Demystifying
Data Science

Venkatesh Vinayakarao
venkateshv@cmi.ac.in
http://vvtesh.co.in

SSN School of Advanced Career Education

The world's most valuable resource is no longer oil, but data. – **Economist Report, 2017.**

What Comes Next?

byte

kilobyte

megabyte

gigabyte

??

???

????

?????

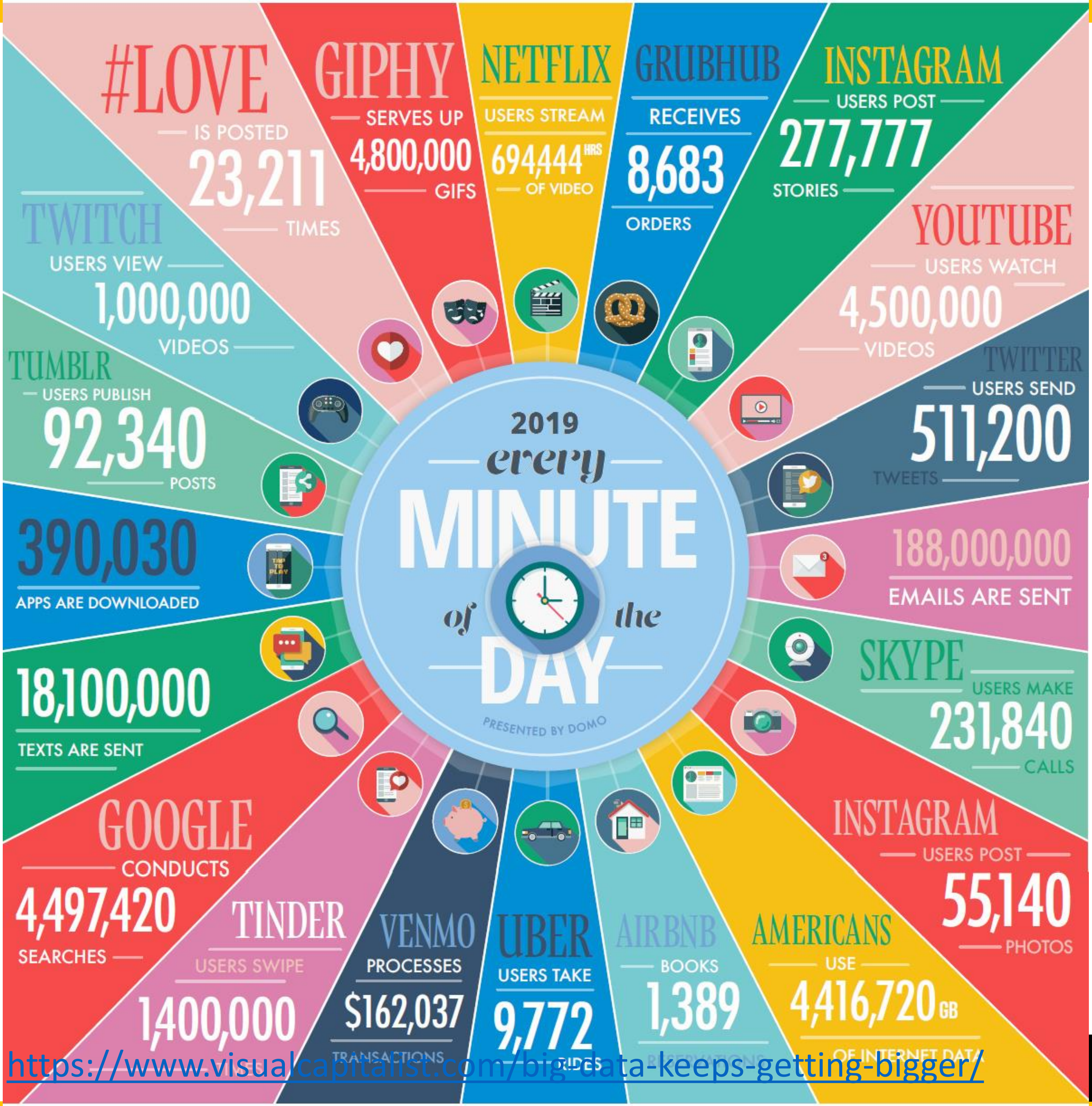
Sizes

Name	Size
Byte	8 bits
Kilobyte	1024 bytes
Megabyte	1024 kilobytes
Gigabyte	1024 megabytes
Terabyte	1024 gigabytes
Petabyte	1024 terabytes
Exabyte	1024 petabytes
Zettabyte	1024 exabytes
Yottabyte	1024 zettabytes

