

Advances & Innovations in Artificial Intelligence

Making Sense of Textual Data



Dr Vivek Kumar Singh
Professor & Head
Department of Computer Science
Banaras Hindu University, Varanasi, India

<http://www.viveksingh.in>

Outline

- Contextual Settings
- Machine Learning
- Text Representation
- ML for Text Processing
- Sentiment Analysis

Some Jobs in the area in India

Research Scientist/Sr. Research Scientist/Sr. Research Engineer: Machine Learning Xerox - Bangalore, IN

- The **Machine Learning & Statistics** group at Xerox Research Center India (XRCI) is seeking researchers with strong technical expertise in the area of Machine Learning and Data Mining to participate in exciting research and technology development projects. Applicants should have a strong background in the general domain of Machine Learning with deep expertise in one or more of text & image mining, graph mining & structure analysis, statistical relational learning, and large scale statistical inference.
- **The Ideal Candidate For The Research Scientist Position Will Have** A PhD (or masters with relevant experience) in Computer Science, Information Systems, Applied Statistics, or related disciplines with sound technical expertise in one or more of the above mentioned areas. Demonstrated ability to formulate scientifically challenging problems and develop and implement novel solutions

Machine Learning/Research Scientist

Amazon - Bangalore

- **Job description**

- The Ad Optimization team in Bangalore is a part of the Advertising Technology group and has the charter to solve optimization problems for ad-programs in Amazon. The Advertising Technology group in Amazon powers online advertising programs for some of the world's largest websites, including Amazon.com and other prime online properties. We supply the technology to show the right ad to the right customer at the right time. Computational Advertising is one of the most challenging areas for algorithmic optimization due to the scale of the problem, direct impact on the business and because of the interplay of multiple areas like machine learning, statistics, data mining, data streams, computational economics and econometrics. We build advanced and highly scalable algorithms to optimize performance for advertiser, publisher, ad-network and user. We are currently focused on solving optimization problems in the areas of traffic quality prediction, fraud and spam detection, bid pricing, performance optimization and attribution.

A Machine Learning Scientist is responsible for solving complex big-data problems in online advertising space using data mining, machine learning, statistical analysis and computational economics. An ideal candidate should have strong depth and breadth knowledge in machine learning, data mining and statistics. The candidate should have reasonable programming and design skills to manipulate unstructured and big data and build prototypes that work on massive datasets. The candidate should be able to apply business knowledge to perform broad data analysis as a precursor to modeling and to provide valuable business intelligence.

- **Desired Skills and Experience**

- - PhD in Machine Learning, Statistics, Optimization or Applied Mathematics
 - Relevant industry or research experience
 - Hands-on experience in predictive modeling and analysis
 - Strong skills in problem solving, programming and computer science fundamentals

Researcher / Computer Scientist (Only PhD) – Adobe Research Labs Adobe - Bangalore

- ...The positions will leverage your expertise in the following set of areas and related areas:
- Analytics & Data mining (Web, Social and Big data)
- Machine Learning and Pattern Recognition
- Natural Language Processing & Computational Linguistics
- Statistical Modelling and Inferencing
- Information Retrieval
- Large Scale Distributed Systems & Cloud Computing
- Econometrics and Quantitative Marketing
- Applied Game Theory and Mechanism Design
- Operations Research and Optimization
- Information Visualization
- **Desired Skills and Experience**
- Demonstrated research excellence.
- **PhD in Computer Science, Computer Engineering, Electrical Engineering, Statistics, Economics, Mathematics or other related quantitative disciplines .**

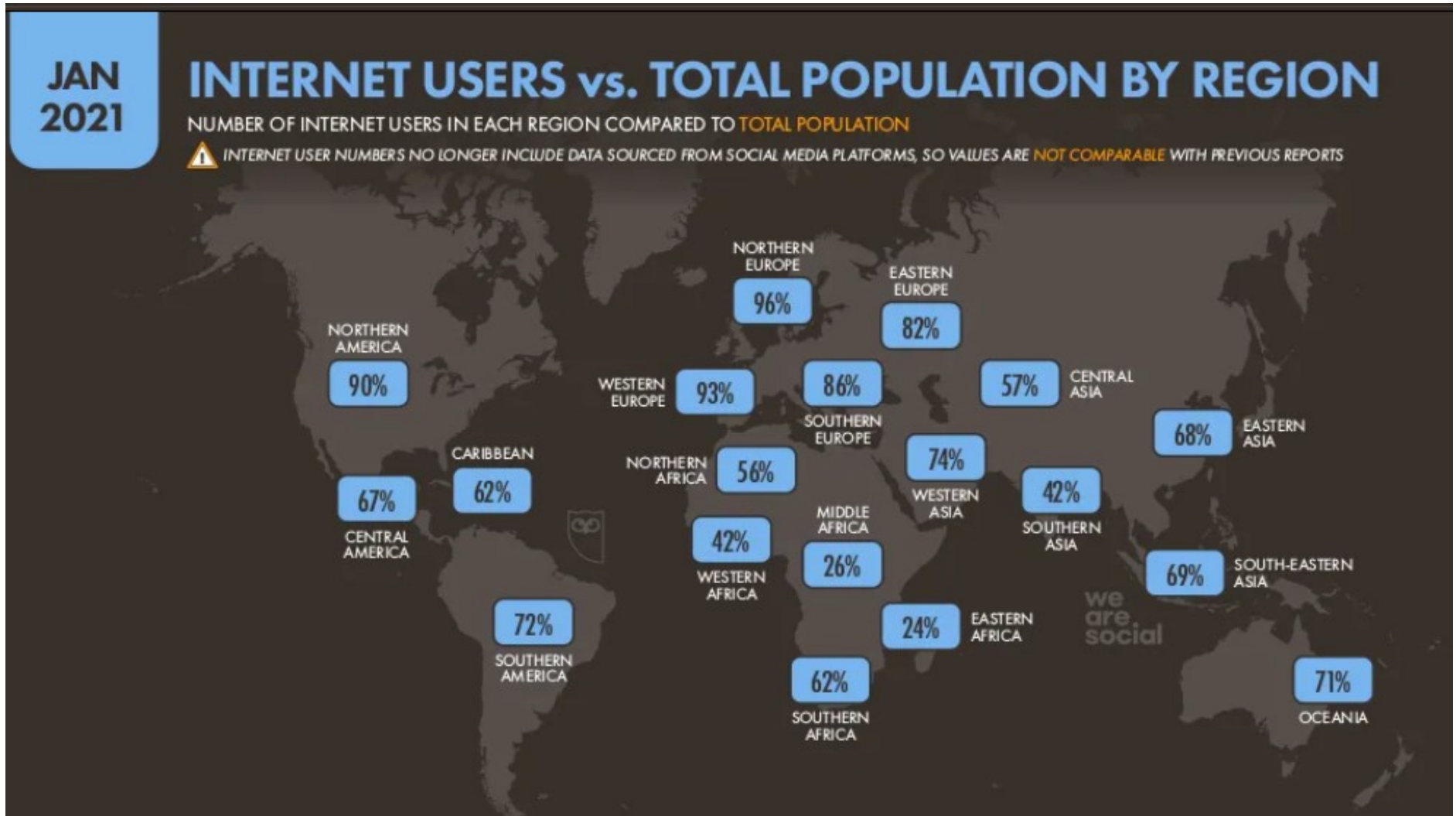
Contextual Settings

Internet & Social Media



(Source: wearesocial.com)

Internet Penetration



(Source: wearesocial.com)

Social Media



- the interaction among people in which they create, share or exchange information and ideas in virtual communities and networks.
- a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content.

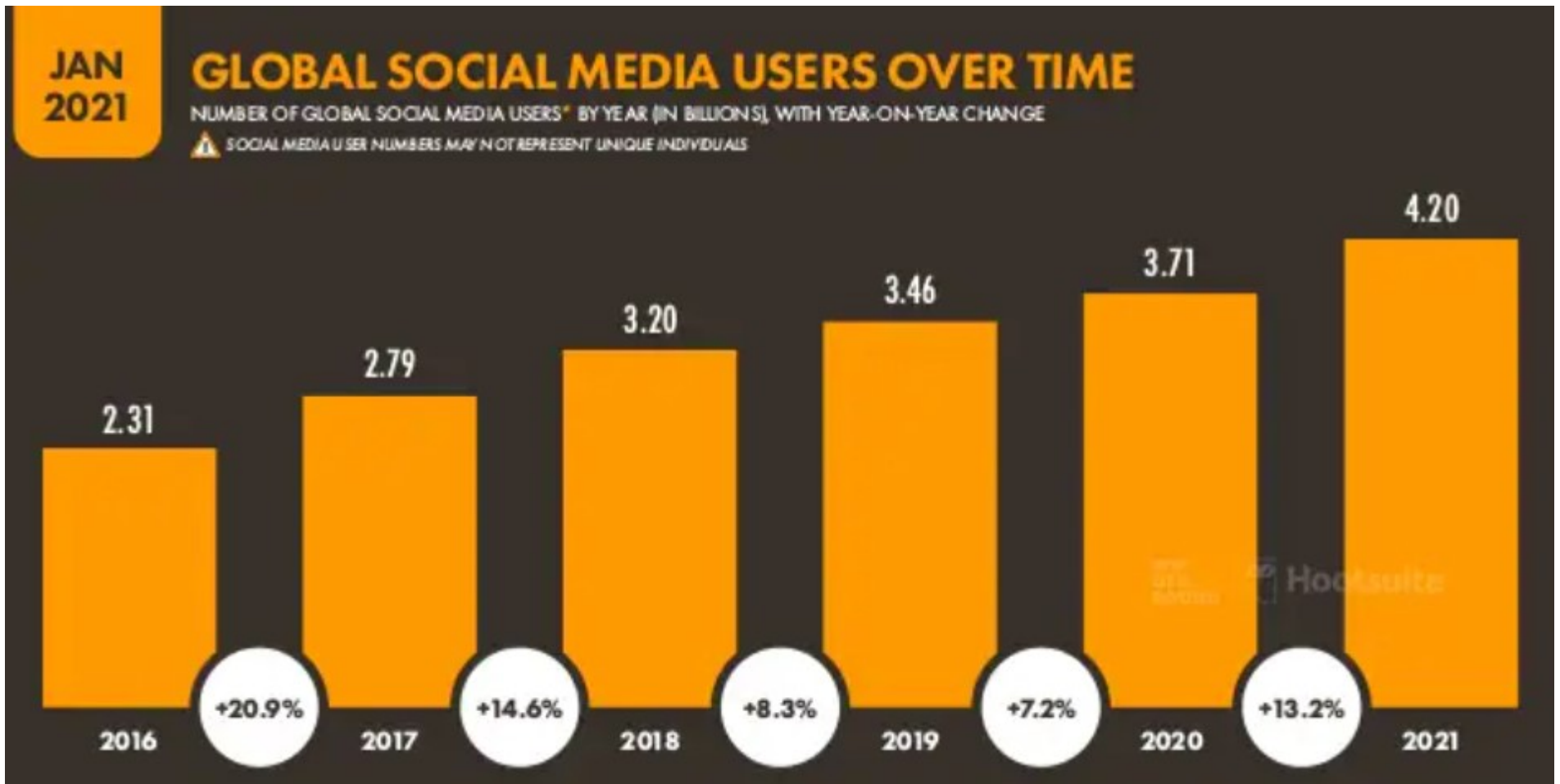
(Source: http://en.wikipedia.org/wiki/Social_media)

Social Media (Contd...)

