DATASTAX ACADEMY

Classification and Clustering

Paired with Wine and Chocolate

Amanda K Moran
Developer Advocate for DataStax

©2019 DataStax.
Use only with permission.

But first… A Little About Amanda

- South Bay!
- Graduated with MS in Computer Science and Engineering from Santa Clara University in 2012
- Worked as a Software Engineer for 6 years at Lockheed Martin, HP, Teradata, Esgyn and now DataStax as a Developer Advocate
- Apache Committer, PMC Member, and initial contributor to all installation and deployment work for Apache Trafodion
- Keywords: Disney, Cloud, Dogs, Veggies, Linux, Databases, Big Data, Analytics, Testing, and Running

What Are We Talking About Today

- What Problem Are We Trying To Solve?
  - Can I use clustering ML algorithms to find which wine comes from which vineyard?
  - Can I use classification ML algorithms to find which country a candy bar comes from?
- Introduction To Apache Spark
- Introduction To Cassandra <-> Spark
- What is KMeans?
  - Demo: KMeans and Wine
- What is Naive Bayes?
  - Demo: Naive Bayes and Chocolate

© DataStax, All Rights Reserved.

What Problem Are We *Really* Trying to Solve?

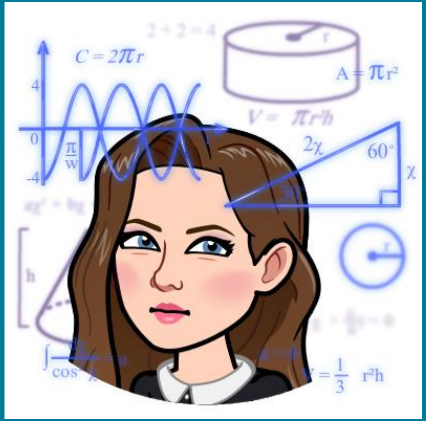
© DataStax, All Rights Reserved.

Can I use Machine Learning with Apache Spark with Wine and Chocolate?

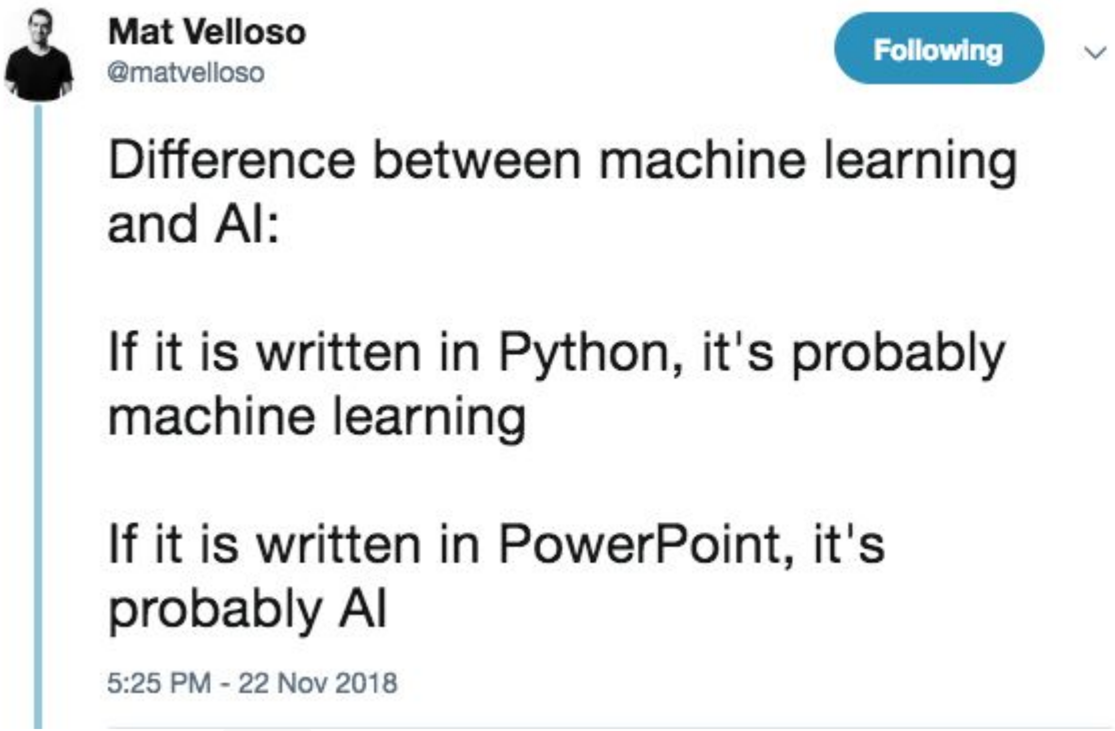
- Data Analytics doesn’t have to be complicated!
- We are going to utilize the power of Big Data using
  - Apache Cassandra
  - Apache Spark
  - Apache Cassandra <-> Apache Spark Connector
  - Spark Machine Learning library
    - Kmeans
    - Naive Bayes
  - Jupyter notebooks
  - Python

