

Introduction to Machine Learning

September 2014 Meetup

Rahul Jain



@rahuldausa

Join us @ For Solr, Lucene, Elasticsearch, Machine Learning, IR

<http://www.meetup.com/Hyderabad-Apache-Solr-Lucene-Group/>

<http://www.meetup.com/DataAnalyticsGroup/>

Join us @ For Hadoop, Spark, Cascading, Scala, NoSQL, Crawlers and all cutting edge technologies.

<http://www.meetup.com/Hyderabad-Programming-Geeks-Group/>

Agenda

- Introduction
- Basics
- Classification
- Clustering
- Regression
- Use-Cases

Quick Questionnaire

How many people have *heard* about Machine Learning

How many people *know* about Machine Learning

How many people are *using* Machine Learning

About

- subfield of Artificial Intelligence (AI)
- name is derived from the concept that it deals with
“construction and study of systems that can learn from data”
- can be seen as building blocks to make computers learn to behave more intelligently
- It is a theoretical concept. There are various *techniques* with various *implementations*.
- http://en.wikipedia.org/wiki/Machine_learning

In other words...

“A computer program is said to learn from experience (E) with some class of tasks (T) and a performance measure (P) if its performance at tasks in T as measured by P improves with E”

Terminology

- Features
 - The number of features or distinct traits that can be used to describe each item in a quantitative manner.
- Samples
 - A sample is an item to process (e.g. classify). It can be a document, a picture, a sound, a video, a row in database or CSV file, or whatever you can describe with a fixed set of quantitative traits.
- Feature vector
 - is an n -dimensional vector of numerical features that represent some object.
- Feature extraction
 - Preparation of feature vector
 - transforms the data in the high-dimensional space to a space of fewer dimensions.
- Training/Evolution set
 - Set of data to discover potentially predictive relationships.

Let's dig deep into it...

What do you mean by

Apple

Learning (Training)



Features:

1. Color: **Radish/Red**
 2. Type : **Fruit**
 3. Shape
- etc...



Features:

1. Sky Blue
 2. **Logo**
 3. Shape
- etc...

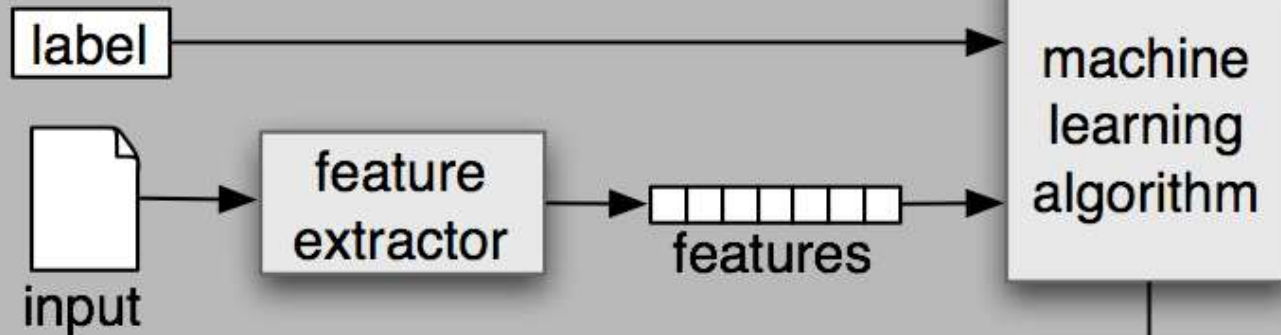


Features:

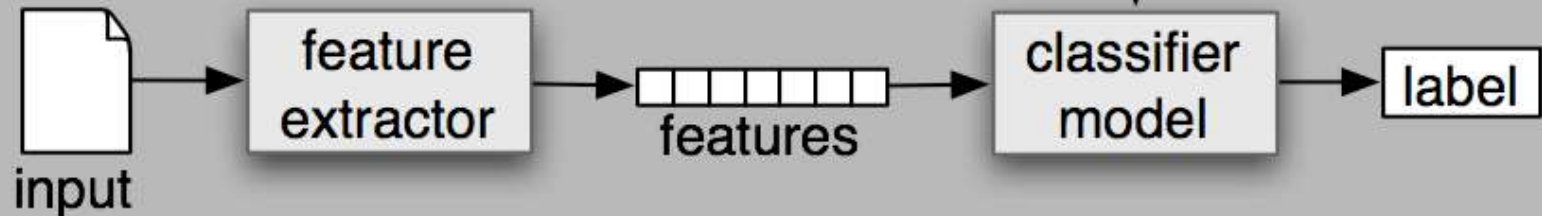
1. **Yellow**
 2. **Fruit**
 3. Shape
- etc...