By applying a set of theories in the field of media research and social processes Kaplan and Haenlein created a classification scheme in their *Business Horizons* (2010) article, with seven different types of social media:

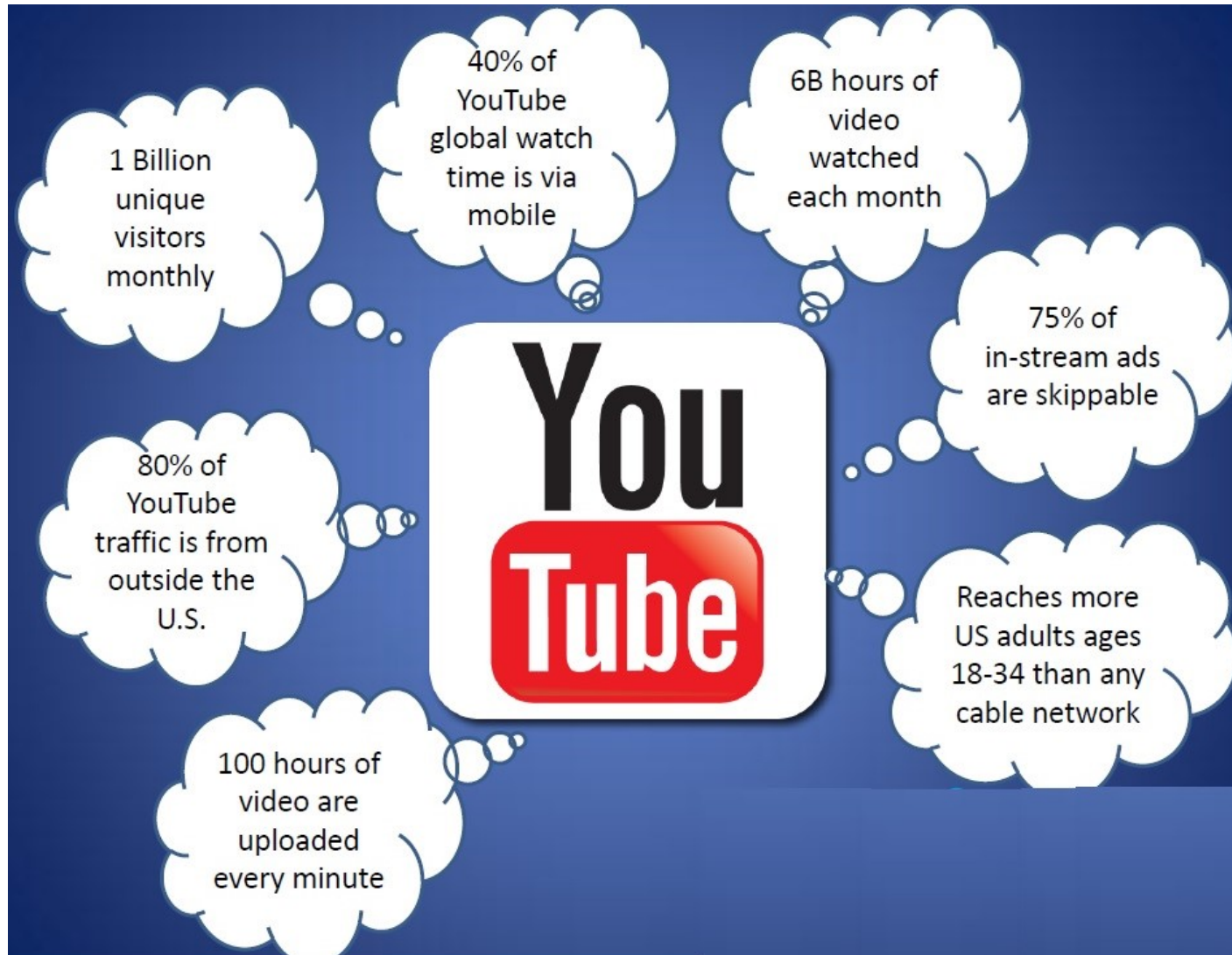
- collaborative projects (for example, [Wikipedia](#))
- blogs and microblogs (for example, [Twitter](#))
- social news networking sites (for example, Digg and Leakernet)
- content communities (for example, [YouTube](#) and [DailyMotion](#))
- social networking sites (for example, [Facebook](#))
- virtual game-worlds (e.g., [World of Warcraft](#))
- virtual social worlds (e.g. [Second Life](#))

Social Media Penetration



(Source: wearesocial.com)

Social Media Statistics- YouTube

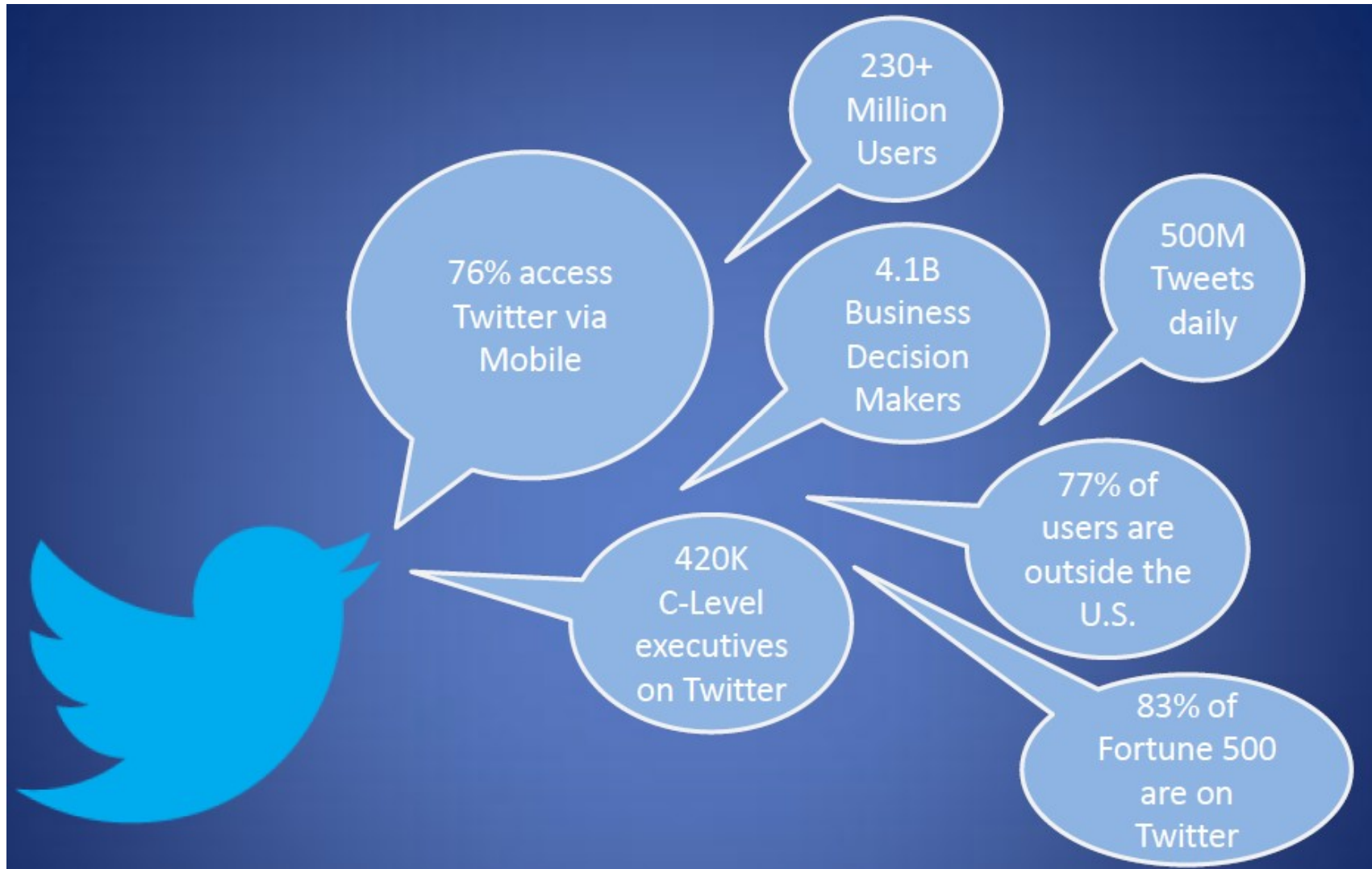


Social Media Statistics- Facebook

JAN 2021				FACEBOOK REACH RANKINGS			
				COUNTRIES AND TERRITORIES* WITH THE LARGEST FACEBOOK ADVERTISING AUDIENCES			
#	COUNTRY / TERRITORY	REACH	▲QOQ	#	COUNTRY / TERRITORY	REACH	▲QOQ
01	INDIA	320,000,000	+3.2%	11	PAKISTAN	40,000,000	+2.6%
02	U.S.A.	190,000,000	0%	12=	TURKEY	38,000,000	+2.7%
03	INDONESIA	140,000,000	0%	12=	U.K.	38,000,000	0%
04	BRAZIL	130,000,000	0%	14	COLOMBIA	36,000,000	0%
05	MEXICO	93,000,000	+1.1%	15	FRANCE	33,000,000	+3.1%
06	PHILIPPINES	83,000,000	+2.5%	16=	ARGENTINA	31,000,000	0%
07	VIETNAM	68,000,000	+4.6%	16=	ITALY	31,000,000	+3.3%
08	THAILAND	51,000,000	+2.0%	18=	GERMANY	29,000,000	+3.6%
09	EGYPT	45,000,000	+2.3%	18=	NIGERIA	29,000,000	+3.6%
10	BANGLADESH	41,000,000	+5.1%	20	MYANMAR	27,000,000	+3.8%

(Source: wearesocial.com)

Social Media Statistics- Twitter



Wikipedia: Power of Crowdsourcing



4,000 experts
80,000 articles
200 years to develop
Annual Updates
“8.8/10.0 Reliability”



32,942,877 users
5.5 Million articles
Began in 2001
Real-Time Updates
“8.0/10.0 Reliability”

Currently, the [English Wikipedia](#) alone has over [5,576,548 articles](#) of any length. All Wikipedias taken together have **40 million articles in 293 languages**.