Big Data is Ubiquitous

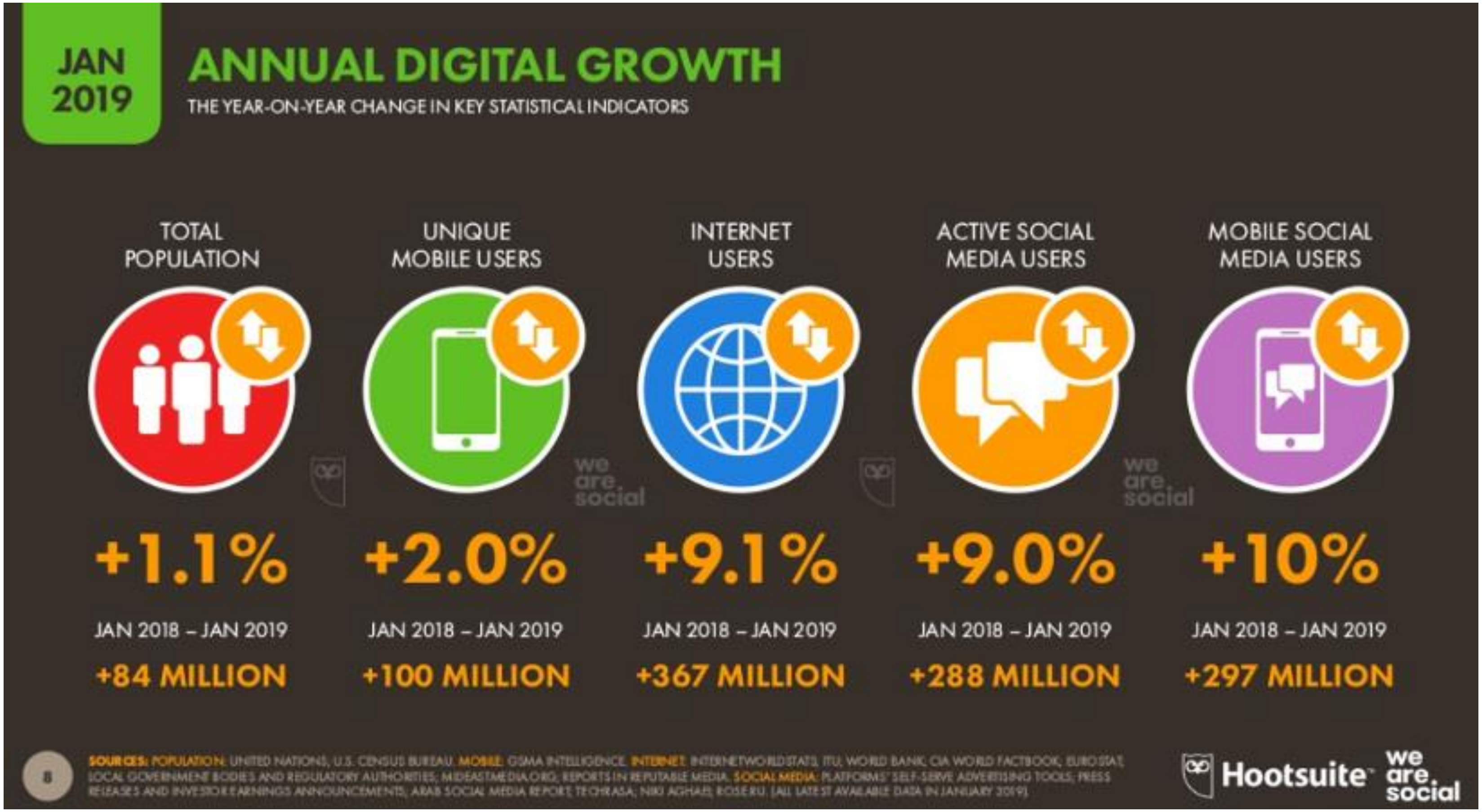
- Facebook Statistics
  - 1.5 billion people are active on Facebook **daily!**
  - **Every minute** there are 510,000 comments posted and 293,000 statuses updated!
  - More than 300 million photos get uploaded **per day**!
  - Totally, more than 2.5 Trillion posts!