Workflow

(a) Training



(b) Prediction

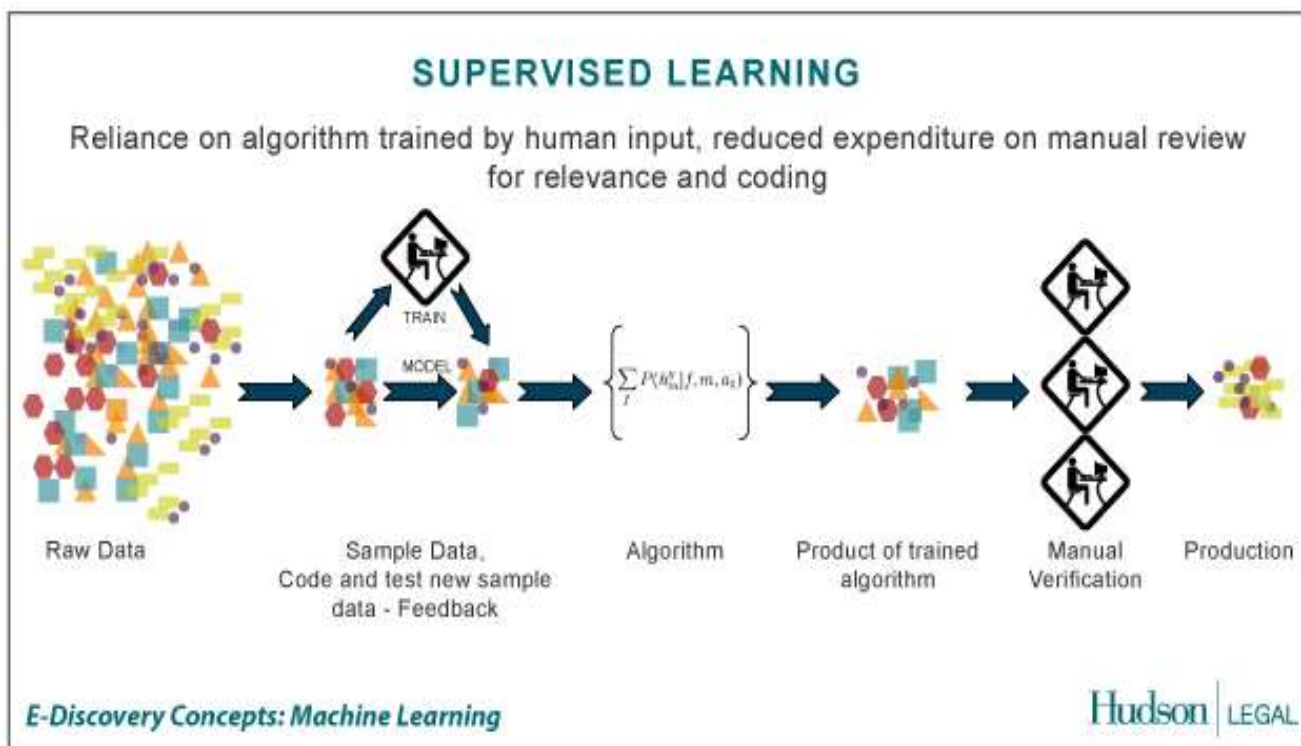


Categories

- Supervised Learning
- Unsupervised Learning
- Semi-Supervised Learning
- Reinforcement Learning

