



Mammoth Scale Machine Learning

Speaker: Robin Anil, Apache Mahout PMC Member

OSCON'10
Portland, OR
July 2010

Quick Show of Hands

- Are you fascinated about ML?
- Have you used ML?
- Do you have Gigabytes and Terabytes of data to analyze?
- Do you have Hadoop or MapReduce experience?

- Thanks for the survey!



Little bit about me

- Apache Mahout PMC member
- A ML Enthusiast 😊
- Software Engineer @ Google
- Google Summer of Code Mentor
- Previous Life: Google Summer of Code student for 2 years.



Agenda

- Introducing Mahout
- Different classes of problems
- And their Mahout based solutions
- Basic data structure
- Usage examples
- Sneak peek at our Next Release



The Mission

To build a *scalable* machine learning library



Scale!

- Scale to large datasets
 - Hadoop MapReduce implementations that scales linearly with data.
 - Fast sequential algorithms whose runtime doesn't depend on the size of the data
 - Goal: To be as fast as possible for any algorithm
- Scalable to support your business case
 - Apache Software License 2
- Scalable community
 - Vibrant, responsive and diverse
 - Come to the mailing list and find out more



The Mission

To build a scalable machine learning *library*



Why a new Library

- Plenty of open source Machine Learning libraries either
 - Lack community
 - Lack scalability
 - Lack documentations and examples
 - Lack Apache licensing
 - Are not well tested
 - Are Research oriented



Agenda

- ~~Introducing Mahout~~
- **Different classes of problems**
- **And their Mahout based solutions**
- Basic data structure
- Usage examples
- Sneak peek at our Next Release



ML on Twitter

- Collection of tweets in the last hour
- Each 140 character or token stream
- We will keep using this example throughout this talk



What is Clustering

- Call it fuzzy grouping based on a notion of similarity

[Spain »](#)

[Spain's World Cup win wasn't pretty, but wipes away its frustration](#)

Los Angeles Times - [Kevin Baxter](#) - 1 hour ago

A goal 28 minutes into extra time by Andres Iniesta defeats the Netherlands, 1-0, allowing Spain to claim sport's most cherished prize after decades of disappointment.

[+](#) [Video: Fans arrive for historic final](#)  [AFP](#)

[Spain beats Netherlands 1-0 to win the World Cup](#)

The Associated Press

[USA Today](#) - [Boston Globe \(blog\)](#) - [CNN International](#) -

[ESPN](#) - [Wikipedia: 2010 FIFA World Cup Final](#)

[all 7,948 news articles »](#)



New York Ti...



Mahout Clustering

- Plenty of Algorithms: K-Means, Fuzzy K-Means, Mean Shift, Canopy, Dirichlet
- Group similar looking objects
- Notion of similarity: Distance measure:
 - Euclidean
 - Cosine
 - Tanimoto
 - Manhattan

Spain »

[Spain's World Cup win wasn't pretty, but wipes away its frustration](#)

Los Angeles Times - Kevin Baxter - 1 hour ago

A goal 28 minutes into extra time by Andres Iniesta defeats the Netherlands, 1-0, allowing Spain to claim sport's most cherished prize after decades of disappointment.

[+](#) [Video: Fans arrive for historic final](#)  [AFP](#)

[Spain beats Netherlands 1-0 to win the World Cup](#)

The Associated Press

[USA Today](#) - [Boston Globe \(blog\)](#) - [CNN International](#) -

[ESPN](#) - [Wikipedia: 2010 FIFA World Cup Final](#)

[all 7,948 news articles »](#)



New York Ti...