(Source: https://en.wikipedia.org/wiki/Wikipedia:Size_comparisons)

Most visited websites

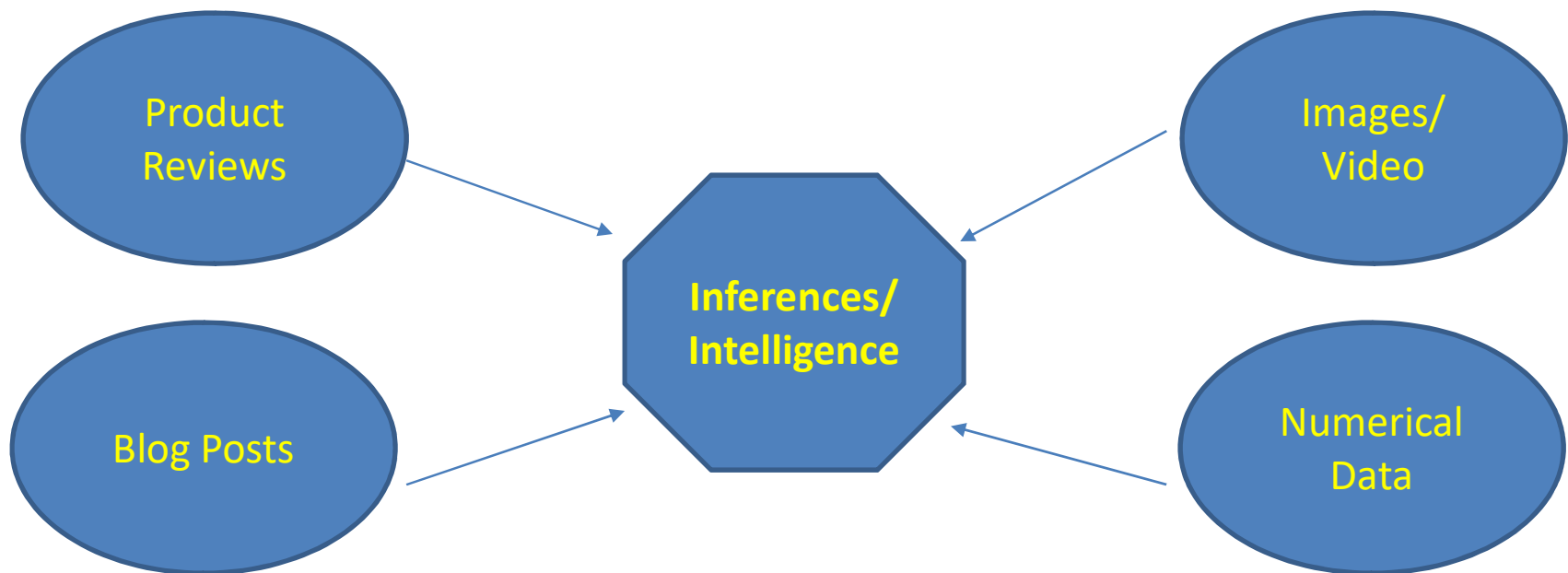
JAN 2021						WORLD'S MOST VISITED WEBSITES (SIMILARWEB)					RANKING OF THE WORLD'S MOST VISITED WEBSITES ACCORDING TO SIMILARWEB, BASED ON TOTAL WEBSITE TRAFFIC				
#	WEBSITE	TOTAL VISITS	UNIQUE VISITS	TIME PER VISIT	PAGES PER VISIT	#	WEBSITE	TOTAL VISITS	UNIQUE VISITS	TIME PER VISIT	PAGES PER VISIT				
01	GOOGLE.COM	92.21B	3,113M	10M 58S	8.3	11	PORNHUB.COM	3.24B	445M	8M 33S	7.2				
02	YOUTUBE.COM	35.75B	1,926M	21M 10S	11.1	12	AMAZON.COM	3.10B	552M	7M 24S	9.6				
03	FACEBOOK.COM	25.33B	2,003M	10M 36S	8.3	13	XNXX.COM	3.08B	382M	8M 27S	11.1				
04	TWITTER.COM	6.54B	902M	10M 49S	11.9	14	WHATSAPP.COM	3.02B	457M	2M 42S	1.5				
05	INSTAGRAM.COM	6.18B	1,009M	7M 45S	10.9	15	NETFLIX.COM	2.66B	261M	9M 54S	4.3				
06	WIKIPEDIA.ORG	5.83B	1,148M	3M 55S	3.0	16	LIVE.COM	2.51B	293M	7M 25S	8.2				
07	BAIDU.COM	5.70B	260M	6M 15S	8.1	17	YAHOO.CO.JP	2.44B	100M	9M 28S	6.7				
08	YAHOO.COM	3.95B	517M	7M 35S	5.8	18	ZOOM.US	2.26B	462M	4M 09S	3.2				
09	XVIDEOS.COM	3.75B	479M	10M 13S	8.9	19	VK.COM	1.81B	128M	16M 51S	19.8				
10	YANDEX.RU	3.27B	183M	11M 06S	9.0	20	REDDIT.COM	1.74B	236M	9M 11S	6.3				

Characterizing the new Web

- Data-driven Applications, data is next Intel-inside
- User at the Centre, Users add value
- Users as Co-creators and not only consumers
- Cooperate, Don't Control
- The perpetual Beta

Data Explosion

- Huge amount of data being generated.
- Large volume of unstructured data.



AI-ML-DL

Artificial Intelligence

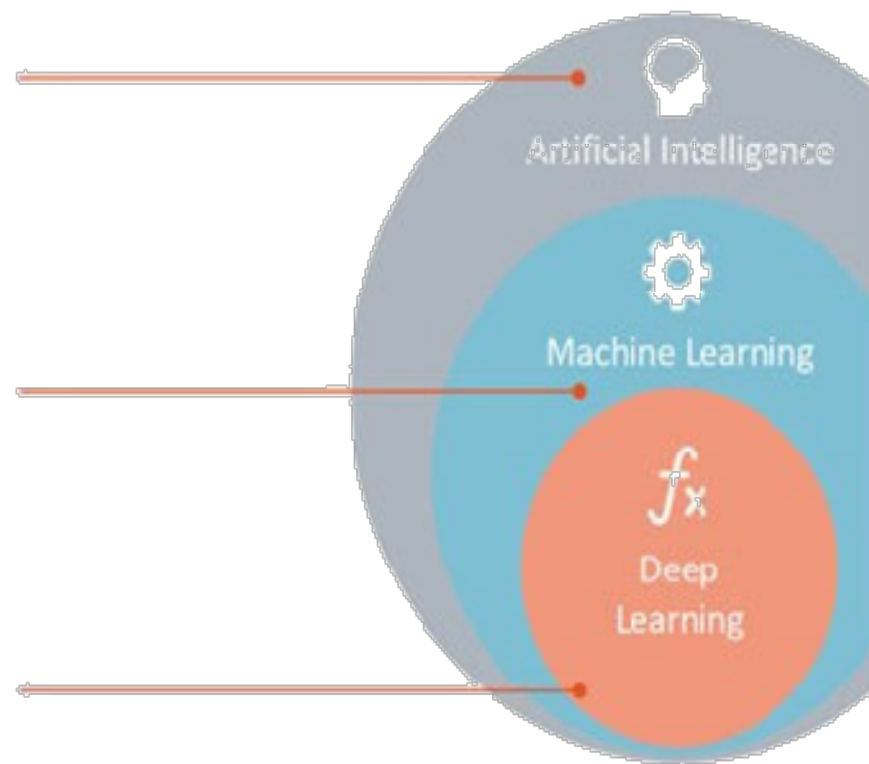
Any technique which enables computers to mimic human behavior.

Machine Learning

Subset of AI techniques which use statistical methods to enable machines to improve with experiences.

Deep Learning

Subset of ML which make the computation of multi-layer neural networks feasible.



Defining AI

- AI is the ability of a digital computer to perform tasks commonly associated with intelligent beings – [Encyclopedia Britannica](#).
- How to make computers do things at which at the moment, people are better. – [Rich & Knight](#).
- The exciting new effort to make computers think.....machines with minds, in the full and literal sense. – [Haugeland](#).
- The study of computations that make it possible to perceive, reason and act. – [Winston](#).

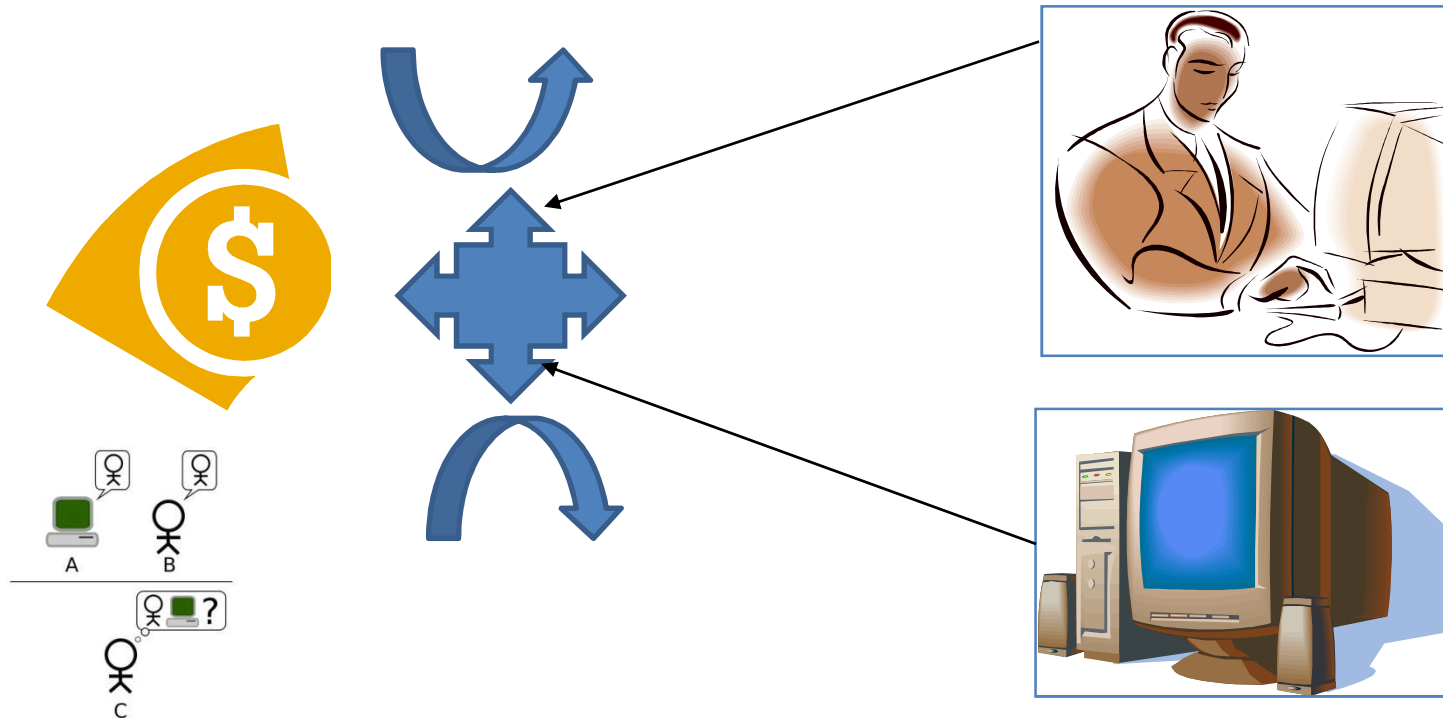
AI Subject Matter

- Knowledge based systems
- Reasoning and Problem Solving
- Machine Learning
- Natural language processing
- Game Playing
- Robotics
- Cognitive Architectures
- Intelligent Agents
- AI and Web 2.0

- Intelligent Information retrieval
- Artificial Neural Networks
- Perception
- Planning
- Vision & Signal Processing
- Swarm & Collective Systems
- Situated Cognition
- Social Computing
- Fuzzy Systems

The Turing's Test

- An Imitation game proposed in 1951 by Alan Turing in his paper 'Computing Machinery & Intelligence'.



Turing's test suggested for inclusion of capabilities of language processing, knowledge representation and automated reasoning for AI.

Searle's Chinese Room

- Turing Test is a behavioral test and not a true test of Intelligence.
- The Chinese room passes Turing's test, though it lacks intelligence.
- Searle's Chinese room argument suggested to first have a sufficiently precise theory of working of human mind and then to express it as computer program.
- Lead to *Thinking humanly* definition of AI.