Source: Forbes



Source: https://www.visualcapitalist.com/big-data-keeps-getting-bigger/

And, It is Growing!



JAN 2019

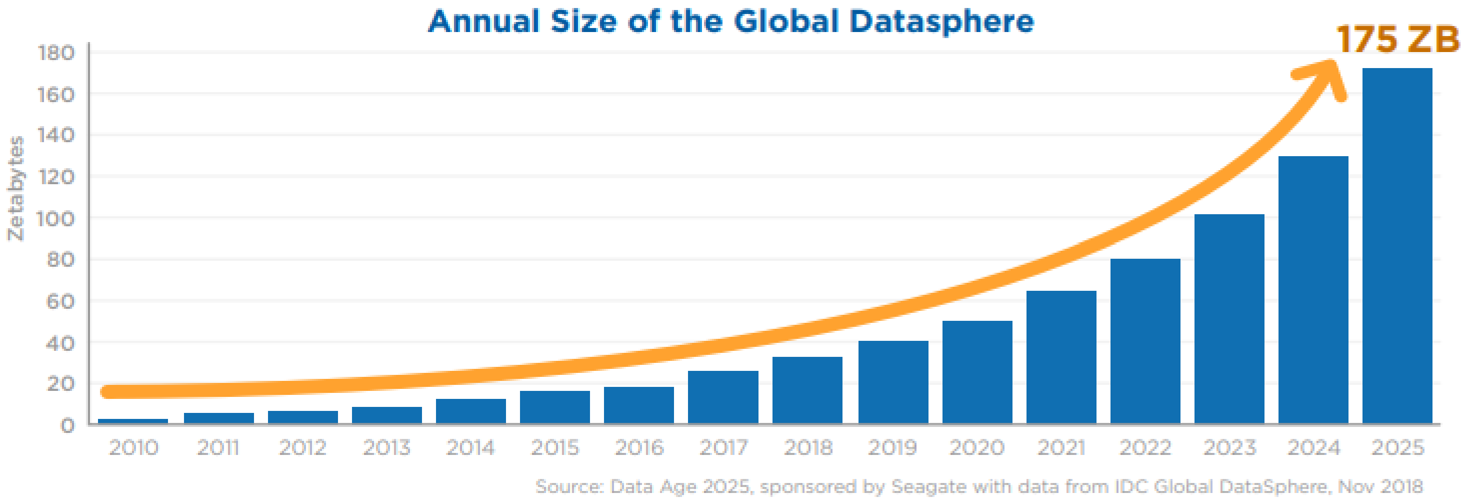
ANNUAL DIGITAL GROWTH

THE YEAR-ON-YEAR CHANGE IN KEY STATISTICAL INDICATORS

TOTAL POPULATION	UNIQUE MOBILE USERS	INTERNET USERS	ACTIVE SOCIAL MEDIA USERS	MOBILE SOCIAL MEDIA USERS
+1.1%	+2.0%	+9.1%	+9.0%	+10%
JAN 2018 – JAN 2019	JAN 2018 – JAN 2019	JAN 2018 – JAN 2019	JAN 2018 – JAN 2019	JAN 2018 – JAN 2019
+84 MILLION	+100 MILLION	+367 MILLION	+288 MILLION	+297 MILLION