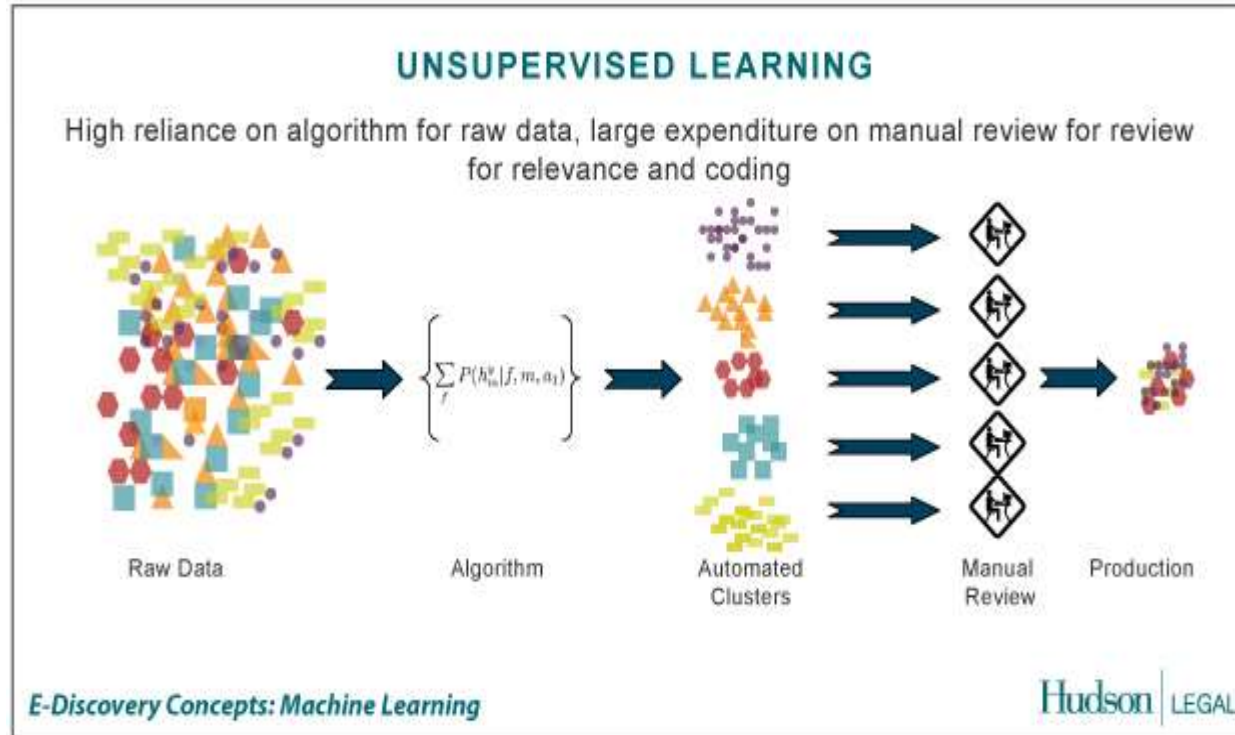
Supervised Learning

- the correct classes of the training data are known



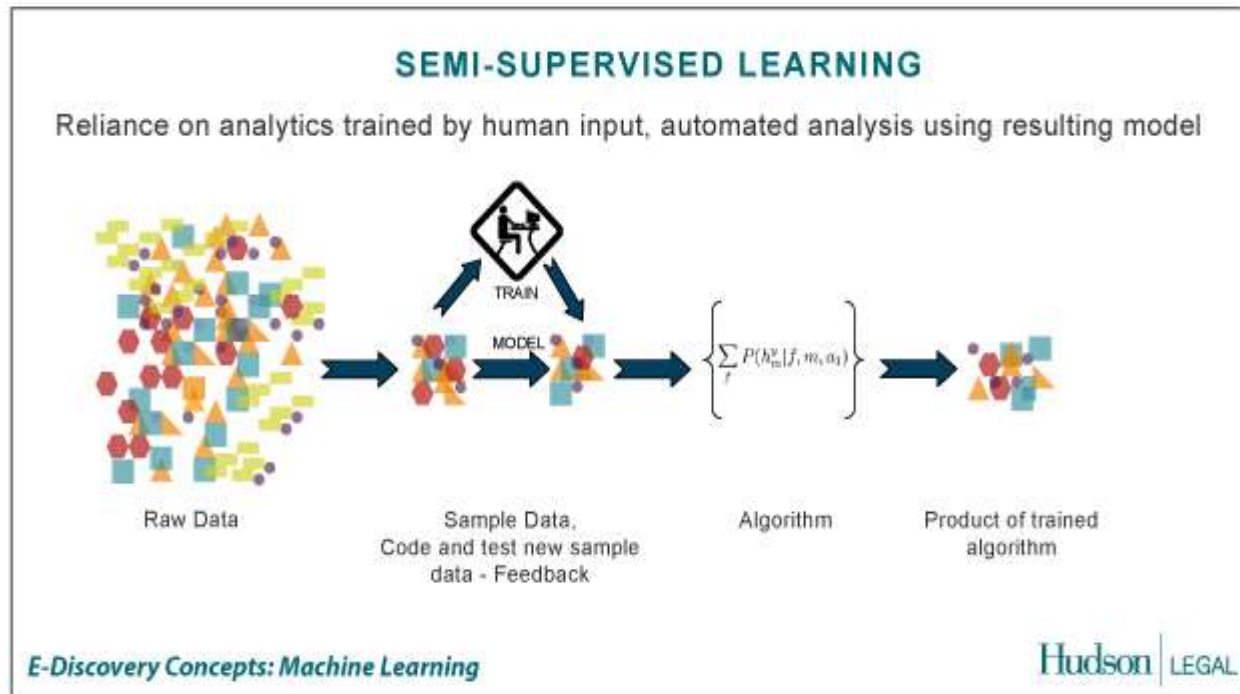
Unsupervised Learning

- the correct classes of the training data are not known



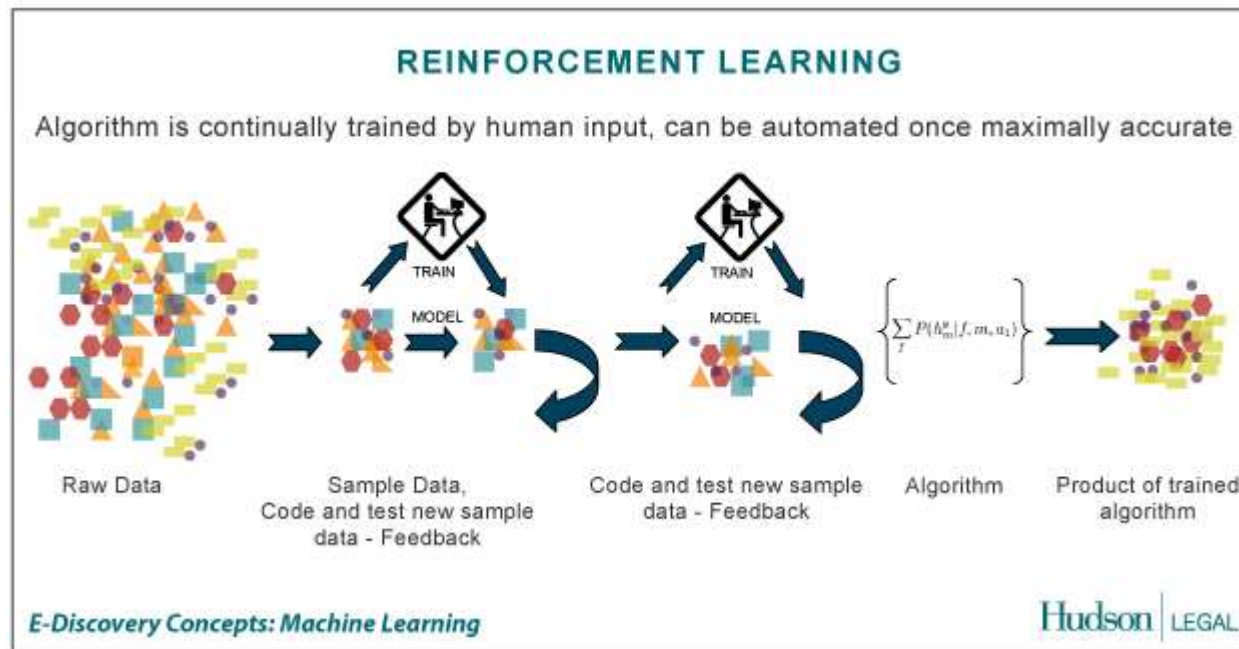
Semi-Supervised Learning

- A Mix of Supervised and Unsupervised learning



Reinforcement Learning

- allows the machine or software agent to learn its behavior based on feedback from the environment.
- This behavior can be learnt once and for all, or keep on adapting as time goes by.



Machine Learning Techniques

Techniques

- **classification**: predict class from observations
- **clustering**: group observations into “meaningful” groups
- **regression (prediction)**: predict value from observations