AI: Journey

- Early Successes & Misplaced Optimisms
- A Dose of reality: AI Winter
- Expert Systems: Revival of Funding
- Engineering & Run
- Augmented Intelligence
- Situated Intelligence
- Collective Intelligence
- Big Data & Machine Learning
- Smart Systems

Machine Learning

Machine Learning

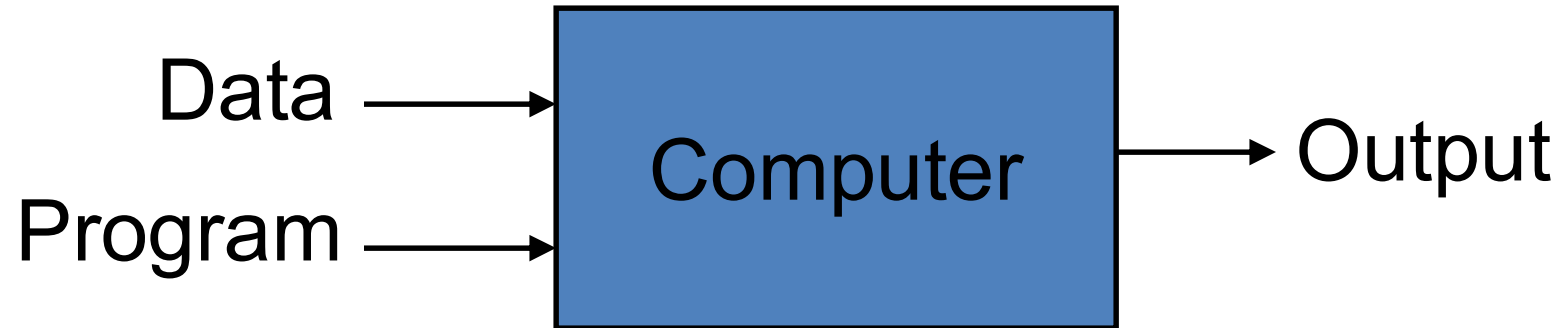
- Algorithms that can learn from data.
- Building models and using that for decisions/predictions rather than following explicitly programmed instructions.
- Supervised, Unsupervised and Semi-supervised.

Machine Learning: A Definition

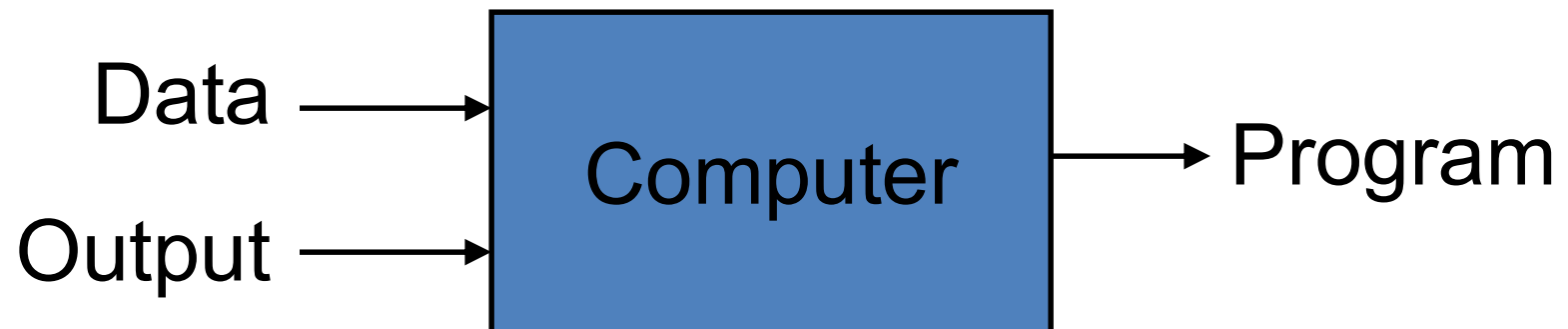
A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T , as measured by P , improves with experience E .

(As per T. Mitchell's Book)

Traditional Programming



Machine Learning



Magic?

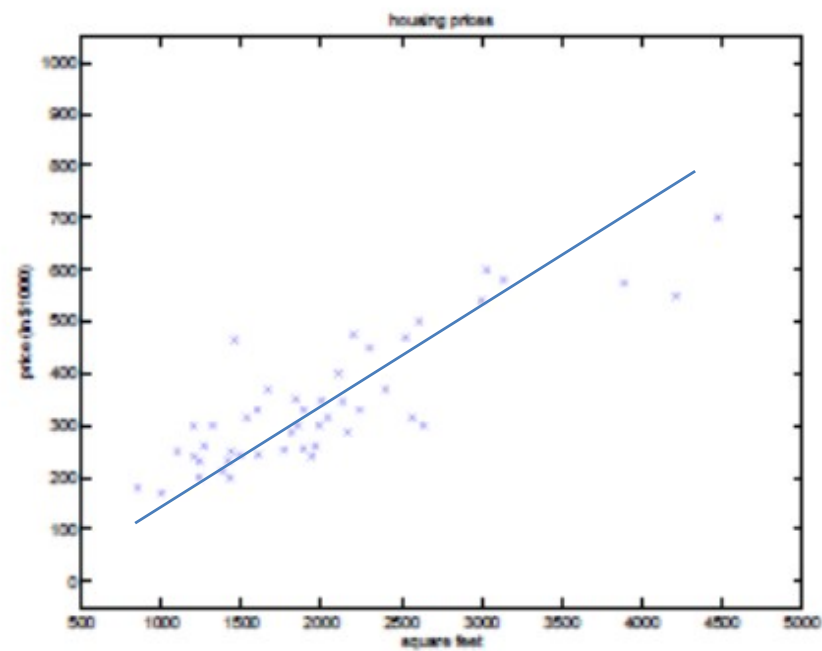
No, more like gardening

- **Seeds** = Algorithms
- **Nutrients** = Data
- **Gardener** = You
- **Plants** = Programs

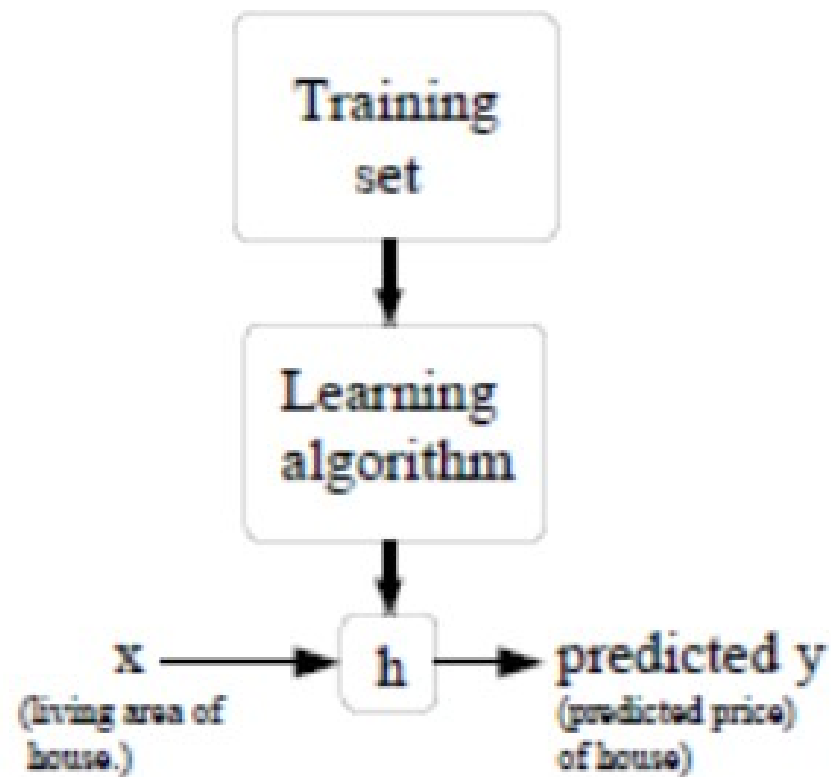


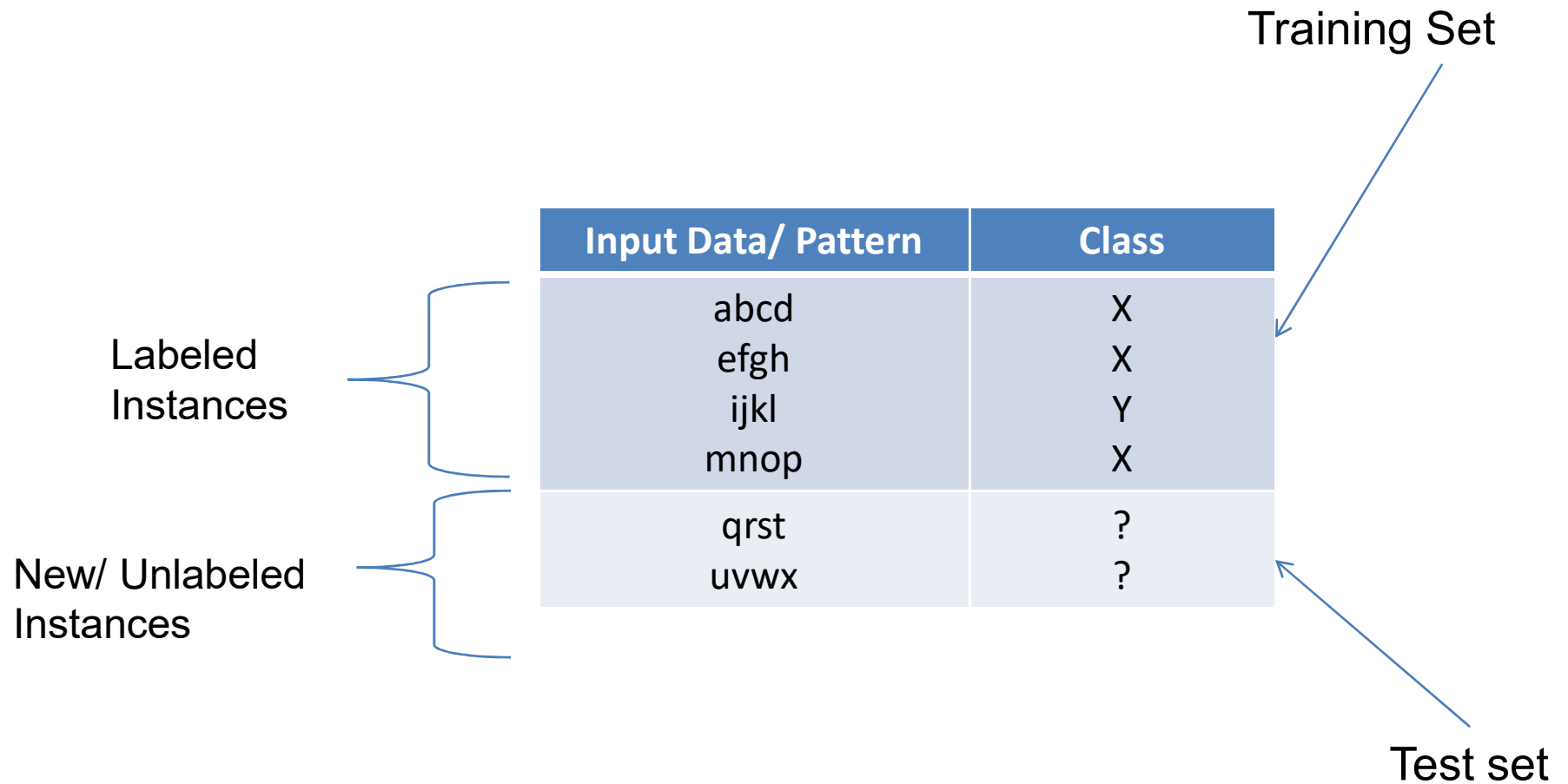
Living area (feet ²)	Price (1000\$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮

We can plot this data:



Machine Learning (Contd.)