SOURCES: POPULATION: UNITED NATIONS; U.S. CENSUS BUREAU. MOBILE: GSMA INTELLIGENCE. INTERNET: INTERNETWORLDSTATS; ITU; WORLD BANK; CIA WORLD FACTBOOK; EUROSTAT; LOCAL GOVERNMENT BODIES AND REGULATORY AUTHORITIES; MIDEASTMEDIA.ORG; REPORTS IN REPUTABLE MEDIA. SOCIAL MEDIA: PLATFORMS' SELF-SERVE ADVERTISING TOOLS; PRESS RELEASES AND INVESTOR EARNINGS ANNOUNCEMENTS; ARAB SOCIAL MEDIA REPORT; TECHRASA; NIKI AGHAEI; ROSE.RU. (ALL LATEST AVAILABLE DATA IN JANUARY 2019).

Data Growth



Mankind's quest to digitize the world!
33 ZB (2018) → 175 ZB (2025)
size of global datasphere*

*Source: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Solitary Confinement is Cruel

DATA & AI LANDSCAPE 2019
INFRASTRUCTURE
HADOOP ON-PREMISE
HADOOP IN THE CLOUD
STREAMING / IN-MEMORY
ANALYTICS
DATA ANALYST PLATFORMS
DATA SCIENCE PLATFORMS
APPLICATIONS – ENTERPRISE
SALES
MARKETING - B2B
MARKETING - B2C
CUSTOMER SERVICE
ENTERPRISE PRODUCTIVITY
NoSQL DATABASES
NewSQL DATABASES
GRAPH DBs
MPP DBs
CLOUD EDW
SERVERLESS
BI PLATFORMS
VISUALIZATION
MACHINE LEARNING
HUMAN CAPITAL
LEGAL
REGTECH & COMPLIANCE
FINANCE
BACK OFFICE AUTOMATION & RPA
SECURITY
DATA TRANSFORMATION
DATA INTEGRATION
DATA GOVERNANCE
MGMT / MONITORING
COMPUTER VISION
HORIZONTAL AI
SPEECH & NLP
APPLICATIONS – INDUSTRY
ADVERTISING
EDUCATION
GOVERNMENT
REAL ESTATE
FINANCE - INVESTING
FINANCE - LENDING
INSURANCE
STORAGE
CLUSTER SVCS
DATA GENERATION
HARDWARE
AGRICULTURE
COMMERCE
INDUSTRIAL
OTHER
FRAMEWORK
QUERY / DATA FLOW
DATA ACCESS & DATABASES
ORCHESTRATION & MANAGEMENT
STREAMING & MESSAGING
STAT TOOLS & LANGUAGES
AI / MACHINE LEARNING / DEEP LEARNING
SEARCH
LOGGING & MONITORING
VISUALIZATION
COLLABORATION
SECURITY
DATA SOURCES & APIs
HEALTH
IOT
FINANCIAL & ECONOMIC DATA
AIR / SPACE / SEA
PEOPLE / ENTITIES
LOCATION INTELLIGENCE
OTHER
DATA RESOURCES
DATA SERVICES
INCUBATORS & SCHOOLS
RESEARCH
Jun 27, 2019
© Matt Turck (@mattturck), Lisa Xu (@lisaxu92), & FirstMark (@firstmarkcap)
mattturck.com/bigdata2019
FIRSTMARK EARLY STAGE VENTURE CAPITAL

World needs data scientists!