Classification

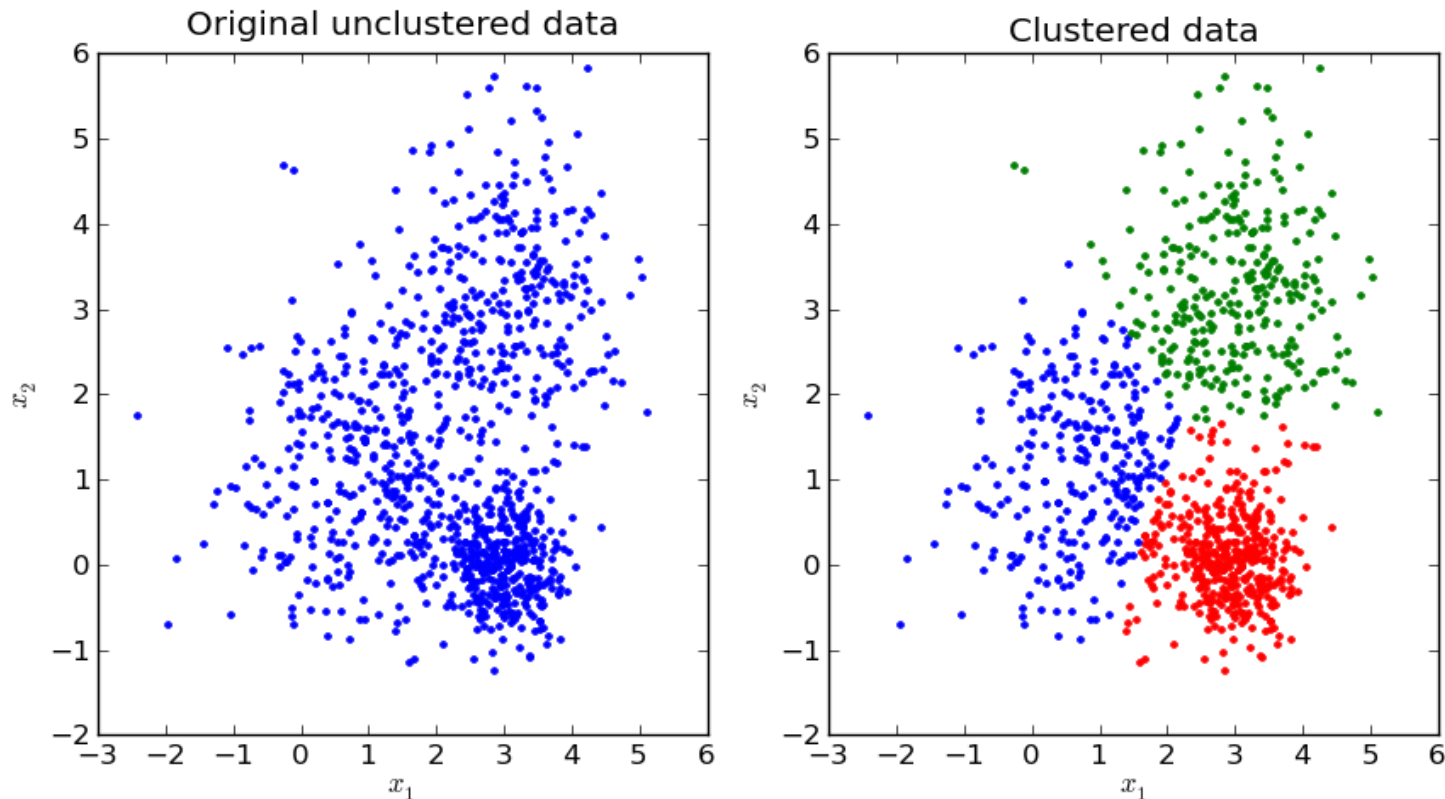
- classify a document into a predefined category.
- documents can be text, images
- Popular one is Naive Bayes Classifier.
- Steps:
 - Step1 : Train the program (Building a Model) using a training set with a category for e.g. sports, cricket, news,
 - Classifier will compute probability for each word, the probability that it makes a document belong to each of considered categories
 - Step2 : Test with a test data set against this Model
- http://en.wikipedia.org/wiki/Naive_Bayes_classifier

Clustering

- **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 to each other
- objects are not predefined
- For e.g. these keywords
 - “man’s shoe”
 - “women’s shoe”
 - “women’s t-shirt”
 - “man’s t-shirt”
 - can be cluster into 2 categories “shoe” and “t-shirt” or “man” and “women”
- Popular ones are **K-means clustering** and **Hierarchical clustering**

K-means Clustering

- partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.
- http://en.wikipedia.org/wiki/K-means_clustering



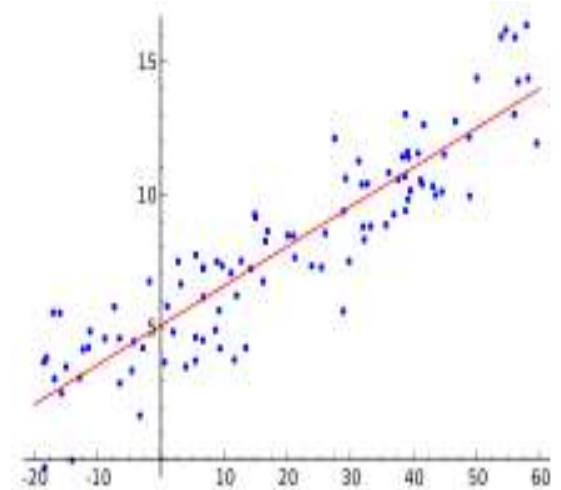
<http://pypr.sourceforge.net/kmeans.html>