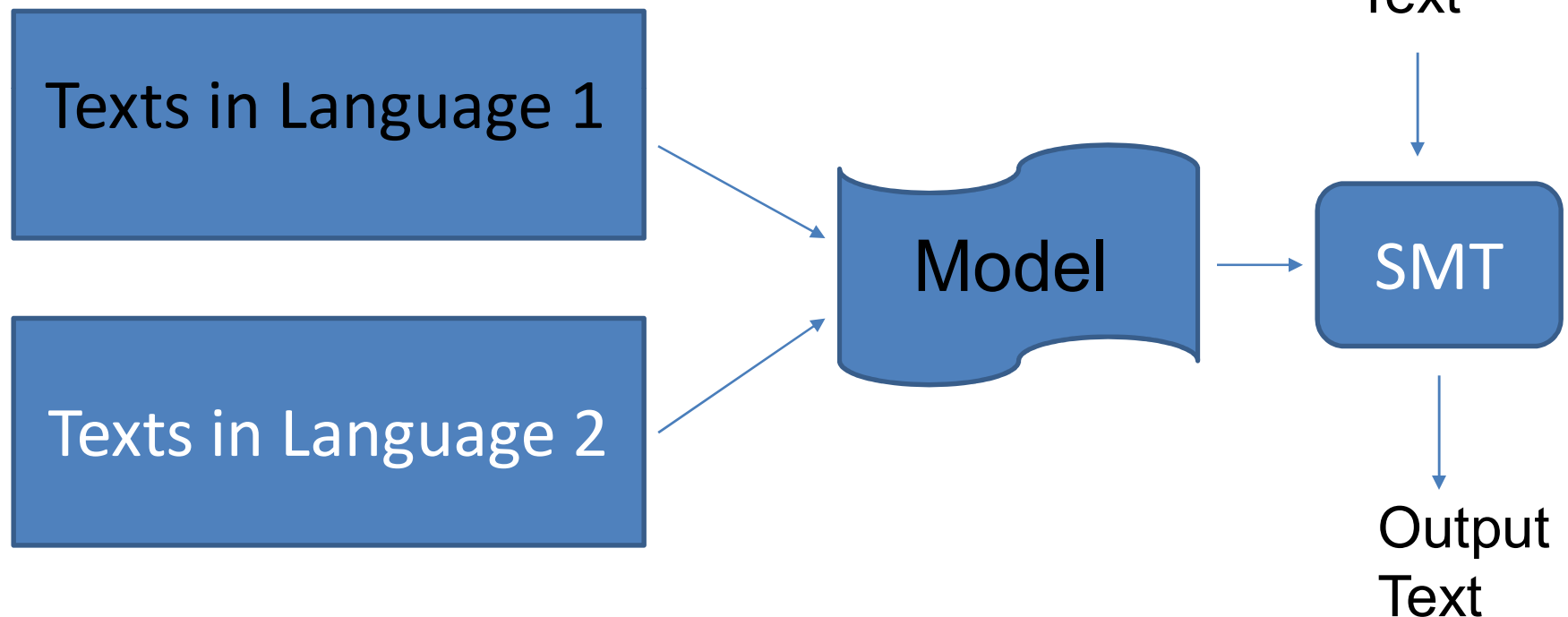


Why Machine Learning is popular?

- Large amount of data
- Cheap Computing Power
- Higher Storage (Memories)
- New kind of problems

Example: Machine Translation

- Statistical Machine Translation
- Parallel Corpora with suitable model



Machine Learning



what society thinks I
do



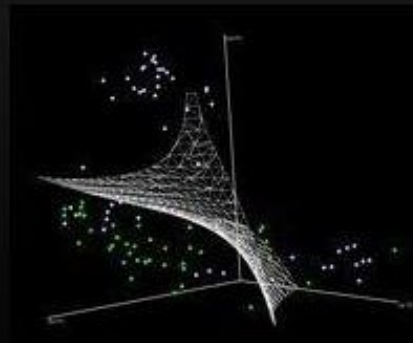
what my friends think
I do



what my parents think
I do

$$\begin{aligned}
 L_p &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^n \alpha_i \\
 \alpha_i &\geq 0, \forall i \\
 \mathbf{w} &= \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i, \sum_{i=1}^n \alpha_i y_i = 0 \\
 \nabla \hat{g}(\theta_t) &= \frac{1}{n} \sum_{i=1}^n \nabla \ell(x_i, y_i; \theta_t) + \nabla r(\theta_t) \\
 \theta_{t+1} &= \theta_t - \eta_t \nabla \ell(x_{i(t)}, y_{i(t)}; \theta_t) - \eta_t \cdot \nabla r(\theta_t) \\
 \mathbb{E}_{i(t)}[\ell(x_{i(t)}, y_{i(t)}; \theta_t)] &= \frac{1}{n} \sum_i \ell(x_i, y_i; \theta_t)
 \end{aligned}$$

what other programmers
think I do



what I think I do

```
>>> from scipy import svm
```

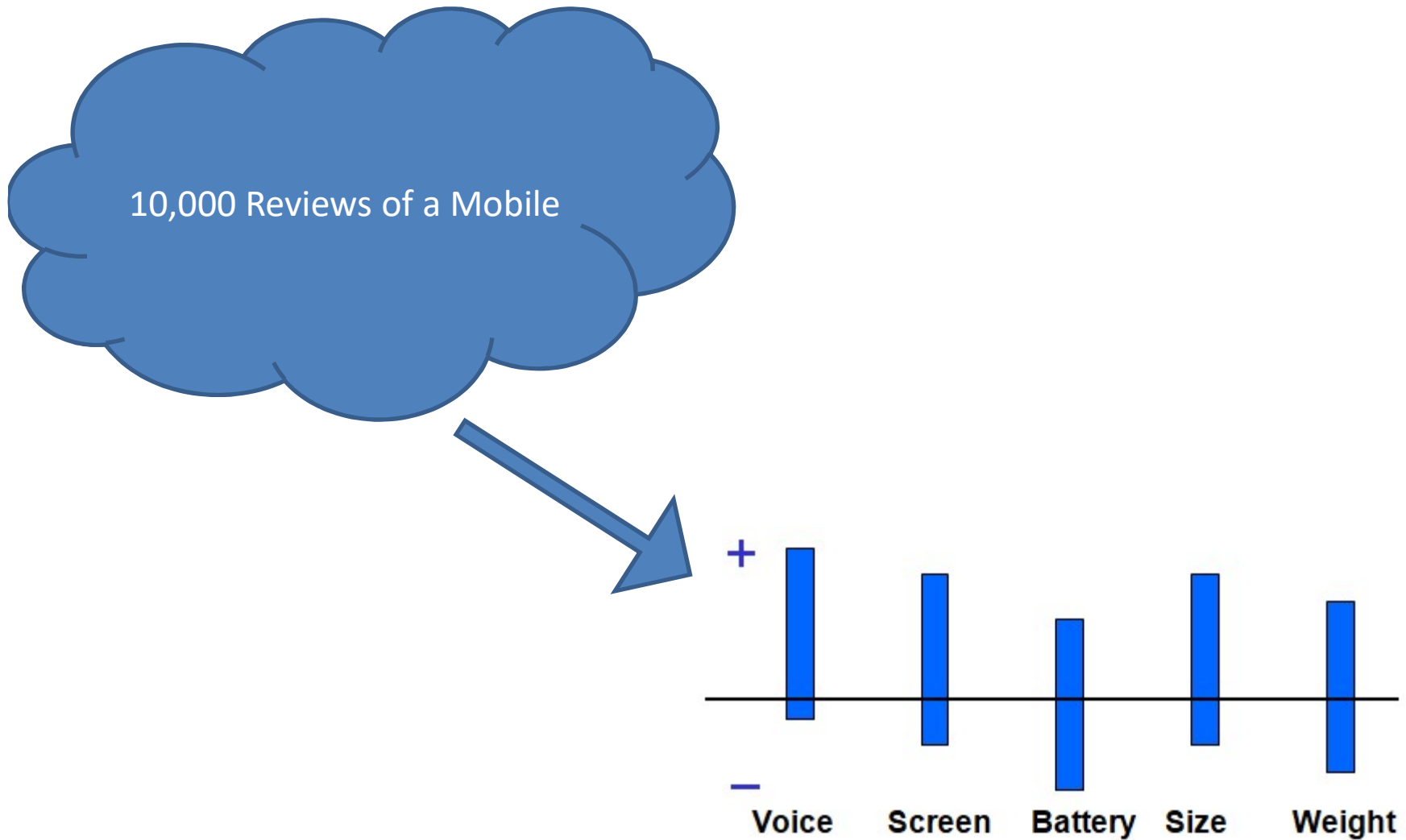
what I really do

Introductory Reference Books on ML

- David Barber, “Bayesian Reasoning and Machine Learning”, Cambridge University Press, 2013.
- T. Mitchell, “Machine Learning”, McGrawHill, 1997
- Kevin Murphy, “Machine Learning: A Probabilistic Perspective”, MIT Press, 2012.
- Christopher M Bishop, “Pattern Recognition and Machine Learning”, Springer, 2013