Data Science

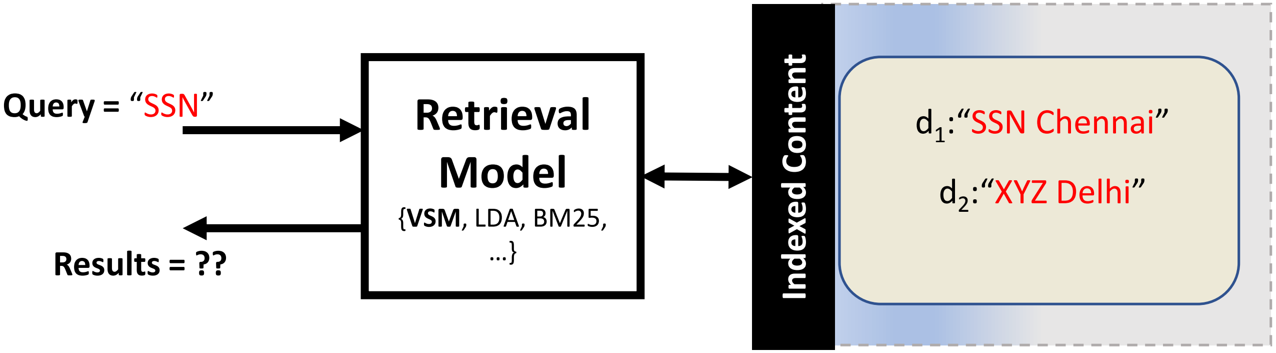
Loads of (structured and unstructured) data available.

Need scientifically sound methods to capture, maintain, process, communicate and analyze data.

Modern Text Processing

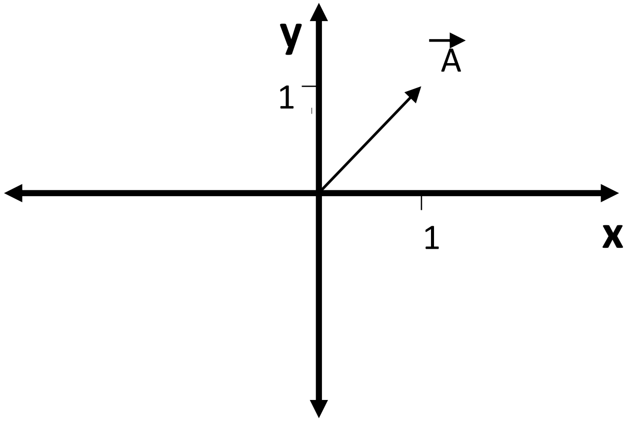
Vector Space Model

Which Document to Retrieve?



