataFrame and produces Transformer
- **Pipeline** - chains Transformers and Estimators

Transformer.**transform()**

Estimator**.fit()**

Spark ML Pipeline (Topic Detection)



Spark ML Pipeline

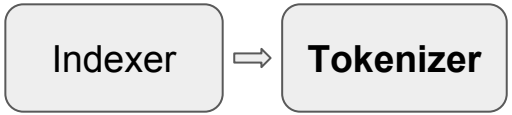
Indexer

text	label
some text 1	music
some text 2	space
some text 3	music
some text 4	finance

⇨

text	index
some text 1	0.0
some text 2	1.0
some text 3	0.0
some text 4	2.0

Spark ML Pipeline

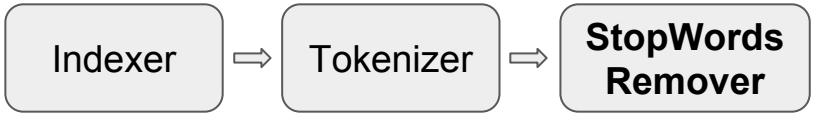


text
some text 1
some text 2
some text 3
some text 4

⇨

text	words
some text 1	(some, text, 1)
some text 2	(some, text, 2)
some text 3	(some, text, 3)
some text 4	(some, text, 4)

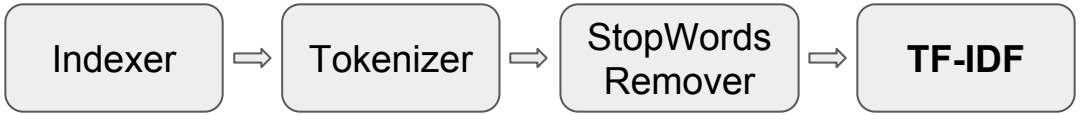
Spark ML Pipeline



text
Korolyov was a Ukrainian rocket engineer who launched the first human into space
Massive Attack are a British musical group formed in 1988 in Bristol

filtered
(Korolyov, Ukrainian, rocket, engineer, launched, human, space)
(Massive, Attack, British, musical, group, formed, 1988, Bristol)

Spark ML Pipeline



TF - Term Frequency

N of term / N of all terms

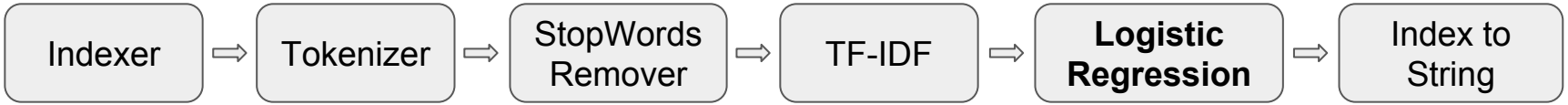
0.034	a
0.031	the
0.027	is
0.024	that
0.021	when
0.017	where

IDF - Inverse Document Frequency

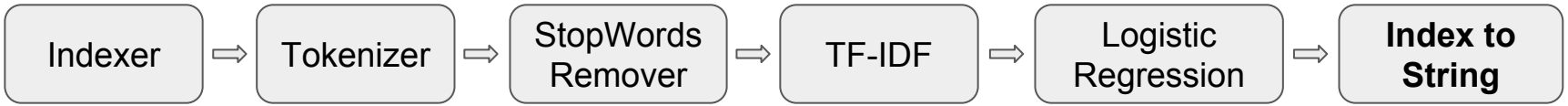
log (N of all docs / N of docs with term in it)

0.009	space
0.008	rocket
0.008	astronaut
0.007	spacecraft
0.006	earth
0.005	engineering