Clustering Tweets

“Identify tweets that are similar and group them”



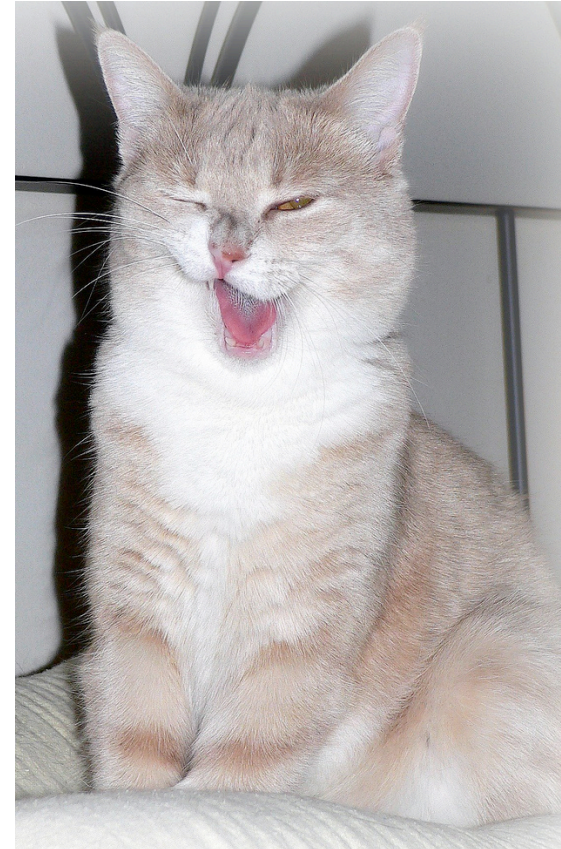
Topic modeling

- Grouping similar or co-occurring features into a topic
 - Topic “Lol Cat”:
 - Cat
 - Meow
 - Purr
 - Haz
 - Cheeseburger
 - Lol



Mahout Topic Modeling

- Algorithm: Latent Dirichlet Allocation
 - Input a set of documents
 - Output top K prominent topics and the features in each topic



Filtering Topics from Tweets

“Identify emerging topics in a collection of tweets”

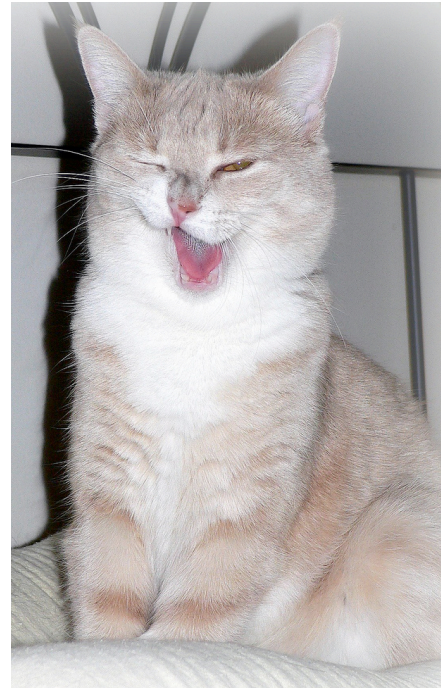


Classification

- Predicting the type of a new object based on its features
- The types are predetermined



Dog

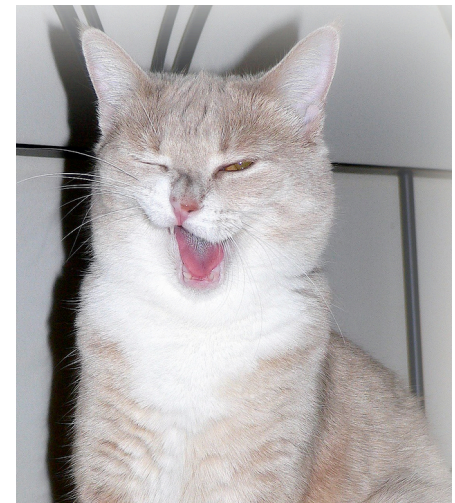
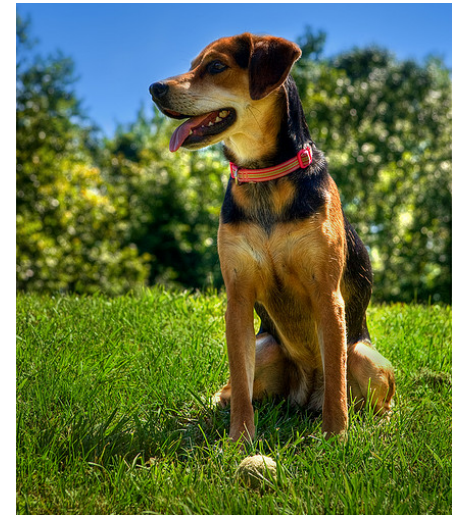


