livelessons
video instruction from technology experts

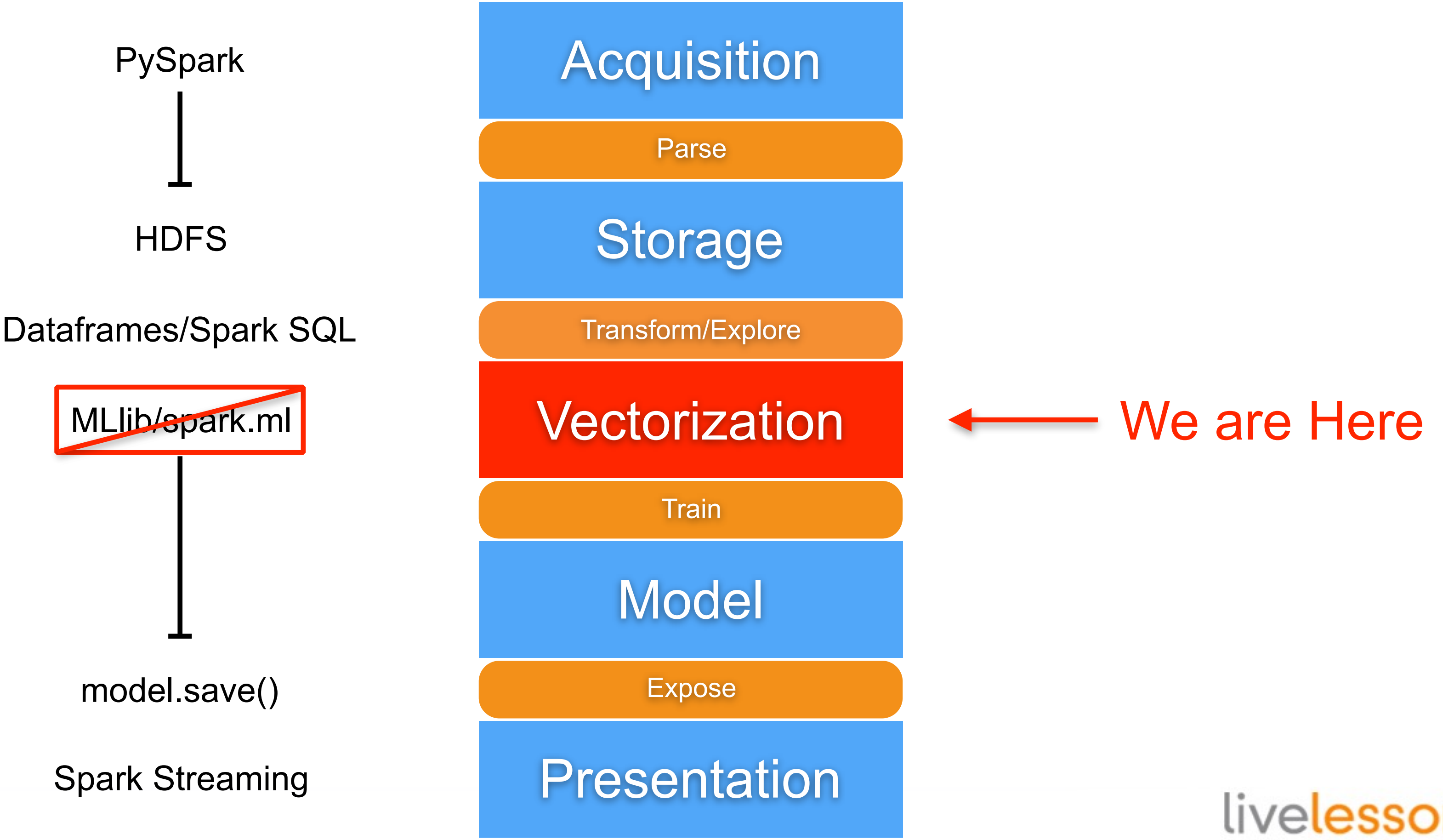
Lesson 3: Your First Spark Application

3.6 Working with Text: Introduction to NLP

livelessons

At Scale

Data Pipeline



PySpark

HDFS

Dataframes/Spark SQL

~~MLlib/spark.ml~~

model.save()

Spark Streaming

Acquisition

Parse

Storage

Transform/Explore

Vectorization

Train

Model

Expose

Presentation

We are Here

Natural Language Processing

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

[1, 3, 1, 1, 2, 0, 1, 0]

[0, 1, 4, 0, 0, 1, 1, 1]

[3, 0, 1, 1, 2, 2, 3, 2]