© DataStax, All Rights Reserved.

Let’s Talk about AI/ML



ML vs AI



Mat Velloso
@matvelloso

Difference between machine learning and AI:

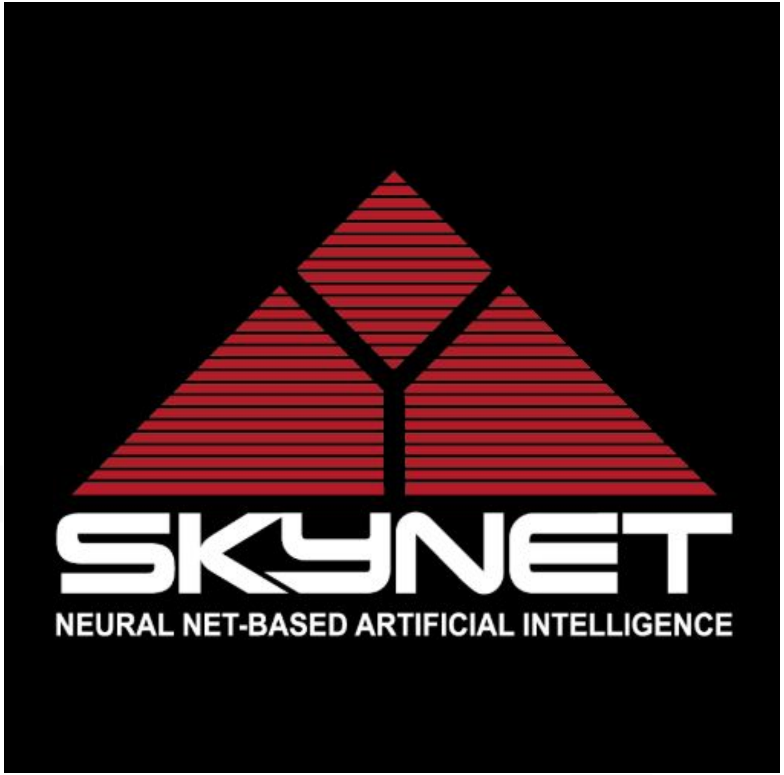
If it is written in Python, it's probably machine learning

If it is written in PowerPoint, it's probably AI

5:25 PM - 22 Nov 2018

© DataStax, All Rights Reserved.

AI



© DataStax, All Rights Reserved.

AL

- Article Published Yesterday:
- 40% of all AI based startups in Europe have nothing to do with AI.
  - https://www.ft.com/content/21b19010-3e9f-11e9-b896-fe36ec32aece

Machine Learning

- We are going to focus on the practical.
- How we can actually use *Machine Learning* to get some information from our data!