Spark ML Pipeline



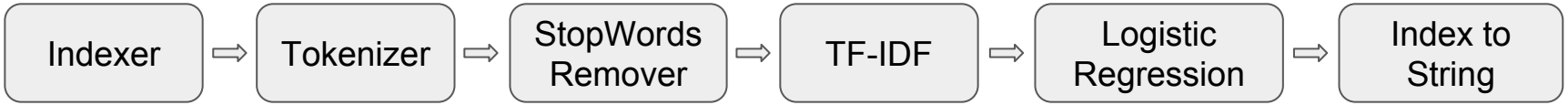
Spark ML Pipeline



text	index
some text 1	0.0
some text 2	1.0
some text 3	0.0
some text 4	2.0

text	label
some text 1	music
some text 2	space
some text 3	music
some text 4	finance

Spark ML Pipeline




```
StringIndexer RegexTokenizer StopWordsRemover  CountVectorizer IDF  LogisticRegression  IndexToString
```

Talk is cheap. Show me the code

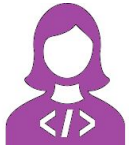
Collaborative Filtering

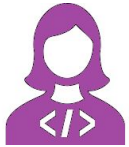
- Learns without information about users or products attributes
- Learns by collecting interactions between users and products

Collaborative Filtering

Recommended restaurants nearby my location

Recommended songs played by other users which played (bought) other same songs as me

	KAZKA Плакала	**Lady Gaga** Alexandro	**Rammstein** Ich Will	**SOAD** Chop Suey
	1	3	16	25
	25	18	4	3
	34	35	3	2
	1	2	36	23
	2	?	45	?

	KAZKA Плакала	**Lady Gaga** Alexandro	**Rammstein** Ich Will	**SOAD** Chop Suey
	1	3	**16**	**25**
	**25**	**18**	4	3
	**34**	**35**	3	2
	1	2	**36**	**23**
	2	?	**45**	?

	KAZKA Плакала	**Lady Gaga** Alexandro	**Rammstein** Ich Will	**SOAD** Chop Suey
	-	-	+	+
	+	+	-	-
	+	+	-	-
	-	-	+	+
	-	?	+	?

	KAZKA Плакала	**Lady Gaga** Alexandro	**Rammstein** Ich Will	**SOAD** Chop Suey
	-	-	+	+
	+	+	-	-
	+	+	-	-
	-	-	+	+
	-	-	+	+

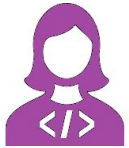
Collaborative Filtering

Recommend new words

Find users wich translated the same words as I

Find words of these users which I didn’t translate

achieve	1
emphasize	2
frequently	4
explicitly	1

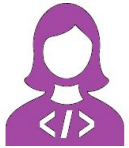


achieve	2
emphasize	3
frequently	1
occur	1

rhythm	3
overdrive	1
significance	5
play	1

rhythm	1
overdrive	1
significance	2
harmony	3

achieve	1
emphasize	2
frequently	4
explicitly	1

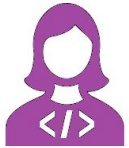


achieve	2
emphasize	3
frequently	1
occur	1

rhythm	3
overdrive	1
significance	5
play	1

rhythm	1
overdrive	1
significance	2
harmony	3

achieve	1
emphasize	2
frequently	4
explicitly	1

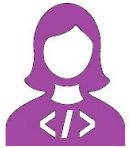


achieve	2
emphasize	3
frequently	1
occur	1

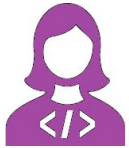
rhythm	3
overdrive	1
significance	5
play	1

rhythm	1
overdrive	1
significance	2
harmony	3

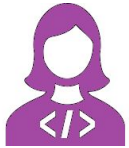
Matrix Factorization

achieve	1	2		
emphasize	2	3		
frequently	4	1		
explicitly	1			
occur		1		
rhythm			3	1
overdrive			1	1
significance			5	2
play			1	
harmony				3

achieve	1	1		
emphasize	1	1		
frequently	1	1		
explicitly	1			
occur		1		
rhythm			1	1
overdrive			1	1
significance			1	1
play			1	
harmony				1

achieve	1	1	?	?
emphasize	1	1	?	?
frequently	1	1	?	?
explicitly	1	?	?	?
occur	?	1	?	?
rhythm	?	?	1	1
overdrive	?	?	1	1
significance	?	?	1	1
play	?	?	1	?
harmony	?	?	?	1