[0, 1, 1, 1, 0, 3, 2, 3]

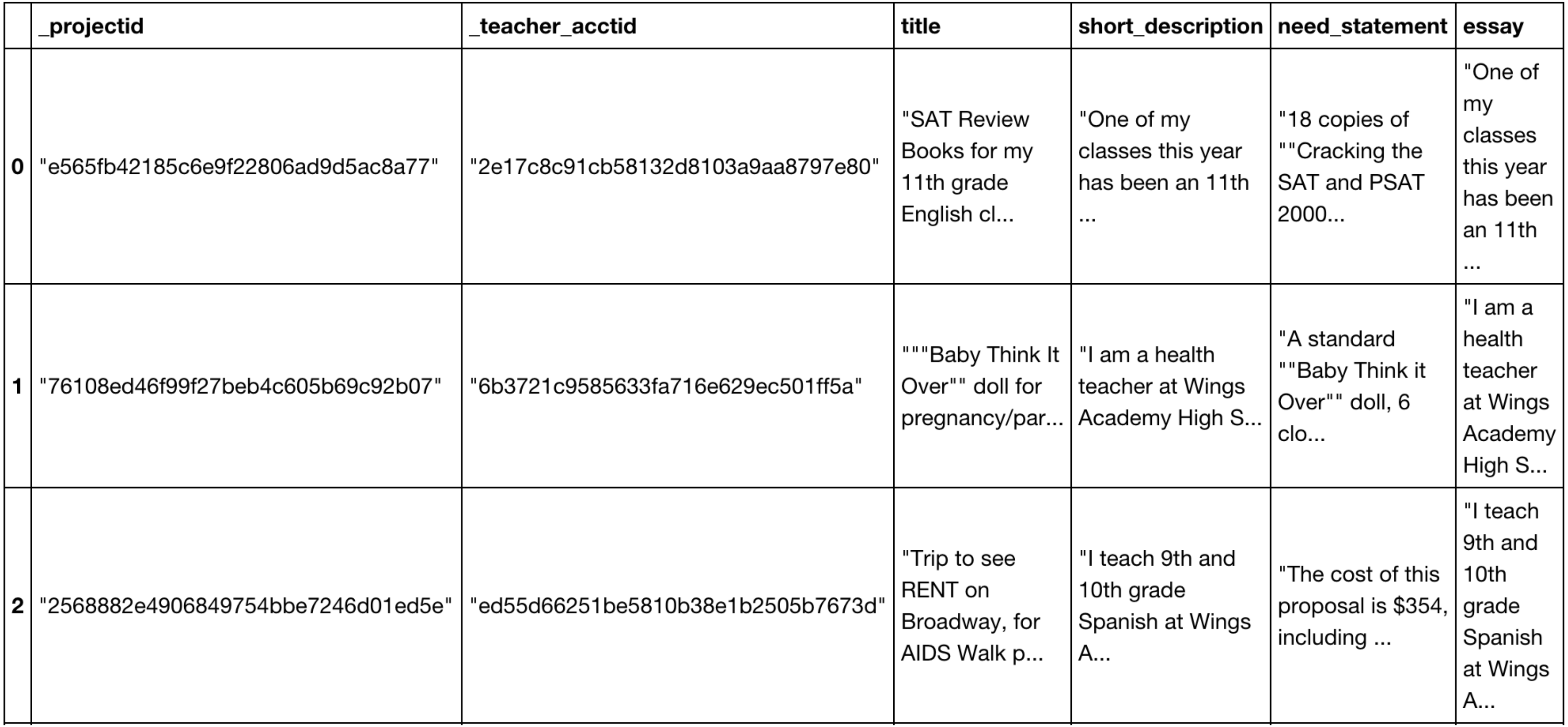
[1, 2, 1, 2, 2, 0, 0, 0]

[1, 0, 1, 1, 0, 1, 1, 1]

[0, 2, 0, 0, 2, 2, 0, 0]

[1, 1, 1, 1, 0, 1, 1, 1]