Making Sense from Textual Data

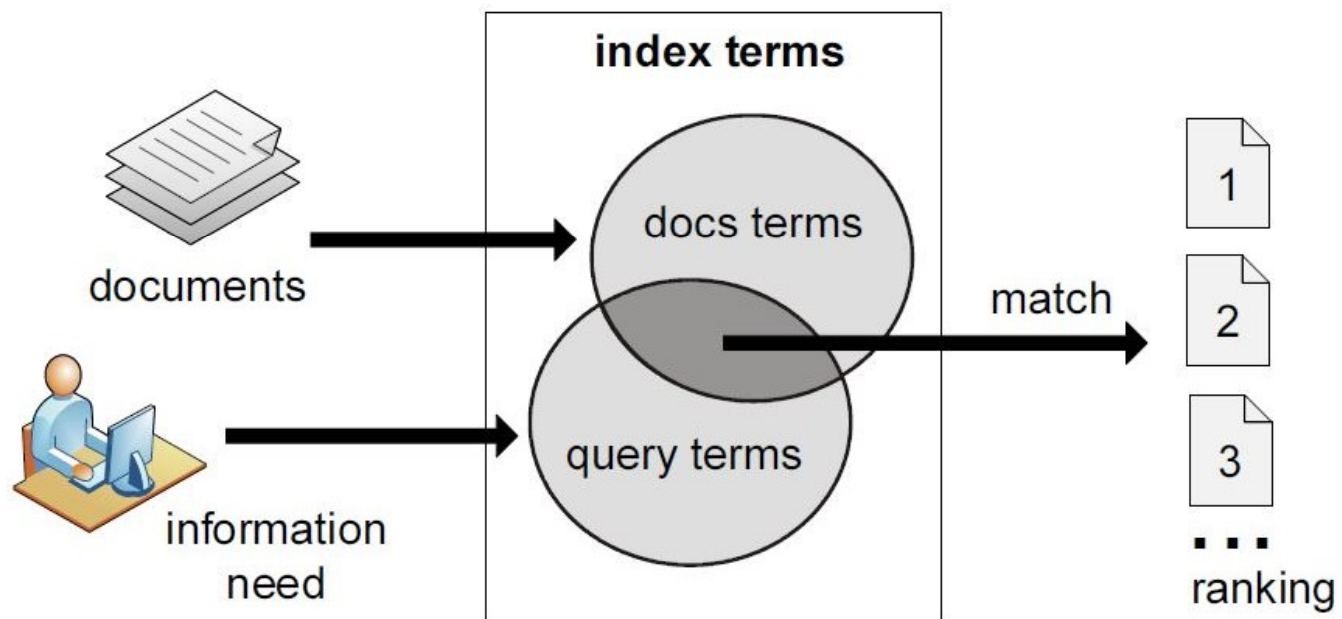
Making Sense from Textual Data

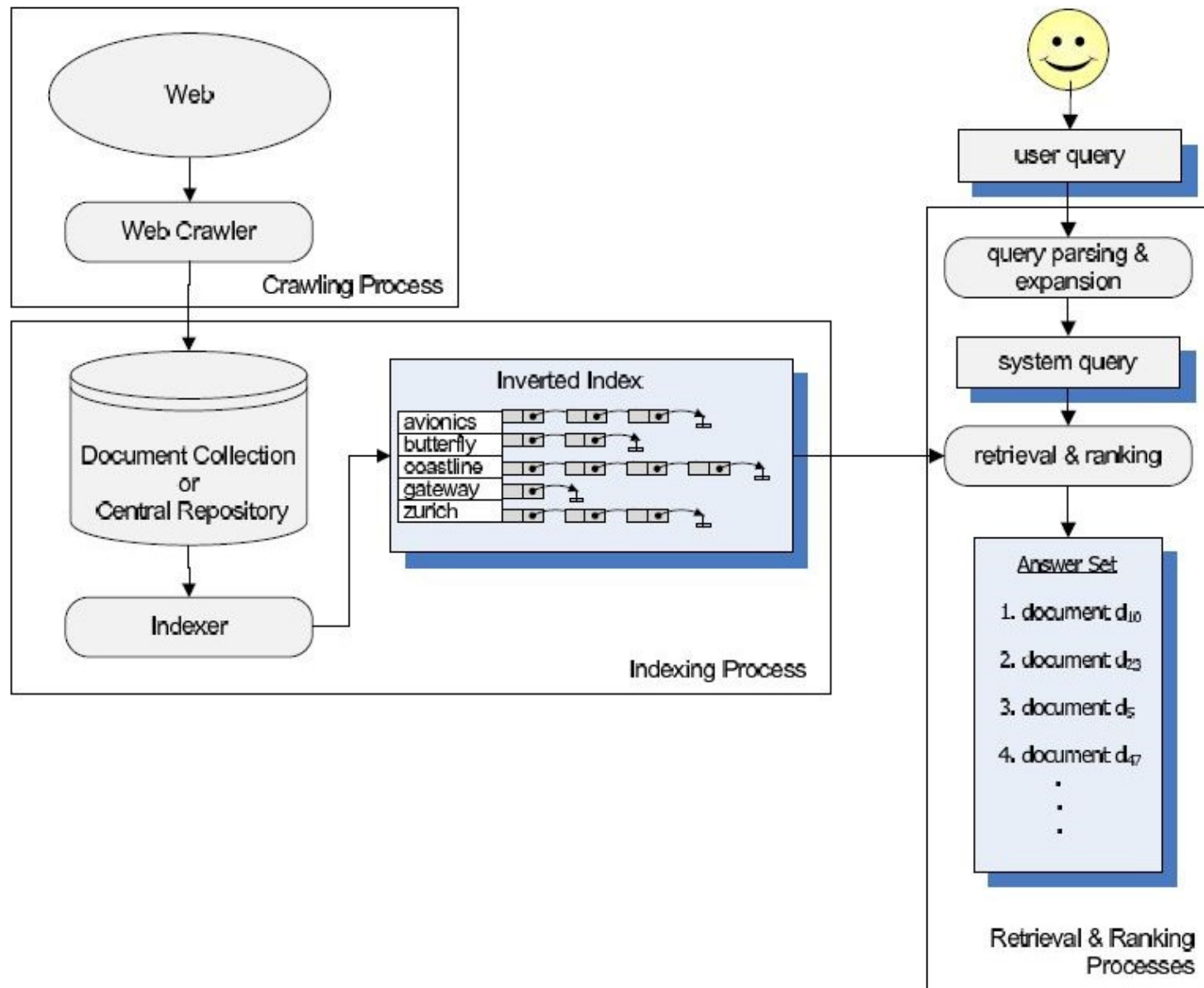


Information Retrieval



IR Process





IR Terminology

- **Document Collection:** Fixed set of documents, aka **corpus**.
- **Information need:** the topic about which the user desires to know more.
- **Relevance:** A document is relevant if it is the one that the user perceives as containing desired information about the topic.
- **Goal:** Retrieve documents with information that is relevant to the user's **information need**.

Text Processing- Representation

Vector Space Model

- Each document is represented as a vector of term values (w_{ij})
- Usually **tf** X **idf** values for each vocab term.
- Thus, **N** docs with vocab size of **M** results in each Doc(*i*) represented as $W_{i1}, W_{i2}, \dots, W_{iM}$.

Vector Space Model

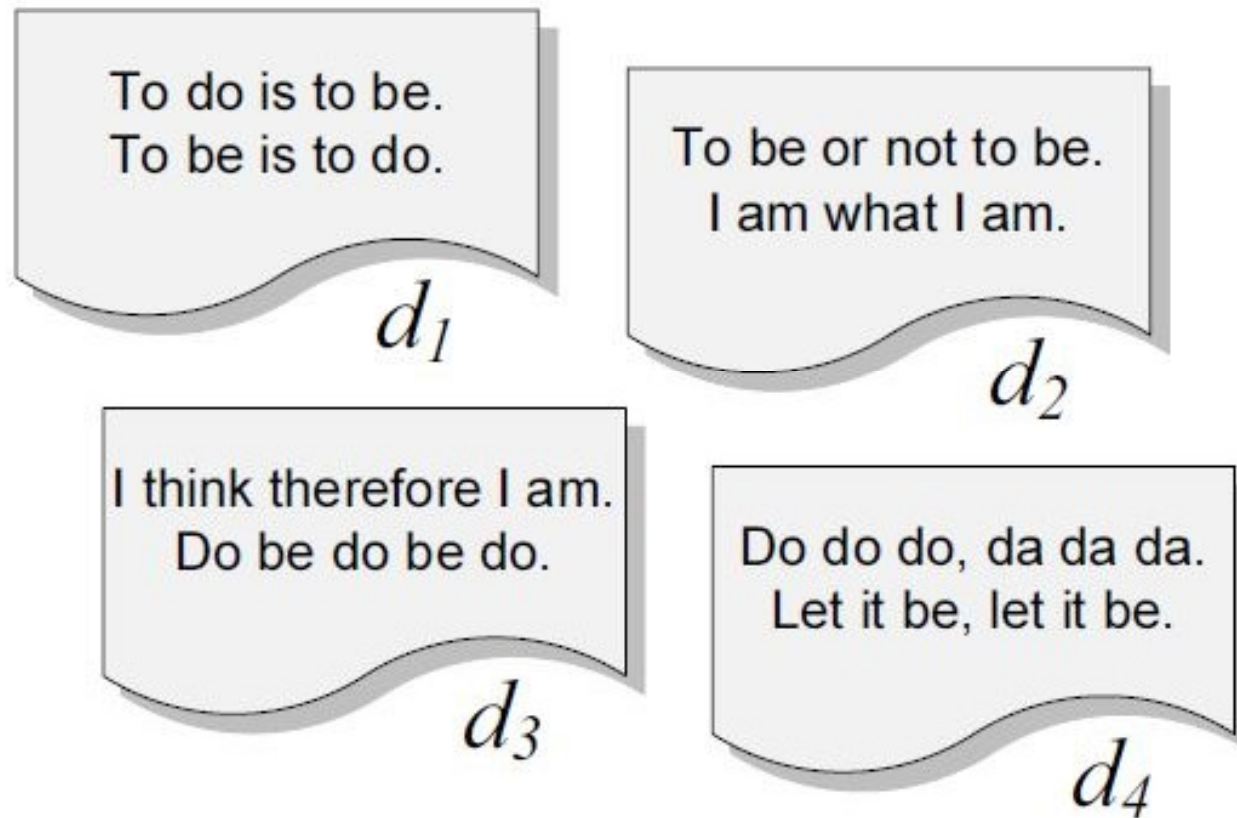
- Usually tf X idf values for each vocab term.

$$tf_{i,j} = \begin{cases} 1 + \log f_{i,j} & \text{if } f_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$idf_i = \log \frac{N}{n_i}$$

$$w_{i,j} = \begin{cases} (1 + \log f_{i,j}) \times \log \frac{N}{n_i} & \text{if } f_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Exercise: Representation



Exercise- Representation- tf values

Vocabulary		$tf_{i,1}$	$tf_{i,2}$	$tf_{i,3}$	$tf_{i,4}$
1	to	3	2	-	-
2	do	2	-	2.585	2.585
3	is	2	-	-	-
4	be	2	2	2	2
5	or	-	1	-	-
6	not	-	1	-	-
7	I	-	2	2	-
8	am	-	2	1	-
9	what	-	1	-	-
10	think	-	-	1	-
11	therefore	-	-	1	-
12	da	-	-	-	2.585
13	let	-	-	-	2
14	it	-	-	-	2

To do is to be.
To be is to do.

d_1

To be or not to be.
I am what I am.

d_2

I think therefore I am.
Do be do be do.

d_3

Do do do, da da da.
Let it be, let it be.

d_4

Exercise- Representation- idf values

To do is to be.
To be is to do.

d_1

To be or not to be.
I am what I am.

d_2

I think therefore I am.
Do be do be do.

d_3

Do do do, da da da.
Let it be, let it be.

d_4

	term	n_i	$idf_i = \log(N/n_i)$
1	to	2	1
2	do	3	0.415
3	is	1	2
4	be	4	0
5	or	1	2
6	not	1	2
7	I	2	1
8	am	2	1
9	what	1	2
10	think	1	2
11	therefore	1	2
12	da	1	2
13	let	1	2
14	it	1	2

Exercise- Representation- tf.idf values

To do is to be.
To be is to do.

d_1

To be or not to be.
I am what I am.

d_2

I think therefore I am.
Do be do be do.

d_3

Do do do, da da da.
Let it be, let it be.

d_4

		d_1	d_2	d_3	d_4
1	to	3	2	-	-
2	do	0.830	-	1.073	1.073
3	is	4	-	-	-
4	be	-	-	-	-
5	or	-	2	-	-
6	not	-	2	-	-
7	I	-	2	2	-
8	am	-	2	1	-
9	what	-	2	-	-
10	think	-	-	2	-
11	therefore	-	-	2	-
12	da	-	-	-	5.170
13	let	-	-	-	4
14	it	-	-	-	4

Notion of Similarity

Semantic and Syntactic

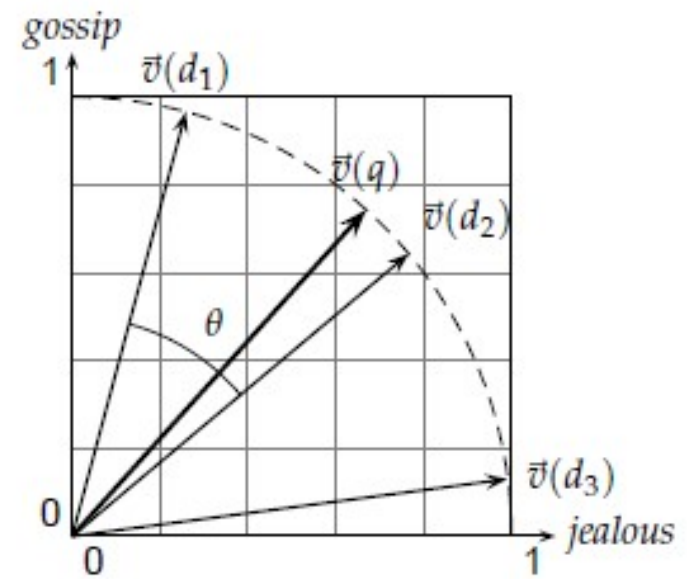
$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t.$$

In other words, $\text{tf-idf}_{t,d}$ assigns to term t a weight in document d that is

1. highest when t occurs many times within a small number of documents (thus lending high discriminating power to those documents);
2. lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal);
3. lowest when the term occurs in virtually all documents.

