



Monsoon 2022

Information Retrieval

- Introduction to IR Systems

Dr. Rajendra Prasath

Indian Institute of Information Technology Sri City, Chittoor



12th August 2022 (<http://rajendra.2power3.com/>)

> Focused Topics

- ▶ Recap
- ▶ Classical Search Engines
- ▶ Text Data
 - ▶ Structured vs Unstructured Text Data
- ▶ Keywords / User Information Needs
- ▶ Relevance / Irrelevance
- ▶ Personalization
- ▶ Words / Term Weighting
- ▶ Text Collection / Corpora
- ▶ Evaluation Strategy
 - ▶ More topics to come up ... Stay tuned ...!!

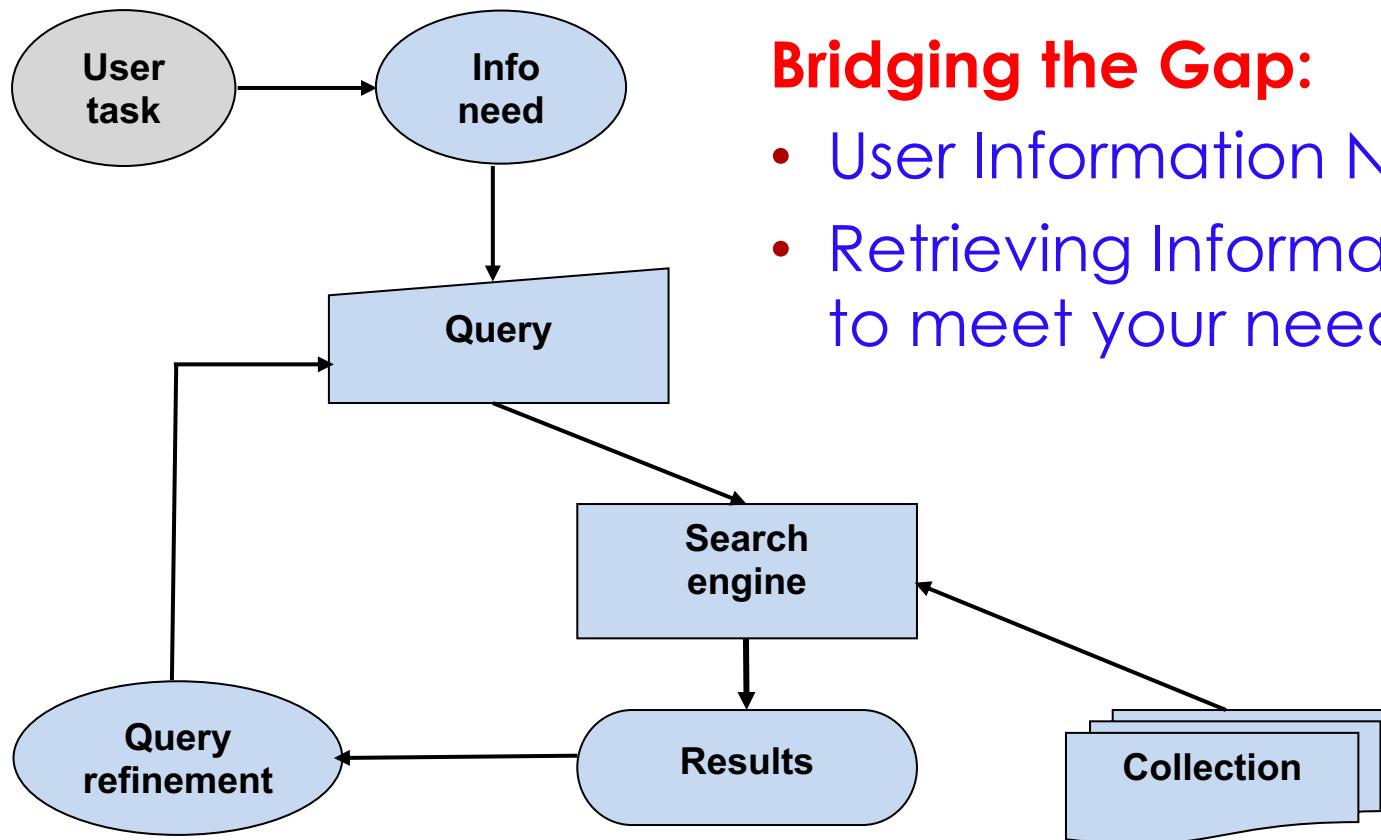


Recap: Information Retrieval

- **Information Retrieval (IR)** is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).
- These days we frequently think first of web search, but there are many other cases:
 - **E-mail Search**
 - **Search Your Files in Laptop / Desktop**
 - **Corporate knowledge bases**
 - **Legal information retrieval**
 - **News Retrieval Systems**
 - **World Wide Webs / Forums / Blogs**
 - **and so on . . .**



Recap: Classical Search Engines



Bridging the Gap:

- User Information Needs
- Retrieving Information to meet your needs

Data

How do we define Data?