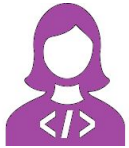
	feature 1	feature 2	feature 3	feature 4
achieve	1	1	0	0
emphasize	1	1	0	0
frequently	1	1	0	0
explicitly	1	1	0	0
occur	1	1	0	0
rhythm	0	0	1	1
overdrive	0	0	1	1
significance	0	0	1	1
play	0	0	1	1
harmony	0	0	1	1

X

feature 1	1	1	0	0
feature 2	1	1	0	0
feature 3	0	0	1	1
feature 4	0	0	1	1

Y^T

$$A \approx XY^T$$

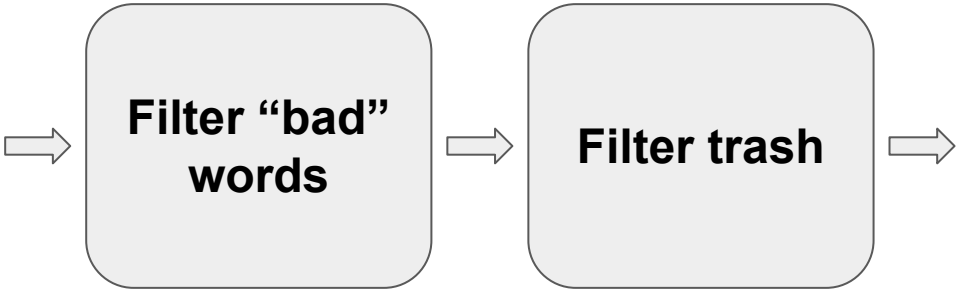



achieve	1	1	0	0
emphasize	1	1	0	0
frequently	1	1	0	0
explicitly	1	**1**	0	0
occur	**1**	1	0	0
rhythm	0	0	1	1
overdrive	0	0	1	1
significance	0	0	1	1
play	0	0	1	**1**
harmony	0	0	**1**	1