DonorsChoose: Project Essays

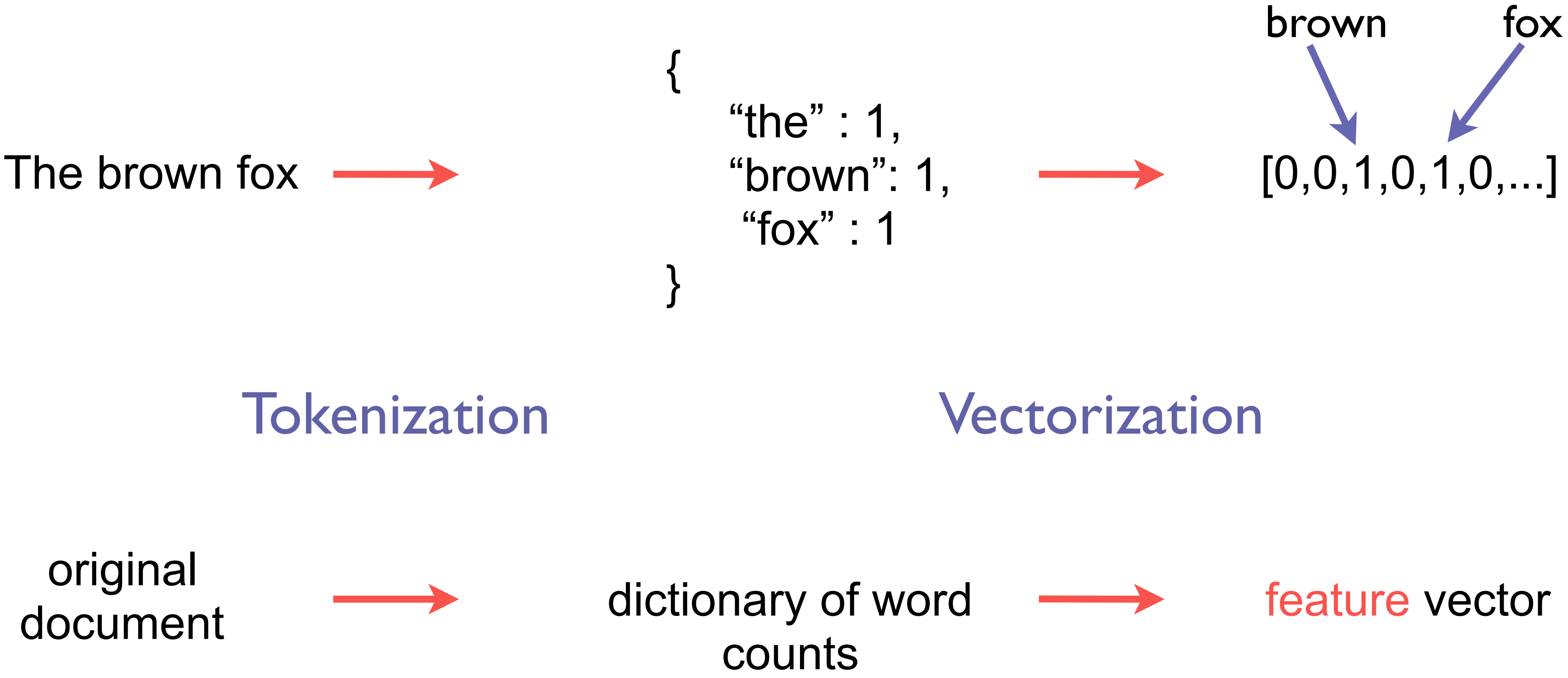


	_projectid	_teacher_acctid	title	short_description	need_statement	essay
0	"e565fb42185c6e9f22806ad9d5ac8a77"	"2e17c8c91cb58132d8103a9aa8797e80"	"SAT Review Books for my 11th grade English cl...	"One of my classes this year has been an 11th ...	"18 copies of ""Cracking the SAT and PSAT 2000...	"One of my classes this year has been an 11th ...
1	"76108ed46f99f27beb4c605b69c92b07"	"6b3721c9585633fa716e629ec501ff5a"	"""Baby Think It Over"" doll for pregnancy/par...	"I am a health teacher at Wings Academy High S...	"A standard ""Baby Think it Over"" doll, 6 clo...	"I am a health teacher at Wings Academy High S...
2	"2568882e4906849754bbe7246d01ed5e"	"ed55d66251be5810b38e1b2505b7673d"	"Trip to see RENT on Broadway, for AIDS Walk p...	"I teach 9th and 10th grade Spanish at Wings A...	"The cost of this proposal is $354, including ...	"I teach 9th and 10th grade Spanish at Wings A...

Bag of Words

- **Document:** Single row of data/corpus
- **Corpus:** Entire set of all documents
- **Vocabulary:** Set of all words in corpus
- **Vector:** Mathematical representation of document (counts of word occurrences)

Bag of Words



The brown fox

{
“the” : 1,
“brown”: 1,
“fox” : 1
}

brown

fox

[0,0,1,0,1,0,...]

Tokenization

Vectorization

original document

dictionary of word counts

feature vector