Hierarchical clustering

- method of cluster analysis which seeks to build a hierarchy of clusters.
- There can be two strategies
 - **Agglomerative:**
 - This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
 - Time complexity is $O(n^3)$
 - **Divisive:**
 - This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.
 - Time complexity is $O(2^n)$
- http://en.wikipedia.org/wiki/Hierarchical_clustering

Regression

- is a measure of the relation between the mean value of one variable (e.g. output) and corresponding values of other variables (e.g. time and cost).
- **regression analysis** is a statistical process for estimating the relationships among variables.
- Regression means to **predict** the output value using training data.
- Popular one is Logistic regression (binary regression)
- http://en.wikipedia.org/wiki/Logistic_regression



Classification vs Regression

- Classification means to group the output into a class.
- classification to **predict** the type of tumor i.e. harmful or not harmful using training data
- if it is discrete/categorical variable, then it is classification problem
- Regression means to predict the output value using training data.
- regression to **predict** the house price from training data
- if it is a real number/continuous, then it is regression problem.

Let's see the usage in Real life

Use-Cases

- Spam Email Detection
- Machine Translation (Language Translation)
- Image Search (Similarity)
- Clustering (KMeans) : Amazon Recommendations
- Classification : Google News

continued...

Use-Cases (contd.)

- Text Summarization - Google News
- Rating a Review/Comment: Yelp
- Fraud detection : Credit card Providers
- Decision Making : e.g. Bank/Insurance sector
- Sentiment Analysis
- Speech Understanding – iPhone with Siri
- Face Detection – Facebook's Photo tagging

Classification in Action
isn't it easy?

it's not (Snapshot of Spam folder)

Delete all spam messages now (messages that have been in Spam more than 30 days will be automatically deleted)

<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	HDFC Bank	LOAN upto Rs 25 lac - Disbursal in 2 days - open in fresh tab -- If you do not want to receive any more newsletters, please click here	9:40 pm
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	iEntry	Welcome iEntry Member - Ultimate Guide To Assessing	9:23 pm
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	New-Zealand-Jobs.067L	Come to New Zealand to find a great job and settle here (Search for all Jobs from diffe... - Search for all Jobs from different kinds of industries Find a Job in Enchanting	8:18 pm
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	CarSizzler	Assured Free Luxurious Ride worth Rs.300 with Uber Cabs - Home Home Buy New Car Buy New Car Sell Car Sell Car Tech Tics Tip & Tale Facebook 41727 others	6:05 pm
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	Supermarket Promotion	Enjoy Rs.1700 voucher valid at any supermarket! - If you are unable to view this mailer Click here HOW TO CONTACT US? BY EMAIL: support@savethedeals.in	4:51 pm
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	Entireweb Newsletter	Hire an SEO the Right Way -- 6 Tips You Must Remember for Life - Unsubscribe me View web version Become a fan on Facebook Follow us on Twitter September 5th, 21	1:24 pm
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	Max Bupa	A policy that understands your family's medical need - open in fresh tab - If you do not want to receive any more newsletters, please	11:08 am
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	Scoop.it	Your Scoop.it Daily Summary - How to Maximize Your LinkedIn Publishing Exposure SME a... - Scoop.it Facebook Twitter G+ H	9:30 am
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	standard charterer Bank	Instant approval on your Credit Card	7:27 am
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	CAR TRADE	Sell your car at no cost at all - If you are having trouble viewing this email, view web version View this message in your mobile	Sep 4
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	Uday	VPS Web Hosting Services Provider - Dear Sir, I am Uday Sharma, Business development executive. We are providing quality VPS hosting for	Sep 4
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	Mark Regan, SPN	How to Find Your Most Valuable Keywords [Free Guide] - This is a SiteProNews/ExactSeek Webmaster Exclusive Mailing! To drop your subscription, use the link	Sep 4
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	HDFC Bank	LOAN upto Rs 25 lac - Disbursal in 2 days - open in fresh tab You have received this mailer from Shop@Best on behalf of HDFC Bank because you	Sep 4
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	CAR TRADE	Sell your car at no cost at all - open in fresh tab -- If you do not want to receive any further newsletters, please click here	Sep 4
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	ICICI Bank	Home Loan Interest Rate starting from 10.15%*. Get Instant Approval! - open in fresh tab -- If you do not want to receive any more newsletters, please Click Here	Sep 4
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	calculateyourwealth	It's good when your bank helps you manage your wealth and fulfill your ambitions - Calculate Now Dreams you wish to realize in your lifetime require enough wealth. C	Sep 4
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	Angel Broking	Get Low Brokerage - Free Demat & Trading Account - open in fresh tab -- If you do not want to receive any further newsletters, please click here	Sep 4
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	Bankbazaar	7 Minute Instant Online Approval for your PESONAL LOAN - Now get instant online Personal Loan approval in 7 minutes by BankBazaar.com from leading Banks in	Sep 4
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	Jayde	Welcome To The Jayde Newsletter! - Welcome To WebProNews Welcome To The Jayde Newsletter! Before we begin, make sure to add	Sep 4
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	ineedhits noreply	[ineedhits] Your ineedhits Account and Password - ACCOUNT CREATION Account ID : A1588368 Dear Rah, Welcome to ineedhits. Yo	Sep 4
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	Rekha	Mobility Apps for Your Business - While we look at the span of last 20 years, we could broadly look at two distinct eras, - Life in	Sep 4
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	SlideShare Newsletter	Top Tips From the World Champions of PowerPoint - View online version Remember to display images Meet the PowerPoint World Champs Top Tips From the	Sep 4
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	Dilshad Pathan	Feeling Hesitate to Discuss personal Health Queries - My Life Care Follow Us on facebook twitter linkedin Google+ Feeling Hesitate to Discuss personal	Sep 4
<input type="checkbox"/>	<input type="star"/>	<input type="trash"/>	Vaishu	TAKE YOUR PICK. Register in SimplyMarry - TAKE YOUR PICK. Register in SimplyMarry -- Regards Vaishu	Sep 4