Alternating Least Squares

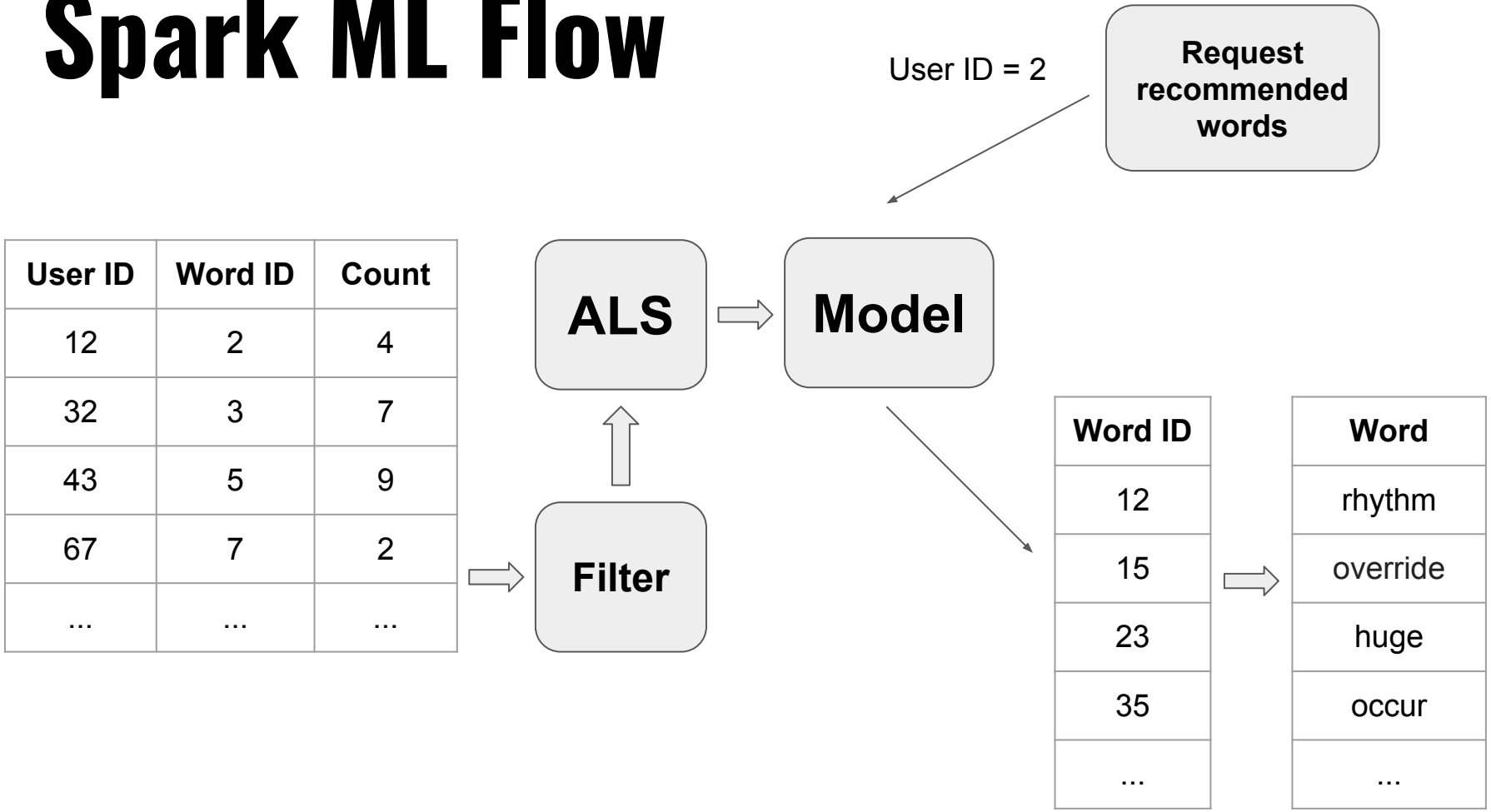
Spark ML Flow

Word ID	Word
2	achieve
3	emphasize
9	f***
5	frequently
11	віiва iадл
7	occur
14	$echo()
...	...



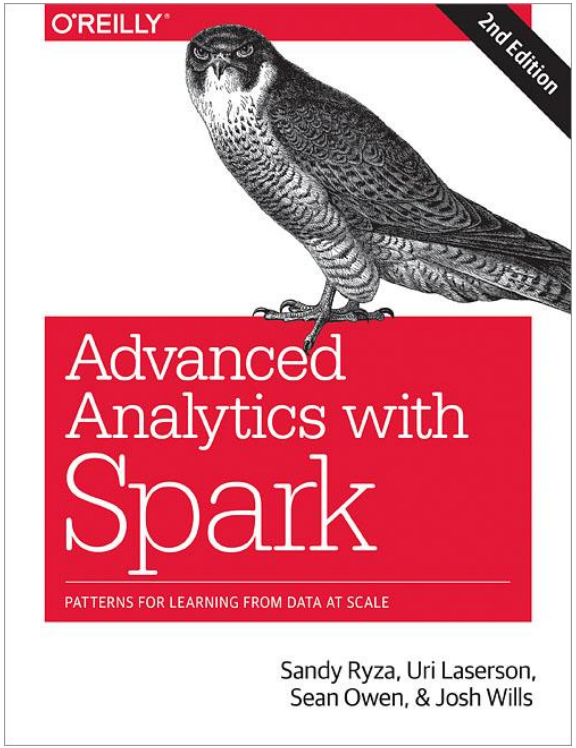
Word ID	Word
2	achieve
3	emphasize
5	frequently
7	occur
...	...

Spark ML Flow



Talk is cheap. Show me the code

Useful Resources



https://www.amazon.com/Advanced-Analytics-Spark-Patterns-Learning/dp/1491912766

Useful Resources

Machine Learning

Stanford University

https://www.coursera.org/learn/machine-learning/home/welcome

Thank you!

/igorock/teach-machine-to-teach