Cat



Mahout Classification

- Plenty of algorithms
 - Naïve Bayes
 - Complementary Naïve Bayes
 - Random Forests
 - Logistic Regression (Almost done)
 - Support Vector Machines (patch ready)
- Learn a model from a manually classified data
- Predict the class of a new object based on its features and the learned model



Detect OSCON Tweets

“Tweets without #OSCON”

Use tweets mentioning #OSCON to train
and
Classify incoming tweets



Recommendations

- Predict what the user likes based on
 - His/Her historical behavior
 - Aggregate behavior of people similar to him

Customers Who Bought This Item Also Bought



[Pattern Recognition and Machine Learning...](#) by Christopher M. Bishop

★★★★☆ (50)

\$76.10



[The Elements of Statistical Learning: Data Minin...](#) by Trevor Hastie

★★★★☆ (38)

\$71.96



[Pattern Classification \(2nd Edition\)](#) by Richard O. Duda

★★★★☆ (29)

\$88.52



Mahout Recommenders

- Different types of recommenders
 - User based
 - Item based
- Full framework for storage, online online and offline computation of recommendations
- Like clustering, there is a notion of similarity in users or items
 - Cosine, Tanimoto, Pearson and LLR

Customers Who Bought This Item Also Bought



Recommended Tweets

“Discover interesting tweets without Re-Tweeting or Replying”



Frequent Pattern Mining

- Find interesting groups of items based on how they co-occur in a dataset



Mahout Parallel FPGrowth

- Identify the most commonly occurring patterns from
 - Sales Transactions
buy “Milk, eggs and bread”
 - Query Logs
ipad -> apple, tablet, iphone
 - Spam Detection
Yahoo! <http://www.slideshare.net/hadoopusergroup/mail-antispam>



Frequent patterns in Tweets

“Identify groups of words that occur together”

Or

“Identify related searches from search logs”



Mahout is Evolving

- Mapreduce enabled fitness functions for Genetic programming
 - Integration with Watchmaker
 - Solves: Travelling salesman, class discovery and many others
- Singular Value decomposition [SVD] of large matrices
 - Reduce a large matrix into a smaller one by identifying the key rows and columns and discarding the others
 - Mapreduce implementation of Lanczos algorithm