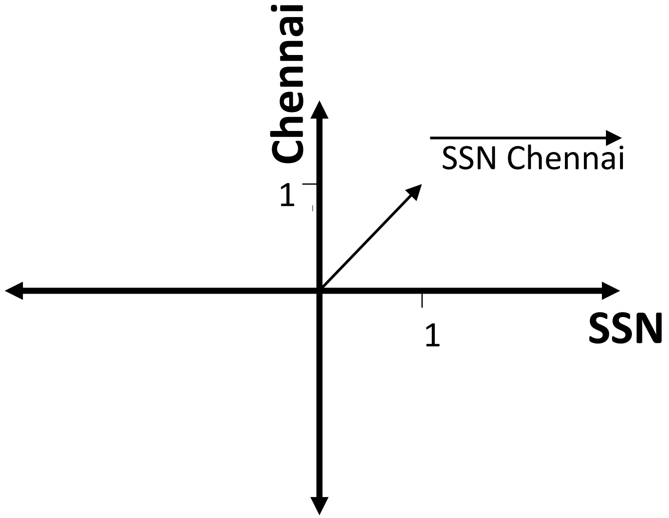
Vectors

- Geometric entity which has magnitude and direction



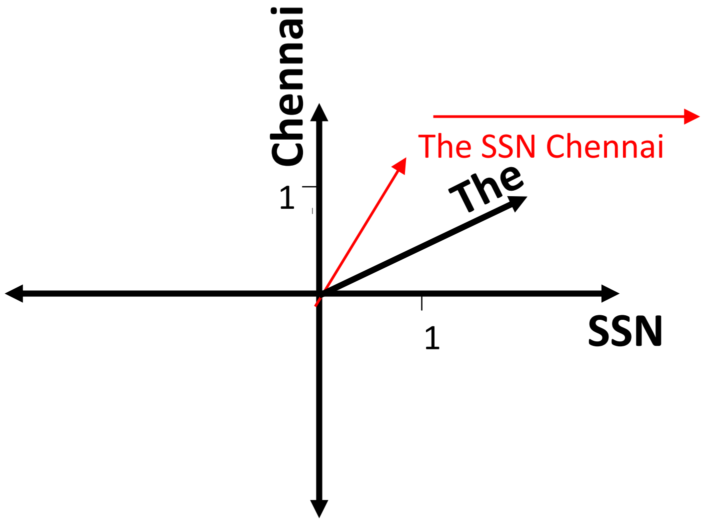
Sentences are vectors

- “SSN Chennai” as a vector



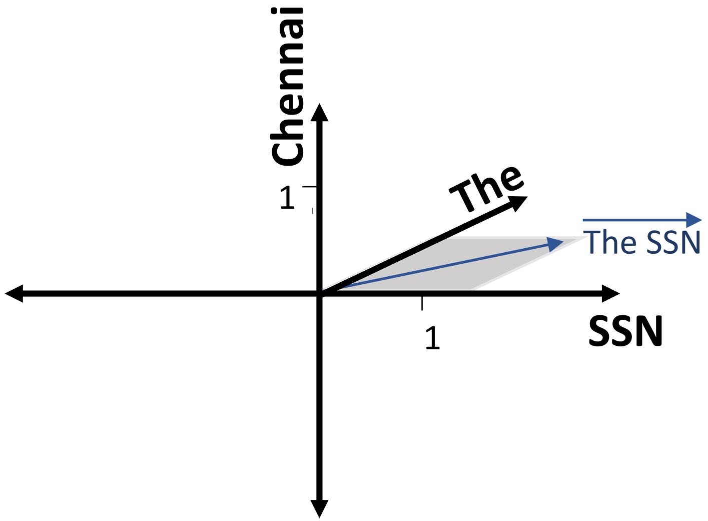
Sentences are vectors

- “The SSN Chennai” is a 3-dimensional vector



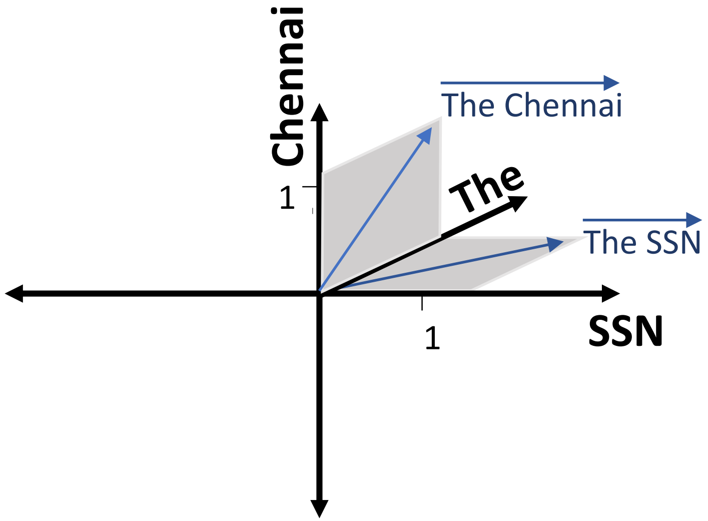
Sentences are vectors

- On this 3D space, “The SSN” vector will lie on the x (The) and z (SSN) plane.



Comparing Sentences

- We can compare sentences using the angle between vectors

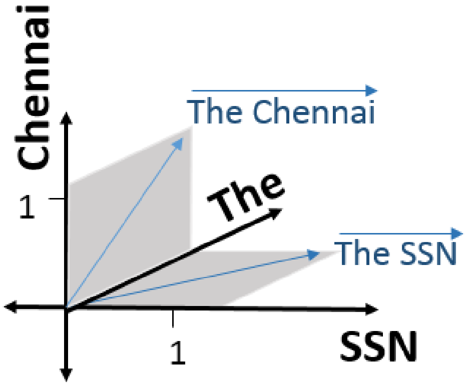


Angle between two vectors

- What is the angle between $\overrightarrow{\text{The}}$ and $\overrightarrow{\text{SSN}}$ vectors?
- What is the angle between $\overrightarrow{\text{SSN}}$ and $\overrightarrow{\text{Chennai}}$ vectors?
- What is the angle between $\overrightarrow{\text{The SSN}}$ and $\overrightarrow{\text{The SSN}}$ vectors?