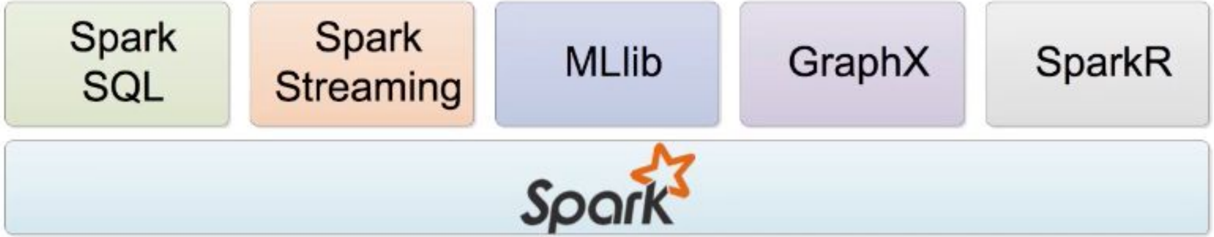
© DataStax, All Rights Reserved.

(Small) Introduction to Apache Spark

© DataStax, All Rights Reserved.

What is Apache Spark?

- “Apache Spark is a unified analytics engine for large-scale data processing” --https://spark.apache.org
- 100 times faster than Hadoop for analytics
  - Utilizes in-memory processing
  - Amazing parallelism
- Machine Learning library -- Spark MLlib



Apache Cassandra and Apache Spark Connector

© DataStax, All Rights Reserved.

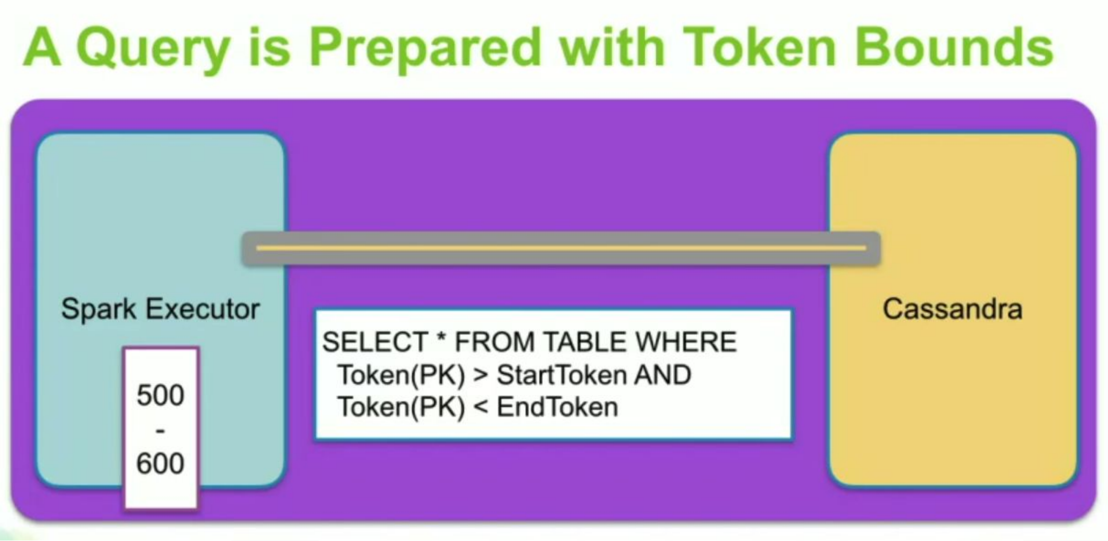
What is Apache Cassandra/Spark Connector

- Co-located (Apache Cassandra and Apache Spark on both on the same node)
- Will gather data from Apache Cassandra and its known token range and page that into the Spark Executor

- Utilized the DataStax Java driver under the hood to move data between Apache Cassandra and Apache Spark
- Keeps paging in data until query is complete

What is Apache Cassandra/Spark Connector

- Reference: https://databricks.com/session/spark-and-cassandra-2-fast-2-furious



A Query is Prepared with Token Bounds

```
SELECT * FROM TABLE WHERE
  Token(PK) > StartToken AND
  Token(PK) < EndToken
```

Demo:

How to setup Apache Cassandra, Apache Spark,

with the Connector and Jupyter

© DataStax, All Rights Reserved.