Not a
Spam

Not a
Spam

NER (Named Entity Recognition)

Stanford Named Entity Tagger

Classifier:

Output Format:

Preserve Spacing:

Please enter your text here:

When Mike Brannigan was 18 months old, he was diagnosed with autism. At the time, his doctors said he would likely need a special school and a group home. His mom, Edie, admits she thought he'd "never be able to function in the world." Fast-forward several years. Brannigan is now 17, and is a senior at Northport High School, a public school in Long Island, New York. He's doing well academically, he has friends -- and he also happens to be one of the best young athletes in the country. Continue Reading...

When **Mike Brannigan** was 18 months old, he was diagnosed with autism. At the time, his doctors said he would likely need a special school and a group home. His mom, **Edie**, admits she thought he'd "never be able to function in the world." Fast-forward several years. Brannigan is now 17, and is a senior at **Northport High School**, a public school in **Long Island, New York**. He's doing well academically, he has friends -- and he also happens to be one of the best young athletes in the country. Continue Reading...

Potential tags:

LOCATION

TIME

PERSON

ORGANIZATION

MONEY

PERCENT

DATE

<http://nlp.stanford.edu:8080/ner/process>

Similar/Duplicate Images

About 81 results (0.70 seconds)



Image size:
250 × 321

No other sizes of this image found.

Best guess for this image: *taj mahal*

Visually similar images

Report images



Remember

Features ?

(Feature Extraction)

Can be :

- Width
- Height
- Contrast
- Brightness
- Position
- Hue
- Colors

Check this :

LIRE (Lucene Image REtrieval)
library -

<https://code.google.com/p/lire/>

Recommendations

More Items to Consider

You looked at



You might also consider



JavaScript: The Good Parts Paperback by Douglas Crockford
~~\$29.99~~ **\$19.79**

JavaScript: The Definitive Guide Paperback by David Flanagan
~~\$49.99~~ **\$31.49**

CSS: The Missing Manual Paperback by David McFarland
~~\$34.99~~ **\$23.09**

Learning jQuery 1.3 Paperback by John Resig
~~\$39.99~~ **\$35.99**

[Find similar items](#)

Related Items You've Viewed

You looked at



You might also consider



Forms that Work: Designing Web Forms for Usability Paperback by Steve Krug
~~\$29.99~~ **\$19.79**




Don't Make Me Think: A Common Sense Approach to Web Usability Paperback by Steve Krug
~~\$49.99~~ **\$31.49**

Letting Go of the Words: Writing Web Content that Works Paperback by Linda Ward Beech
~~\$34.99~~ **\$23.09**

Designing Web Interfaces: Principles for Creating Great User Interfaces Paperback by Steve Krug
~~\$39.99~~ **\$35.99**

Today's Recommendations For You

Here's a daily sample of items recommended for you. Click here to [see all recommendations](#)