Mathematical Notation

- We represent vectors as follows:
  - Vector = (dimension1, dimension2, dimension3, …)
    - First, define the dimensions
    - Next, put “1” if the word is present in the sentence, else “0”
- Example


Chennai
$\overrightarrow{\text{The Chennai}}$
The
1
$\overrightarrow{\text{The SSN}}$
1
SSN

In our example,

vector = (The, SSN, Chennai)

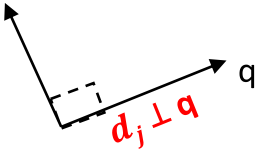
So,

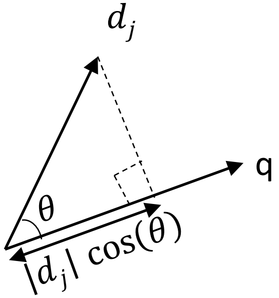
$\overrightarrow{\text{The Chennai}} = (1,0,1)$

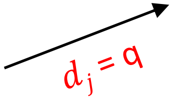
$\overrightarrow{\text{The SSN}} = (1,1,0)$

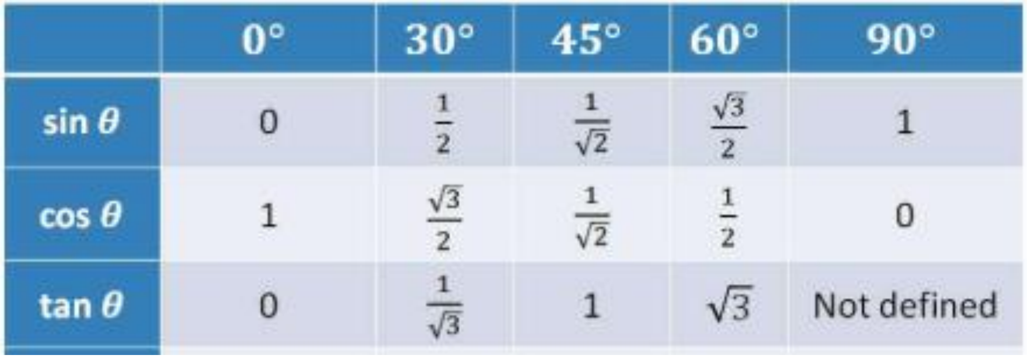
Converting from “0 – 90” to “0 – 1”

- For convenience, We convert the angles 0 – 90 to values 0 – 1
  - When vectors are same, we want to output 1.
  - When vectors are perpendicular, we want to output 0.









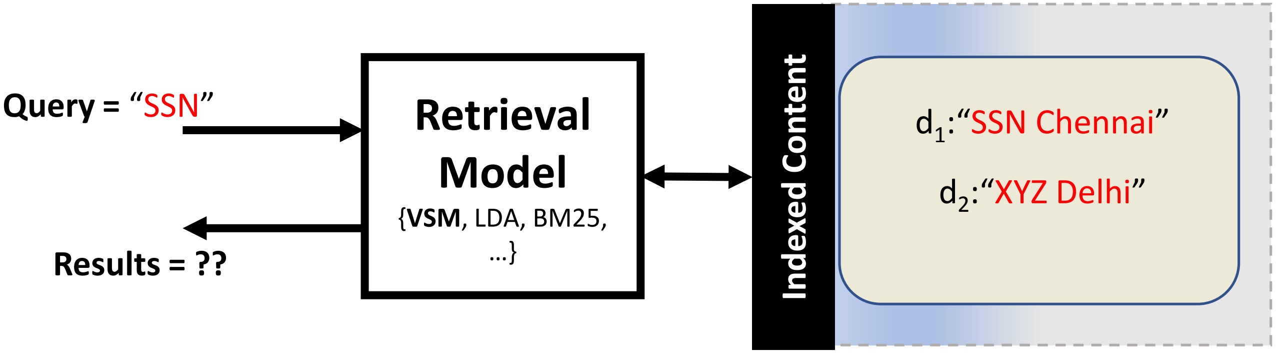
	0°	30°	45°	60°	90°
sin θ	0	$\frac{1}{2}$	$\frac{1}{\sqrt{2}}$	$\frac{\sqrt{3}}{2}$	1
cos θ	1	$\frac{\sqrt{3}}{2}$	$\frac{1}{\sqrt{2}}$	$\frac{1}{2}$	0
tan θ	0	$\frac{1}{\sqrt{3}}$	1	$\sqrt{3}$	Not defined