→ **Look at the following data:**

80.32

80.32

91.54

76.23

76.23

80.32

54.24

54.24

76.23

64.96

64.96

64.96

91.54

91.54

54.24

→ **What do these numbers refer to?**

→ **Marks or Heights or Weights or . . .**



Structured Data

→ Look at the following data:

Export Item	13 – 17 Jan 2018*
Turmeric	80.32
Turmeric	76.23
Turmeric	54.24
Turmeric	64.96
Turmeric	91.54

* In metric tons

Usage	13 – 17 Jan 2018*
Turmeric	80.32
Turmeric	76.23
Turmeric	54.24
Turmeric	64.96
Turmeric	91.54

* Average usage of a nuclear family in Hyd (in milligrams)

→ Does it make any sense now?



Structured Data (contd)

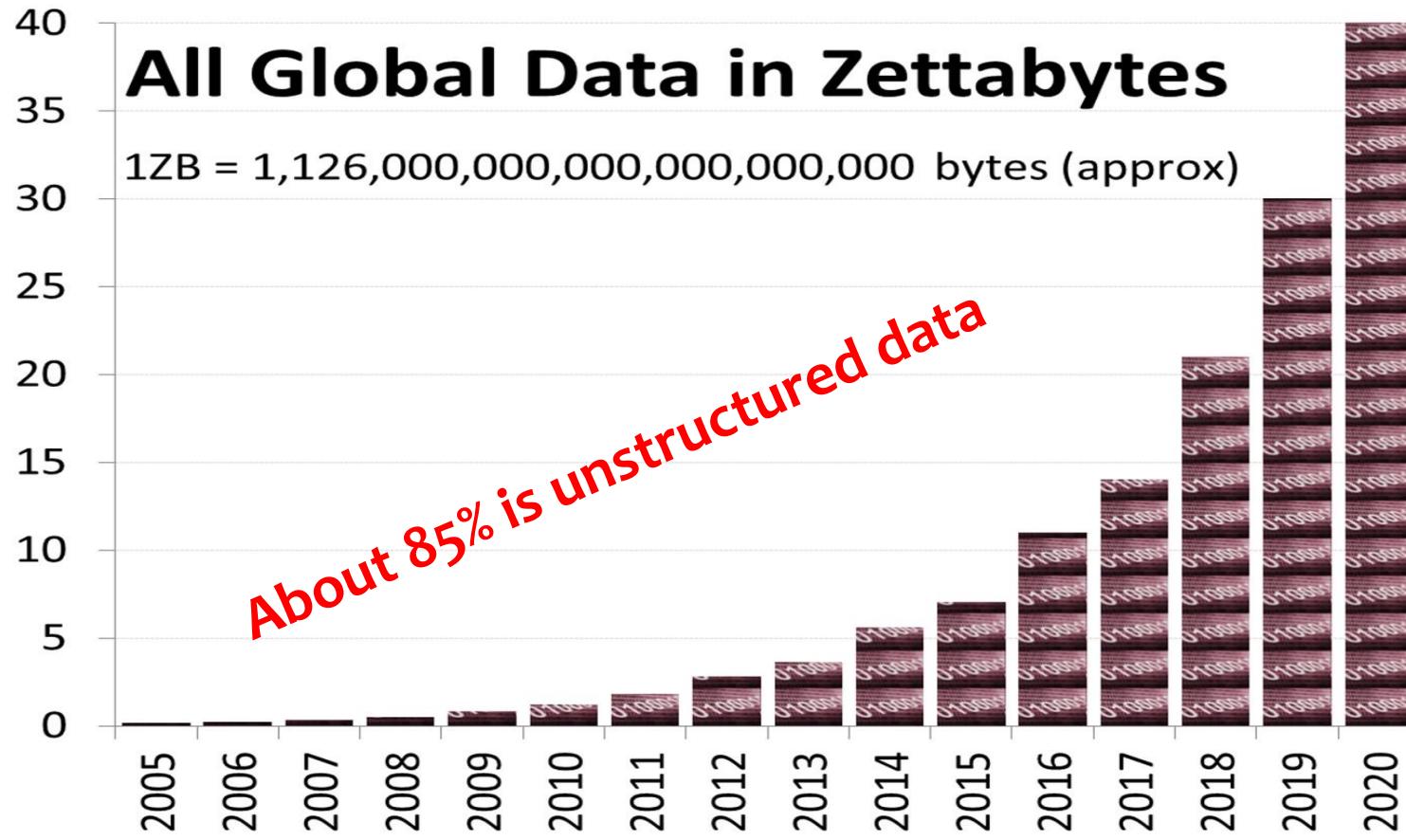
→ Let us look into the details

		Name	Weight	Name	2018 (25 Dec)
Rahul	80.32	Rohan	91.54	Rohan	91.54
Priya	76.23	Priya	80.32	Priya	80.32
Sowmya	54.24	Rahul	76.23	Rahul	76.23
Lucky Vogel	64.96	Lucky Vogel	64.96	Lucky Vogel	64.96
Rohan	91.54	Sowmya	54.24	Sowmya	54.24

→ More Data ... More Features ... More Value(s)
→ More Value → More Money (?!)

Size – Scaling of Text data

Huge Data → mind – blowing size ($1\text{ZB} = 10^{21}\text{bytes}$)



Text Data

No Free Meals: Transaction Fees are not the same.

13 Jan 2018

Mumbai: Manhattan FinServices Inc. has analyzed transaction fees of various mobile payment products. According to per RBI analytical report, PayGO is the top ranked payment products. PayGO ranks top in the list of transaction fees. Rupay is charging the lowest transaction fee of 54.24 paise per transaction. Online seller Amazon charges 64.96 paise per transaction. The transaction fee of PayPal and Paytm is 76.23 and 80.32 respectively. We also observed that PayPaise is charging the highest transaction fee as compared with PayGO.