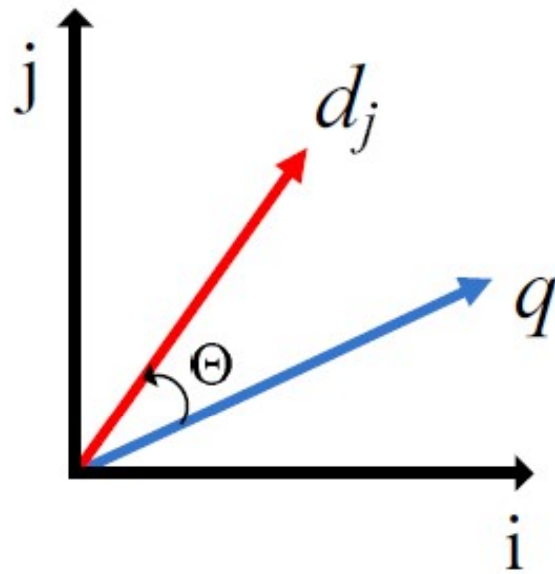
Cosine Similarity

$$\text{sim}(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)| |\vec{V}(d_2)|}$$



Cosine similarity illustrated. $\text{sim}(d_1, d_2) = \cos \theta$.

Cosine Similarity between Query and Document



$$\cos(\theta) = \frac{\vec{d}_j \bullet \vec{q}}{|\vec{d}_j| \times |\vec{q}|}$$

$$\text{sim}(d_j, q) = \frac{\sum_{i=1}^t w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^t w_{i,j}^2} \times \sqrt{\sum_{j=1}^t w_{i,q}^2}}$$

Exercise- Representation- Query

- Query term “to do”
- Vector Representation of Query-
 $\{1, 0.415, 0, 0, \dots\}$
- To compute Cosine Similarity of Query vector with each document vector.

Exercise- Representation- Ranking

To do is to be.
To be is to do.

d_1

To be or not to be.
I am what I am.

d_2

I think therefore I am.
Do be do be do.

d_3

Do do do, da da da.
Let it be, let it be.

d_4

doc	rank computation	rank
d_1	$\frac{1*3+0.415*0.830}{5.068}$	0.660
d_2	$\frac{1*2+0.415*0}{4.899}$	0.408
d_3	$\frac{1*0+0.415*1.073}{3.762}$	0.118
d_4	$\frac{1*0+0.415*1.073}{7.738}$	0.058

Exercise- Cosine Similarity

Documents	Content
d1	new york times
d2	new york post
d3	los angeles times

Query: “new new times”

Compute the score of each document in relative to this query, using the cosine similarity measure.

Exercise- Cosine Similarity

Term Frequency (tf)						
Docs/terms	angeles	los	new	post	times	york
d1	0	0	1	0	1	1
d2	0	0	1	1	0	1
d3	1	1	0	0	1	0

Inverse document Frequency (idf)					
angeles	los	new	post	times	York
$\log_2(3/1)$ = 1.584	$\log_2(3/1)$ = 1.584	$\log_2(3/2)$ = 0.584	$\log_2(3/1)$ = 1.584	$\log_2(3/2)$ = 0.584	$\log_2(3/2)$ = 0.584

Term Frequency (tf*idf)						
Docs/terms	angeles	los	new	post	times	york
d1	0	0	0.584	0	0.584	0.584
d2	0	0	0.584	1.584	0	1.584
d3	1.584	1.584	0	0	0.584	0

Exercise- Cosine Similarity

- Given the following query: “new new times”,

Inverse document Frequency (idf)						
q/terms	angeles	los	new	post	Times	York
query	0	0	$\left(\frac{2}{2}\right) * 0.584 = 0.584$	0	$\left(\frac{1}{2}\right) * 0.584 = 0.292$	0

calculate the length of each document and of the query

$$\text{Length of } d1 = \sqrt{(0.584^2 + 0.584^2 + 0.584^2)} = 1.011$$

$$\text{Length of } d2 = \sqrt{(0.584^2 + 1.584^2 + 0.584^2)} = 1.786$$

$$\text{Length of } d3 = \sqrt{(1.584^2 + 1.584^2 + 0.584^2)} = 2.316$$

$$\text{Length of } d4 = \sqrt{(0.584^2 + 0.292^2)} = 0.652$$

Exercise- Cosine Similarity

- Given the following query: “new new times”,

Then the similarity values are:

$$\cosSim(d1, q) = \frac{(0 * 0 + 0 * 0 + 0.584 * 0.584 + 0 * 0 + 0.584 * 0.292 + 0.584 * 0)}{(1.011 * 0.652)} = 0.776$$

$$\cosSim(d2, q) = \frac{(0 * 0 + 0 * 0 + 0.584 * 0.584 + 1.584 * 0 + 0 * 0.292 + 0.584 * 0)}{(1.786 * 0.652)} = 0.292$$

$$\cosSim(d3, q) = \frac{(1.584 * 0 + 1.584 * 0 + 0 * 0.584 + 0 * 0 + 0.584 * 0.292 + 0 * 0)}{(2.316 * 0.652)} = 0.112$$

- According to the similarity values, result to the query will be: d1, d2, d3

Machine Learning and Text Processing

Major Text Processing Tasks

- Classification
- Clustering
- Sentiment Analysis
- Summarization
- Disambiguation
- Topic Modeling
- Information Extraction/ Machine Reading etc.

Sentiment Analysis

Sentiment Analysis: Levels

- Document-Level
(One topic-one object assumption)
- Sentence-Level
(aggregate picture?)
- Aspect-Level
(Visualization and Decisions?)

Sentiment Analysis: doc-level

- Classify a document (e.g., a review) based on the overall sentiment expressed by opinion holder
 - Classes: Positive or negative (and neutral)
- It assumes
 - Each document focuses on a single object and contains opinions from a single opinion holder.
 - It considers opinion on the object.

Example1

My XYZ CAR was delivered yesterday. It looks faboulous. We went on a long highway drive the very second day of getting the car. It was smooth, comfortable and wonderful drive. Had a wonderful experience with family. Its an awesome car. I am loving it..!

Example 2

Bodyguard (2011) User Review from imdb.com

this is the worst i could ever imagined...stale story line up, stale expressions, awful and mindless south Indian action... Bodyguard featured Salmaan...after giving a crap like Ready what anyone could expect from this. This movie is better than ready but not upto mark of the star cast.... movie is inspired from a south Indian movie and also from one of Hindi movie's love story... movie contains action, comedy, drama but the only good part in the movie is the music....few songs are good n hummable....this movie is a total waste of money and time because it is awe-fully bad n STALE... so better keep out of this one.....

Example 3

Maruti Suzuki Swift – User Review from carwale.com:

i have been driving swift for past 2 years now and i have been hearing a lot of noise in the car. I think the parts have become loose or something and other then that the car is awesome it has tremendous power but one more disappointing fact is the mileage it hardly gives 10 in the city i think maruti can do a better job and even the suspensions are not great..

Sentiment Analysis : Sentence Level

- Sentence-level sentiment analysis has two tasks:
 - **Subjectivity classification**: Subjective or objective.
 - **Objective**: e.g., *I bought an iPhone a few days ago.*
 - **Subjective**: e.g., *It is such a nice phone.*
 - **Sentiment classification**: For subjective sentences or clauses, classify positive or negative.
 - **Positive**: *It is such a nice phone.*
- **However.**
 - subjective sentences **≠** +ve or –ve opinions
 - E.g., *I think he came yesterday.*
 - Objective sentence **≠** no opinion
 - **Implied –ve opinion**: *My phone broke in the second day.*

Example

- I bought an iPhone a few days ago.
- It was such a nice phone.
- The touch screen was really cool.
- The voice quality was clear too.
- Although the battery life was not long, that is ok for me.
- However, my mother was mad with me as I did not tell her before I bought the phone.
- She also thought the phone was too expensive, and wanted me to return it to the shop.....

Sentiment Analysis: Aspect Level

- Sentiment classification at both document and sentence (or clause) levels are not sufficient,
 - they do not tell what people like and/or dislike
 - A positive opinion on an object does not mean that the opinion holder likes everything.
 - An negative opinion on an object does not mean
- Objective: Discovering all quintuples.
- With all quintuples, all kinds of analyses become possible.

Example 1

iPhone- User Review:

I bought an iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life was not long, that is ok for me. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, and wanted me to return it to the shop. ...