A Way to Calculate cosθ

- $\cos(\theta) = \frac{x.y}{||x||\ ||y||}$
- Here,
  - x.y is the “dot product” of x and y vectors.
- So, similarity between “The SSN” and “SSN Chennai”

$$= \frac{1.0 + 1.1 + 0.1}{\sqrt{1^2+1^2+0^2}\sqrt{0^2+1^2+1^2}} = \frac{1}{\sqrt{2}\sqrt{2}} = 0.5$$

Which Document to Retrieve?


Query = “SSN”

Results = ??

Retrieval Model

{**VSM**, LDA, BM25, …}

Indexed Content

d_1:“SSN Chennai”

d_2:“XYZ Delhi”

Example

Let query q = “SSN”.

Let document, d_1 = “SSN Chennai” and d_2 = “XYZ Delhi”.

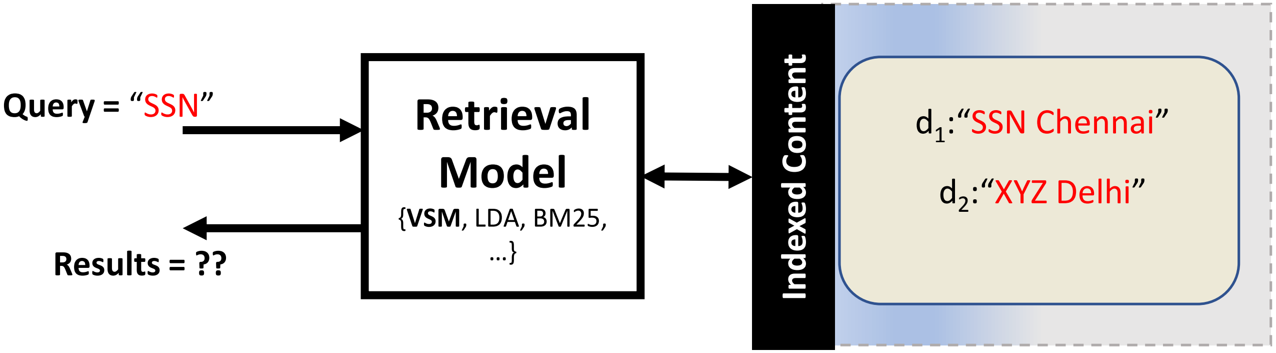
	SSN	Chennai	XYZ	Delhi
q	1	0	0	0
d_1	1	1	0	0
d_2	0	0	1	1

In our VSM, q = (1,0,0,0), d_1= (1,1,0,0) and d_2 = (0,0,1,1)

similarity(d_1, q) = $\frac{d_1.q}{||d_1||\ ||q||} = \frac{1.1+1.0+0.0+0.0}{\sqrt{1^2+1^2}\sqrt{1^2}} = \frac{1}{\sqrt{2}} = 0.71$

similarity(d_2, q) = $\frac{d_2.q}{||d_2||\ ||q||} = \frac{1.0+0.0+0.1+0.1}{\sqrt{1^2+1^2}\sqrt{1^2}}$ = 0.

Which Document to Retrieve?



Summary

- Data is Ubiquitous
  - and it is growing too!
- Modern Text Processing
  - Vector Space Model
- Remember
  - Data processing goes beyond common sense… we need techniques and tools.
  - Products are good to learn. Principles are even more important. Don’t ignore them.

Memories