Many interesting facts revealed in this report which can help you to choose the payment service which is interested to opt for the specific payment service. With mess transaction fee.

Product	13 Jan 2018 *
Paytm	80.32
PayPal	76.23
Rupay	54.24
Amazon Pay	64.96
PayGO	91.54

* Fee in Paise(INR) per transaction in Mumbai

→ Does this table give a better understanding?

What do you understand?

Non-Banking Financial Companies (NBFC) are establishments that provide financial services and banking facilities without meeting the legal definition of a Bank. The top 10 companies are listed by The Banking Finance Post on November 17, 2018 says that Power Finance Corporation Limited ranks topper with INR 267377.40 in terms of annual turn over which is measure in millions of rupees. The second and third places are held by Rural Electrification Corporation Limited (INR 224403.10 millions) and Bajaj Finance Limited (INR 133292.20 millions). The remaining companies in the top 12 list are given as follows with their annual turnover in the brackets respectively: Shriram Transport Finance Company Limited (122768.30), Indian Railway Finance Corporation Limited (110202.32) Mahindra & Mahindra Financial Services Limited (72061.20) HDB Financial Services Limited (70619.90) Muthoot Finance Limited (62432.00) Cholamandalam Investment and Finance Company Limited ((54257.60) L&T Finance Limited (52460.00) Shriram City Union Finance Limited (51015.70), and Tata Capital Financial Services Limited (45553.70).