Agenda

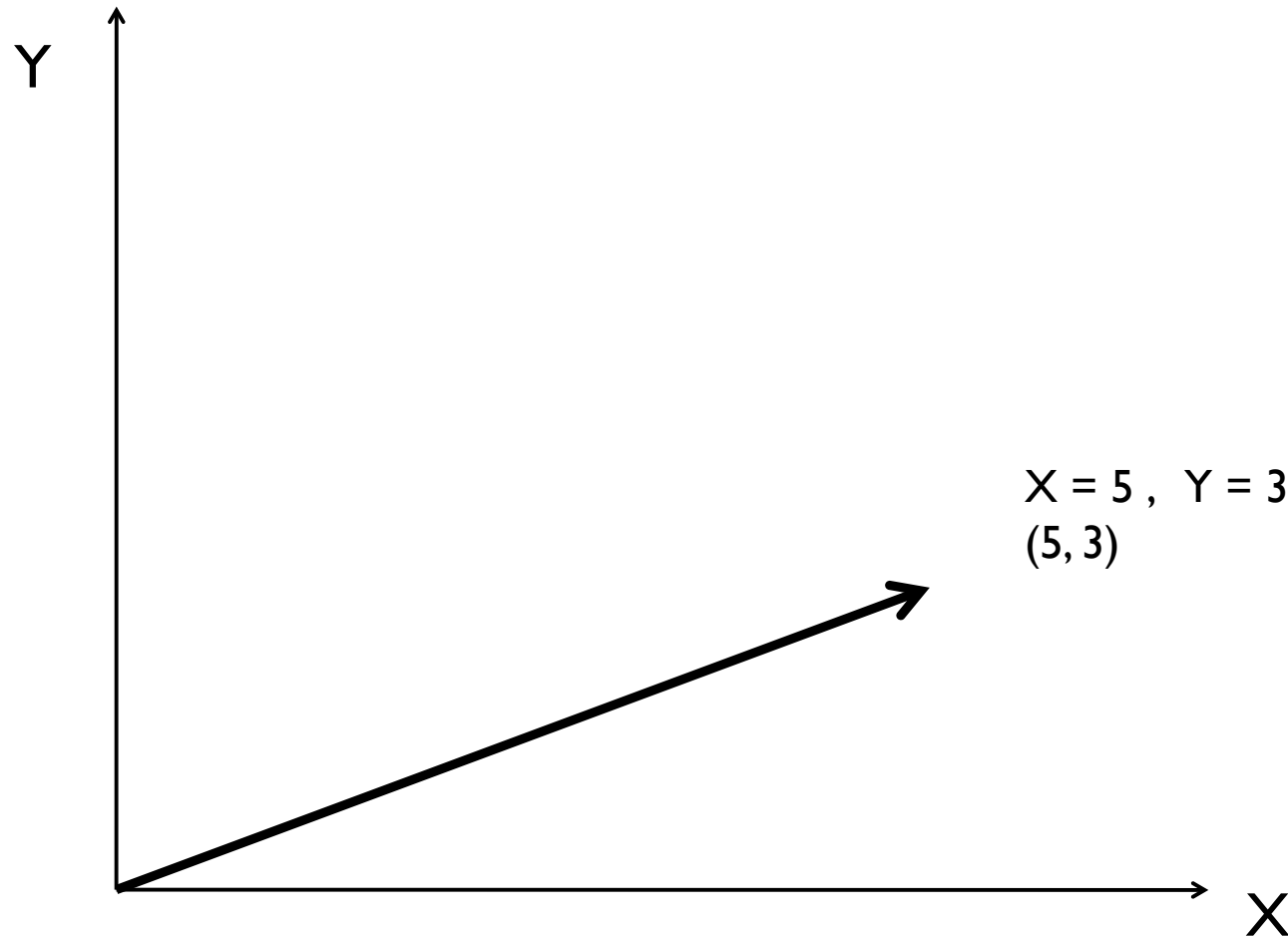
- ~~Introducing Mahout~~
- ~~Different classes of problems~~
- ~~And their Mahout based solutions~~
- **Basic data structure**
- Usage examples
- Sneak peek at our Next Release



Vector



Representing Data as Vectors



- The vector denoted by point (5, 3) is simply `Array([5, 3])` or `HashMap([0 => 5], [1 => 3])`



Representing Vectors – The basics

- Now think 3, 4, 5, n-dimensional
- Think of a document as a bag of words.

“she sells sea shells on the sea shore”

- Now map them to integers

she => 0

sells => 1

sea => 2

and so on

- The resulting vector [1.0, 1.0, 2.0, ...]



Vectorizer tools

- Map/Reduce tools to convert text data to vectors
 - Use collate multiple words (n-grams) eg: “San Francisco”
 - Normalization
 - Optimize for sequential or random access
 - TF-IDF calculation
 - Pruning
 - Stop words removal



Agenda

- ~~Introducing Mahout~~
- ~~Different classes of problems~~
- ~~And their Mahout based solutions~~
- ~~Basic data structure~~
- **Usage examples**
- Sneak peek at our Next Release



How to use mahout

- Command line launcher **bin/mahout**
- See the list of tools and algorithms by running **bin/mahout**
- Run any algorithm by its shortname:
 - `bin/mahout kmeans -help`
- By default runs locally
- `export HADOOP_HOME = /pathto/hadoop-0.20.2/`
 - Runs on the cluster configured as per the conf files in the hadoop directory
- Use driver classes to launch jobs:
 - `KMeansDriver.runjob(Path input, Path output ...)`