Even Faster Web Sites: Performance Tuning Techniques (Paperback) by Steve Souders
★★★★★ (7) \$23.10
[Fix this recommendation](#)

Simply JavaScript (Paperback) by Kevin Yank
★★★★★ (19) \$26.37
[Fix this recommendation](#)

The Art & Science of JavaScript (Paperback) by John Resig
★★★★★ (3)
[Fix this recommendation](#)

Any Category Algorithms Boxed Sets Business & Culture Java
Graphic Design Microsoft Networking Networks, Protocols & APIs New SQL

Popular Frameworks/Tools

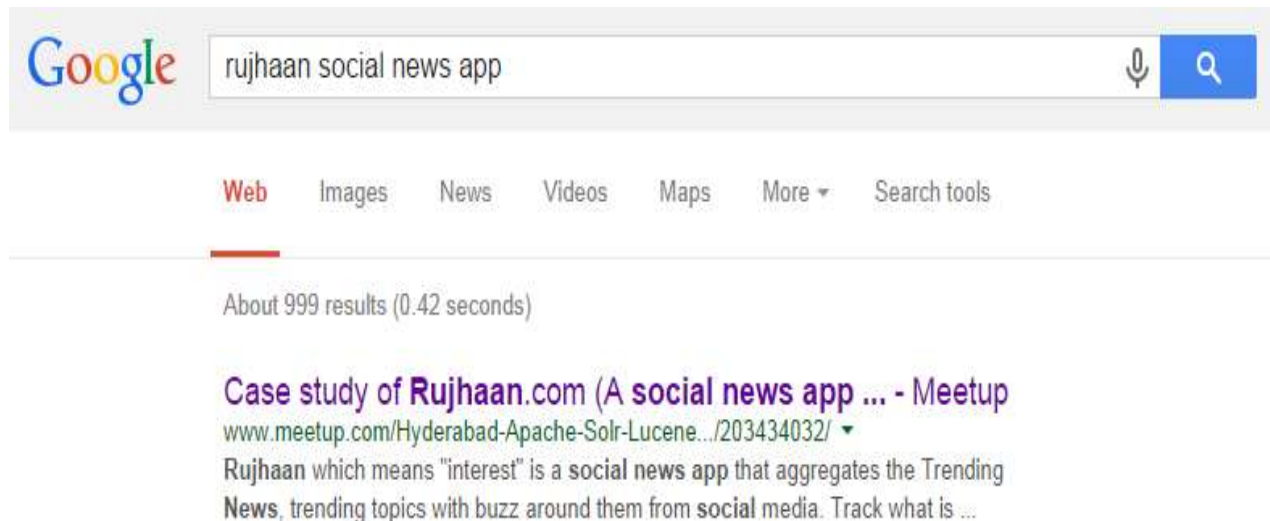
- Weka
- Carrot2
- Gate
- OpenNLP
- LingPipe
- Stanford NLP
- Mallet – Topic Modelling
- Gensim – Topic Modelling (Python)
- Apache Mahout
- MLib – Apache Spark
- scikit-learn - Python
- LIBSVM : Support Vector Machines
- and many more...

Advanced concepts (related to IR)

- Topic Modelling
- Latent Dirichlet allocation (LDA)
- Latent semantic analysis (LSA/LSI) - Semantic Search
- Singular Value Decomposition (SVD)
- Summarization (without Training)

Solr/Lucene Meetup

- Case study of [Rujhaan.com](http://www.meetup.com/Hyderabad-Apache-Solr-Lucene-Group/events/203434032/)
(A social news app)
 - Saturday, Sep 27, 2014 10:00 AM
 - IIIT Hyderabad
 - URL: <http://www.meetup.com/Hyderabad-Apache-Solr-Lucene-Group/events/203434032/>
- OR**
- Search on Google ...



Topics of Talk

- ☐ Crawler(Crawler4j)
 - ☐ MongoDB
 - ☐ Solr
 - ☐ Nginx, ApacheTomcat
 - ☐ Redis
 - ☐ Machine Learning
1. Classification - Classification of News, Tweets - Lingpipe
 2. Clustering, - Similar Items - carrot2 (Near Future: Hadoop and Apache Spark)
 3. Summarization - Extracting the main text with Automatic Summary of article
 4. Topics Extraction from text

Questions ?

Thanks!

@rahuldausa on twitter and slideshare
<http://www.linkedin.com/in/rahuldausa>

Interested in Search/Information Retrieval ?

Join us @ <http://www.meetup.com/Hyderabad-Apache-Solr-Lucene-Group/>