What is Clustering?

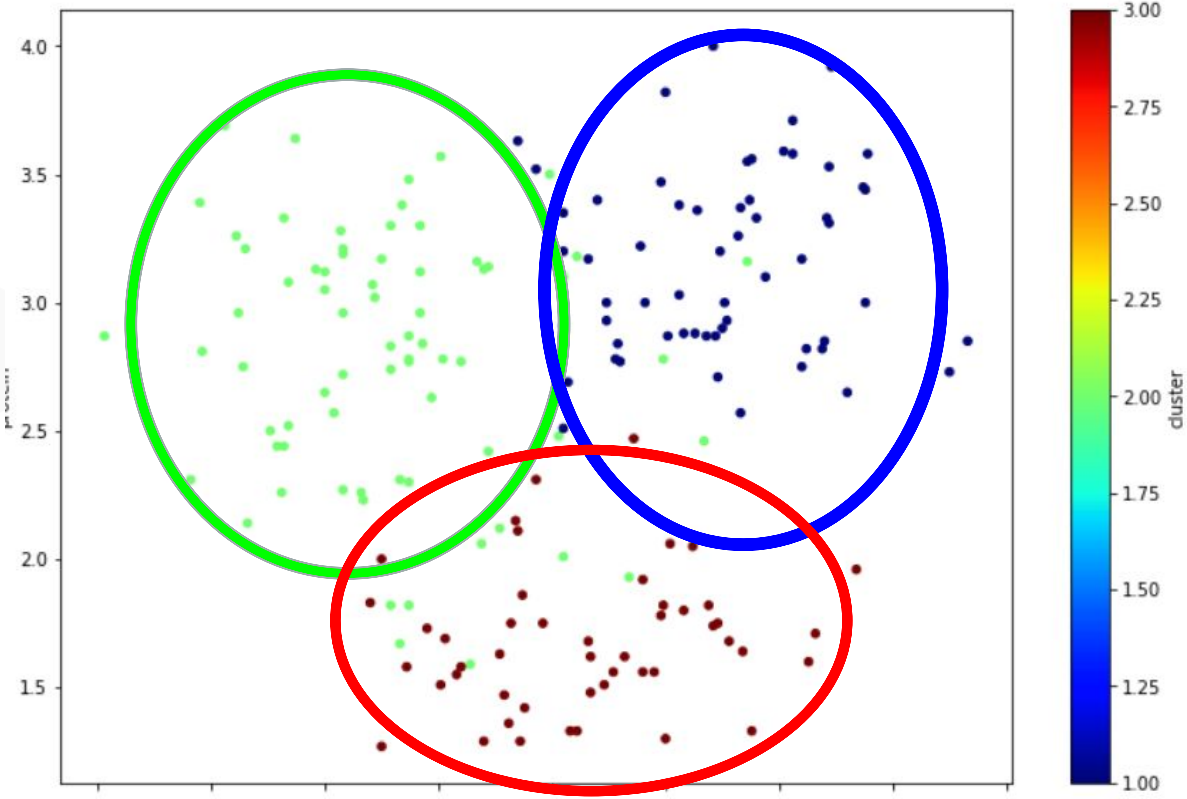
© DataStax, All Rights Reserved.

What are Clustering Algorithms ?

- **Cluster analysis** or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters).

© DataStax, All Rights Reserved.

What are Clustering Algorithms ?



© DataStax, All Rights Reserved.

What are Clustering Algorithms ?

- **Cluster analysis** or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters).

© DataStax, All Rights Reserved.

What is KMeans ?

- K-means clustering is a simple unsupervised learning algorithm that is used to solve clustering problems.
- It follows a simple procedure of classifying a given data set into a number of clusters, defined by the letter "k," which is fixed beforehand.

What is KMeans ?

- The clusters are then positioned as points and all observations or data points are associated with the nearest cluster, computed, adjusted and then the process starts over using the new adjustments until a desired result is reached.