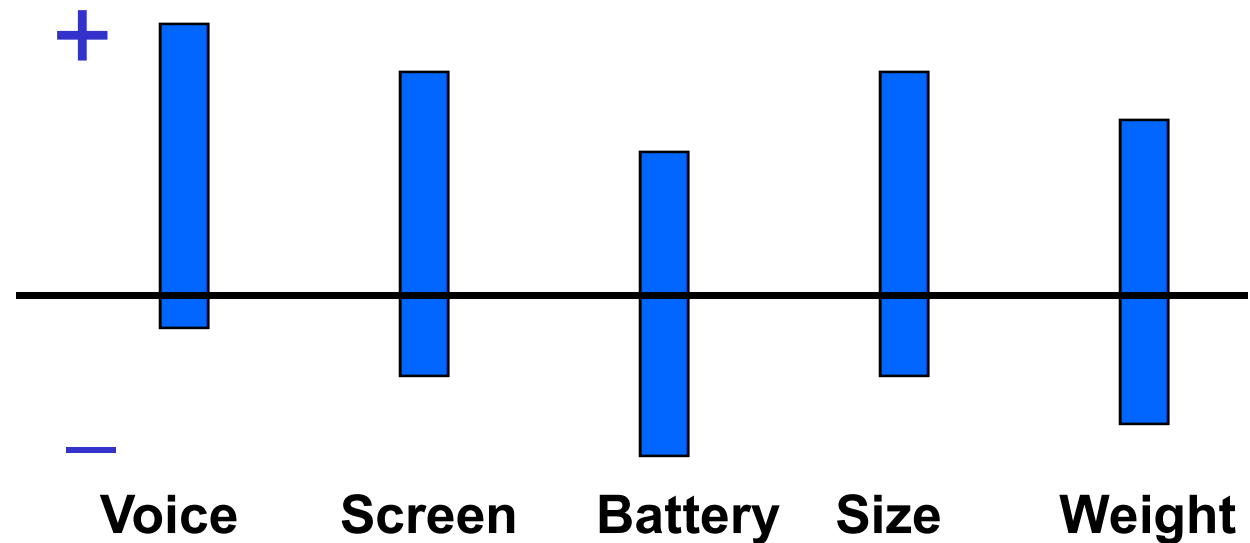
Example2

Maruti Suzuki Swift – User Review from carwale.com:

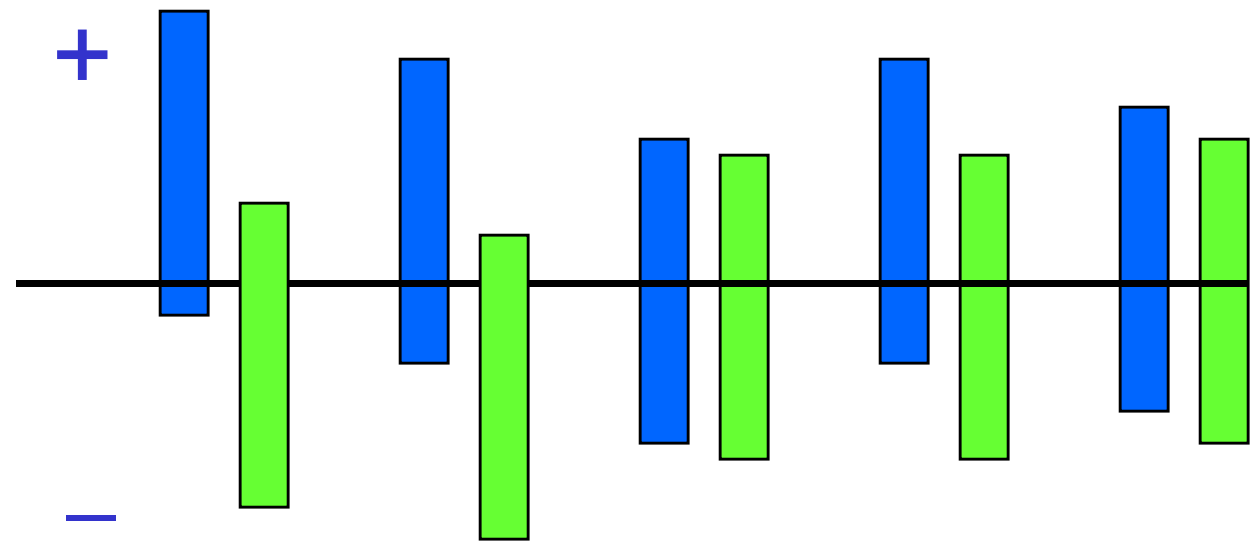
i have been driving swift for past 2 years now and i have been hearing a lot of noise in the car. I think the parts have become loose or something and other then that the car is awesome it has tremendous power but one more disappointing fact is the mileage it hardly gives 10 in the city i think maruti can do a better job and even the suspensions are not great..

Visual Comparison of Feature Based Analysis

- Summary of reviews of **Cell Phone 1**



- Comparison of reviews of **Cell Phone 1** and **Cell Phone 2**



Sentiment Analysis: Approaches

- Machine Learning Classifier Approach

(Naïve Bayes, Max Entropy, SVM etc.)

- Lexicon-based Approach

(SO-PMI-IR, SO-LSA, SentiWordNet etc.)

Some Online SA Tools

<https://www.uclassify.com/browse/uclassify/sentiment>

[uClassify](#) [Classifiers](#) [Docs](#) [Pricing](#) [About](#)



Sentiment

This classifier determines if a texts is positive or negative. It is well suited for million documents with data from Twitter, Amazon product reviews and movies. You may access the sentiment analysis api by signing up (free)! [Read more](#)

by  [uClassify](#)

Classify Text

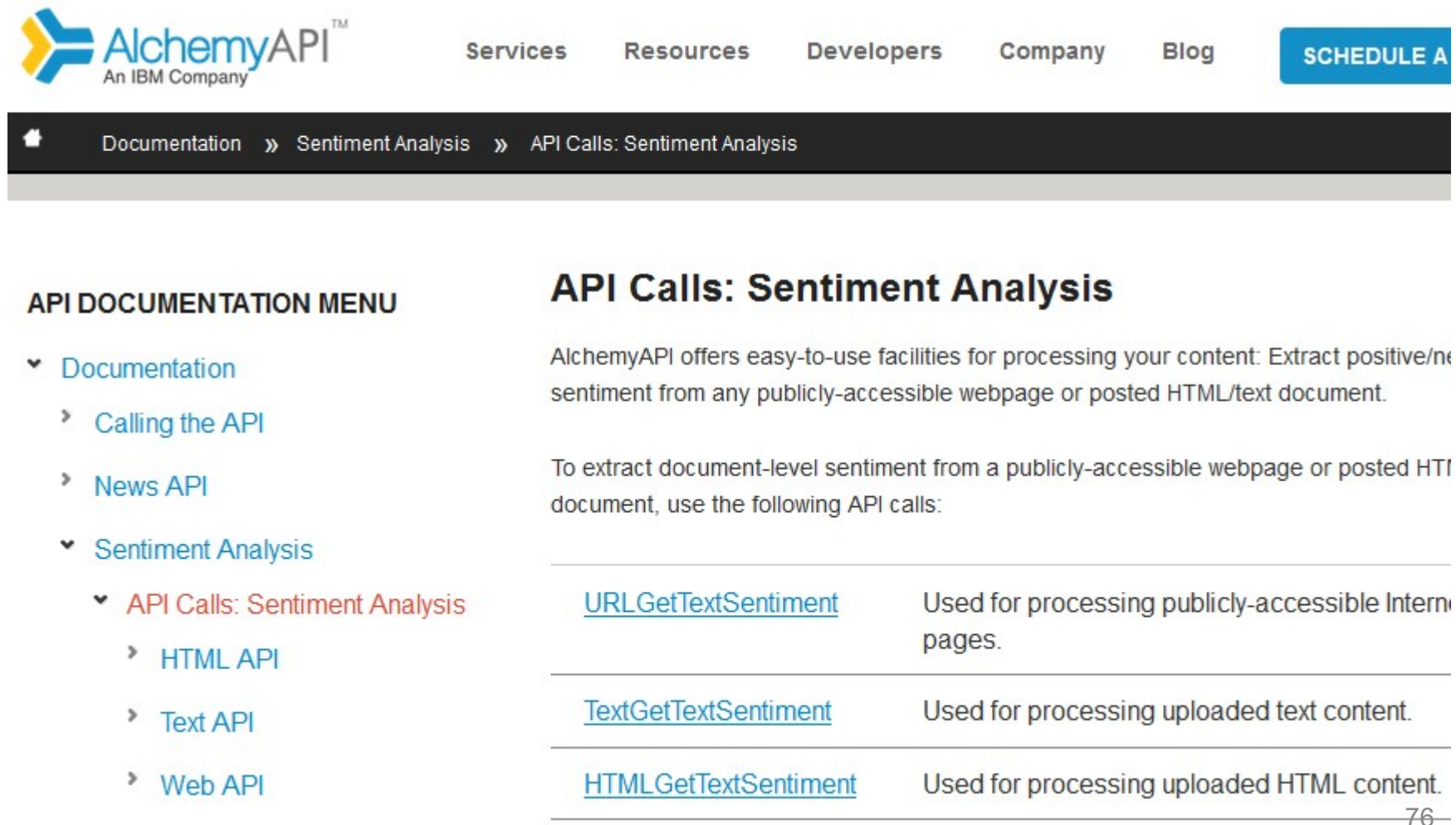
Classify Url

Classify Text

|Text to classify (no html)

Some Online SA Tools

- <http://www.alchemyapi.com/api/sentiment/proc.html>



The screenshot shows the AlchemyAPI website, an IBM Company. The navigation bar includes links for Services, Resources, Developers, Company, and Blog, along with a blue 'SCHEDULE A' button. The breadcrumb trail indicates the current location: Documentation » Sentiment Analysis » API Calls: Sentiment Analysis.

API DOCUMENTATION MENU

- ▼ Documentation
 - › Calling the API
 - › News API
- ▼ Sentiment Analysis
 - ▼ API Calls: Sentiment Analysis
 - › HTML API
 - › Text API
 - › Web API

API Calls: Sentiment Analysis

AlchemyAPI offers easy-to-use facilities for processing your content: Extract positive/negative sentiment from any publicly-accessible webpage or posted HTML/text document.

To extract document-level sentiment from a publicly-accessible webpage or posted HTML document, use the following API calls:

URLGetTextSentiment	Used for processing publicly-accessible Internet pages.
TextGetTextSentiment	Used for processing uploaded text content.
HTMLGetTextSentiment	Used for processing uploaded HTML content.



Suggested Readings for Sentiment Analysis

- B. Liu, Sentiment Analysis and Subjectivity. A Chapter in *Handbook of Natural Language Processing*, 2nd Edition, 2010. (An earlier version) B. Liu, Opinion Mining, A Chapter in the book: Web Data Mining, Springer, 2006. Download from: <http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>
- B. Pang, L. Lee & S. Vaidyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002. Download from: <http://www.cs.cornell.edu/home/llee/papers/sentiment.pdf>
- P.D. Turney, Thumbs up or Thumbs down? Semantic orientation applied to unsupervised classification of reviews, In Proc. Of ACL. Download from: <http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?action=rtdoc&an=8914166>
- A. Esuli & F. Sebastiani, SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining, Download from: <http://nmis.isti.cnr.it/sebastiani/Publications/LREC06.pdf>

- Rajesh Piryani, Madhavi Devaraj and Vivek Kumar Singh, “[Analytical Mapping of Opinion Mining and Sentiment Analysis Research during 2000-2015](#)”, *Information Processing and Management*, Vol. 53, No. 1, pp. 122-150, Jan. 2017, Elsevier. DOI: 10.1016/j.ipm.2016.07.001 (ISSN: 0306-4573, IF: 3.44)
- Singh, V. K. et al. 2013, “[Sentiment Analysis of Movie Reviews A new Feature-based Heuristic for Aspect-level Sentiment Classification](#)”, In Proceedings of International Multi Conference on Automation, Computing, Control, Communication and Compressed Sensing, Kerala, India, Mar. 2013. IEEE Xplore.
- Madhavi Devaraj, Rajesh Piryani and Vivek Kumar Singh, “[Lexicon Ensemble and Lexicon Pooling for Sentiment Polarity Detection](#)”, *IETE Technical Review*, Vol. 33, No. 03, pp. 332-340, Aug. 2015, Taylor and Francis. (ISSN: 0975-1084, IF: 1.33)

Concluding Remarks

- Huge volume of unstructured data being generated can be tapped for useful inferences/ decisions.
- Processing the large volume of data needs automated models.
- AI-ML-DL are being increasingly applied for data understanding and processing.

THANK YOU!