BIG DATA

David Kauchak
CS158 – Spring 2022

1

Admin

Assignment 8

Assignment 9

2

Big Data

What is "big data"?

What are some sources of big data?

What are the challenges of dealing with big data?

What are some of the tools you've heard of?

3

Big data and ML

Why talk about it in a course like this?

4

Machine Learning is…

Machine learning is about predicting the future based on the past.
-- Hal Daume III

5

Machine Learning is…

Machine learning is about predicting the future based on the past.
-- Hal Daume III

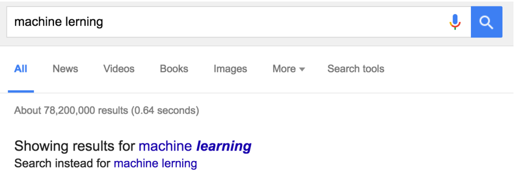
If the "past" has lots of data, then
we need tools to process it!

6

Big data and ML

Why talk about it in a course like this?

Many "machine learning" problems become
much easier when you have lots of data


machine lerning

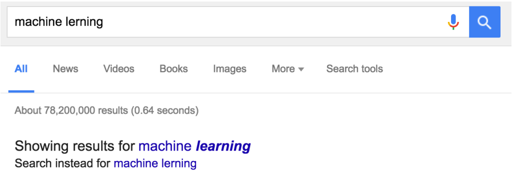
About 78,200,000 results (0.64 seconds)

Showing results for machine ***learning***
Search instead for machine lerning

7

Big data and ML

How would you do it?


machine lerning

About 78,200,000 results (0.64 seconds)

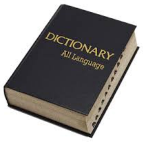
Showing results for machine ***learning***
Search instead for machine lerning

8

Big data and ML

How would you do it?

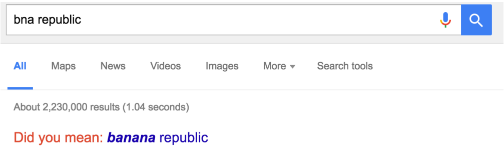
edit distance



9

Big data and ML

How would you do it?



bna republic

About 2,230,000 results (1.04 seconds)

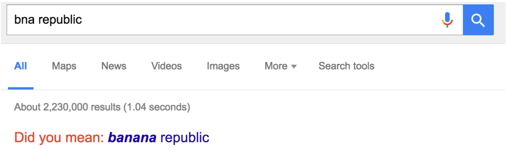
Did you mean: ***banana*** republic

May not get example like this!

10

Big data and ML

How would they do it?
(small company)



bna republic

About 2,230,000 results (1.04 seconds)

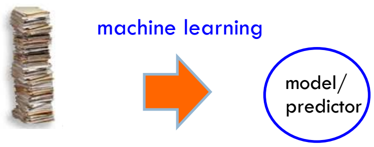
Did you mean: ***banana*** republic

May not get example like this correct!

11

Big data and ML

How would they do it?
(small company)



machine learning

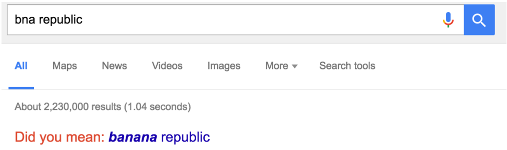
model/
predictor

text corpus

12

Big data and ML

How does Google do it?



bna republic

About 2,230,000 results (1.04 seconds)

Did you mean: ***banana*** republic

May not get example like this!

13

Big data and ML

Google now handles at least 2 trillion searches per year

The search giant won't say exactly how many trillions of queries it processes, other than it's now two or more. It last claimed 1.2 trillion in 2012.

http://searchengineland.com/google-now-handles-2-999-trillion-searches-per-year-250247

14

Big data and ML

Search logs

user_id	time	query
	…	
131524	t	bna republic
	…	
131524	t+5s	banana republic
	…	

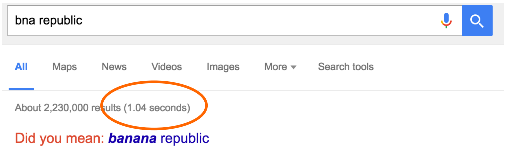
Many problems get easy when you have lots of data!

15

Big data and ML

Many **problems** get easy when you have lots of data!

Challenge: processing all this data in an efficient way

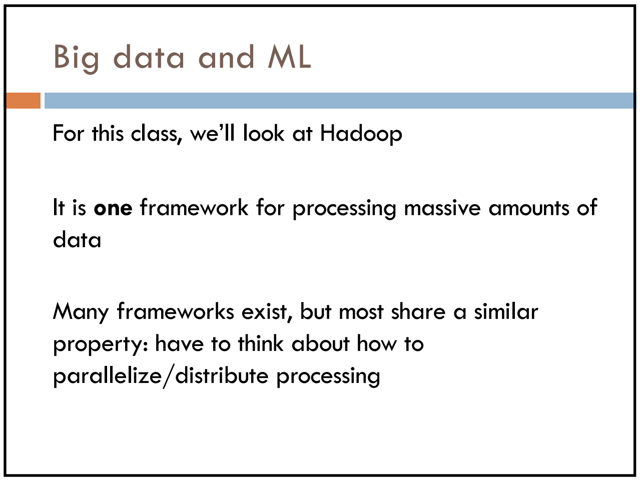


bna republic

About 2,230,000 results (1.04 seconds)

Did you mean: ***banana*** republic

16



Big data and ML

For this class, we'll look at Hadoop

It is **one** framework for processing massive amounts of data

Many frameworks exist, but most share a similar property: have to think about how to parallelize/distribute processing

17