© DataStax, All Rights Reserved.

When to use KMeans?

- The *K*-means clustering algorithm is used to find groups which have not been explicitly labeled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets.
  - Behavior Segmentation of customers
    - Segment by purchase history
  - Finding anomalies

© DataStax, All Rights Reserved.

Demo: KMeans and Wine

Can I use clustering ML algorithms to find which wine comes from which vineyard?

© DataStax, All Rights Reserved.

What is Classification?

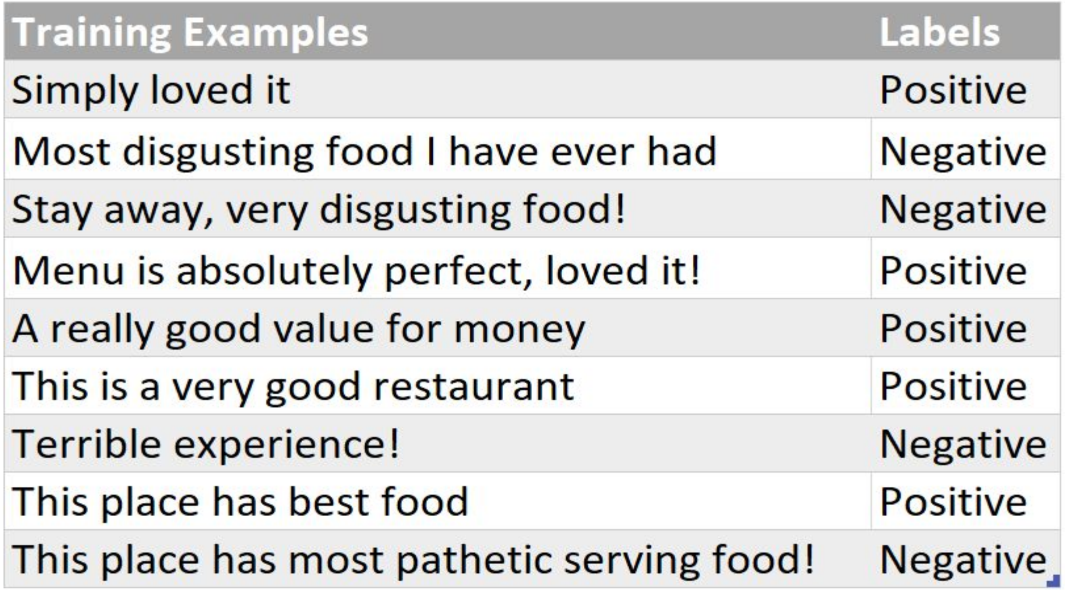
© DataStax, All Rights Reserved.

What are Classification Algorithms ?

- In machine learning and statistics, **classification** is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

What are Classification Algorithms ?

- Examples are assigning a given email to the "spam" or "non-spam" class. Classification is an example of pattern recognition.



Training Examples	Labels
Simply loved it	Positive
Most disgusting food I have ever had	Negative
Stay away, very disgusting food!	Negative
Menu is absolutely perfect, loved it!	Positive
A really good value for money	Positive
This is a very good restaurant	Positive
Terrible experience!	Negative
This place has best food	Positive
This place has most pathetic serving food!	Negative

© DataStax, All Rights Reserved.

What is Naive Bayes ?

- Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.

© DataStax, All Rights Reserved.

Demo: Naive Bayes and Chocolate

Can I use classification ML algorithms to find which country a candy bar comes from?

© DataStax, All Rights Reserved.

Okay, this was awesome! What now?

© DataStax, All Rights Reserved.

Information and Links

- Learn more about Cassandra: https://academy.datastax.com/
- Learn more about Spark: https://spark.apache.org/
- Learn more about DataStax: https://www.datastax.com/
- Follow me on Twitter: @AmandaDataStax
- Get this demo on Github: https://github.com/amandamoran/wineAndChocolate

DATASTAX ACADEMY

Amanda

@AmandaDataStax

© DataStax, All Rights Reserved.