Clustering Walkthrough (tiny example)

- Input: set of text files in a directory
- Download Mahout and unzip
 - `mvn install`
 - `bin/mahout seqdirectory -i <input> -o <seq-output>`
 - `bin/mahout seq2sparse -i seq-output -o <vector-output>`
 - `bin/mahout kmeans -i<vector-output>`
`-c <cluster-temp> -o <cluster-output> -k 10 -cd`
`0.01 -x 20`



Clustering Walkthrough (a bit more)

- Use bigrams: `-ng 2`
- Prune low frequency: `-s 10`
- Normalize: `-n 2`

- Use a distance measure : `-dm`
`org.apache.mahout.common.distance.CosineDistanceMeasure`



Clustering Walkthrough (viewing results)

- `bin/mahout clusterdump`
 - `-s cluster-output/clusters-9/part-00000`
 - `-d vector-output/dictionary.file-*`
 - `-dt sequencefile -n 5 -b 100`

- Top terms in a typical cluster

<code>comic</code>	<code>=></code>	<code>9.793121272867376</code>
<code>comics</code>	<code>=></code>	<code>6.115341078151356</code>
<code>con</code>	<code>=></code>	<code>5.015090566692931</code>
<code>sdcc</code>	<code>=></code>	<code>3.927590843402978</code>
<code>webcomics</code>	<code>=></code>	<code>2.916910980686997</code>



Agenda

- ~~Introducing Mahout~~
- ~~Different classes of problems~~
- ~~And their Mahout based solutions~~
- ~~Basic data structure~~
- ~~Usage examples~~
- **Sneak peek at our Next Release**



Mahout 0.4 (trunk)

- New breed of classifiers:
 - Stochastic Gradient Descent (SGD)
 - Pegasos SVM (Order of magnitude faster than SVM Perf)
 - Lib Linear (Winner, ICML 2008)
- New Recommenders:
 - Restricted Boltzmann Machine (RBM) based recommender
 - SVD++ recommender
- New Clustering algorithms:
 - Spectral Clustering
 - K-Means++
- Full Hadoop 0.20 API compliance and performance improvements



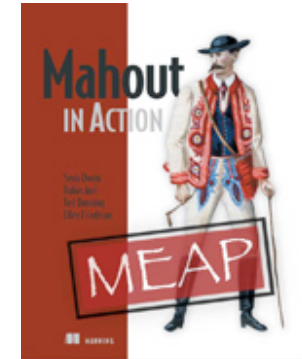
Get Started

- <http://mahout.apache.org>
- dev@mahout.apache.org - Developer mailing list
- user@mahout.apache.org - User mailing list
- Check out the documentations and wiki for quickstart
- <http://svn.apache.org/repos/asf/mahout/trunk/> Browse Code

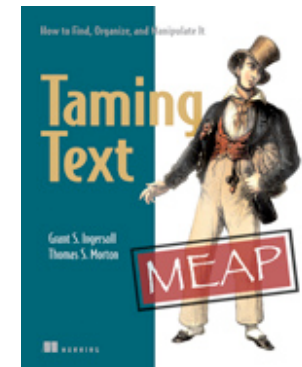


Resources

- “Mahout in Action” Owen, Anil, Dunning, Friedman
<http://www.manning.com/owen>



- “Taming Text” Ingersoll, Morton, Farris
<http://www.manning.com/ingersoll>



- “Introducing Apache Mahout”
<http://www.ibm.com/developerworks/java/library/j-mahout/>



Thanks to

- Apache Foundation
- Mahout Committers
- Google Summer of Code Organizers
- And Students
- OSCON
- Open source!



References

- news.google.com
- Cat <http://www.flickr.com/photos/gattou/3178745634/>
- Dog <http://www.flickr.com/photos/30800139@N04/3879737638/>
- Milk Eggs Bread
<http://www.flickr.com/photos/nauright/4792775946/>
- Amazon Recommendations
- twitter