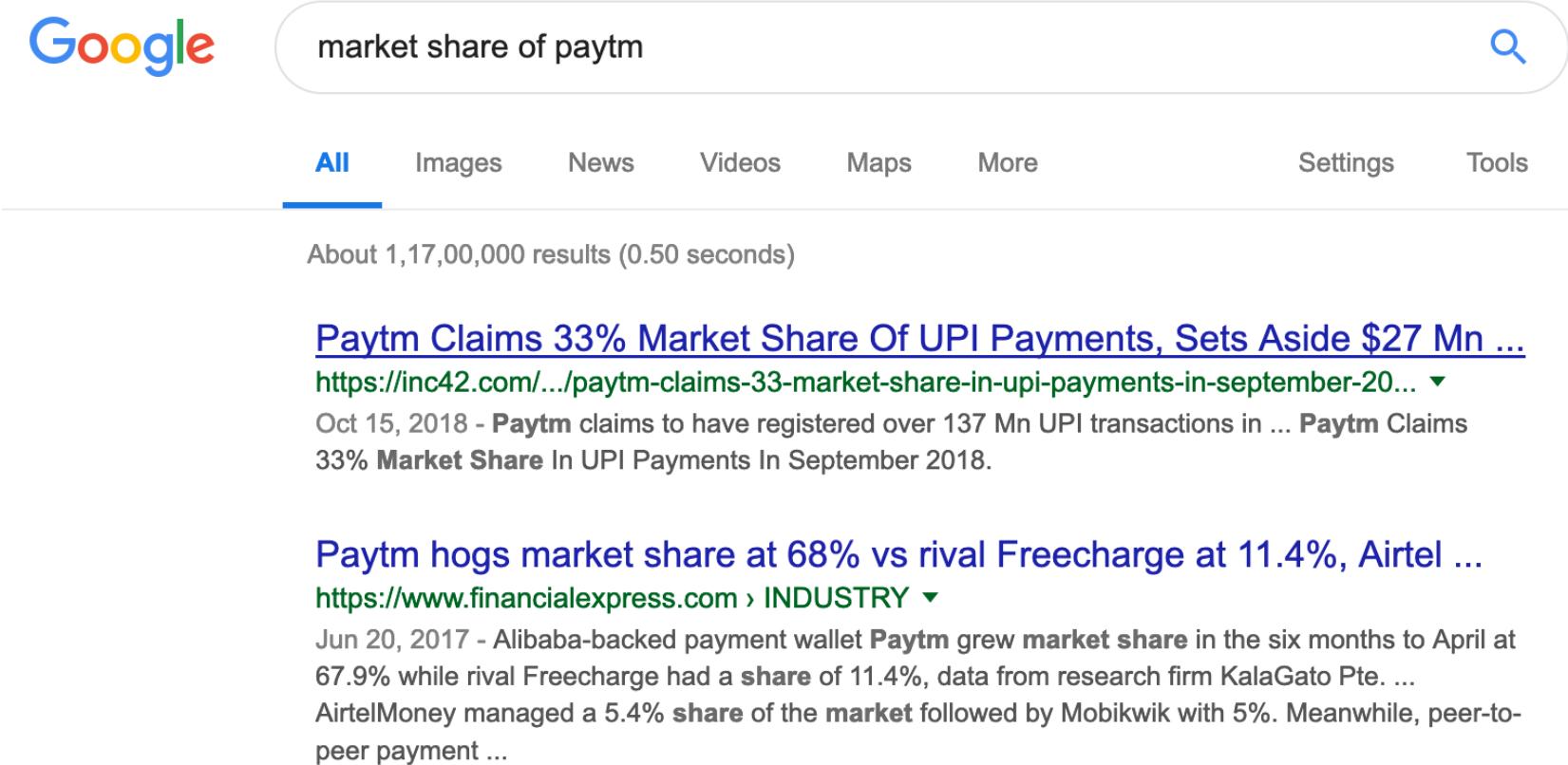
Try with this!!

Top 50 NBFCs' Ranking Based on Annual Turnover

NBFCs List	Total Income	Rank
Power Finance Corporation Limited	267377.40	1
Rural Electrification Corporation Limited	224403.10	2
Bajaj Finance Limited	133292.20	3
Shriram Transport Finance Company Limited	122768.30	4
Indian Railway Finance Corporation Limited	110202.32	5
Mahindra & Mahindra Financial Services Limited	72061.20	6
HDB Financial Services Limited	70619.90	7
Muthoot Finance Limited	62432.00	8
Cholamandalam Investment and Finance Company Limited	54257.60	9
L&T Finance Limited (erstwhile Family Credit Limited)	52460.00	10
Shriram City Union Finance Limited	51015.70	11
Tata Capital Financial Services Limited	45553.70	12

Words, Relevance & Personalization

- ❖ Market Share of **paytm**



A screenshot of a Google search results page. The search query "market share of paytm" is entered in the search bar. The results are filtered under the "All" tab. The first result is a news article from Inc42.com titled "Paytm Claims 33% Market Share Of UPI Payments, Sets Aside \$27 Mn ...". The second result is a news article from Financial Express titled "Paytm hogs market share at 68% vs rival Freecharge at 11.4%, Airtel ...". Both results show a snippet of the article's content.

market share of paytm

All Images News Videos Maps More Settings Tools

About 1,17,00,000 results (0.50 seconds)

[Paytm Claims 33% Market Share Of UPI Payments, Sets Aside \\$27 Mn ...](https://inc42.com/.../paytm-claims-33-market-share-in-upi-payments-in-september-20...)
https://inc42.com/.../paytm-claims-33-market-share-in-upi-payments-in-september-20... ▾
Oct 15, 2018 - Paytm claims to have registered over 137 Mn UPI transactions in ... Paytm Claims 33% Market Share In UPI Payments In September 2018.

[Paytm hogs market share at 68% vs rival Freecharge at 11.4%, Airtel ...](https://www.financialexpress.com/industry/paytm-hogs-market-share-at-68-vs-rival-freecharge-at-11-4-airtel-money-managed-a-5-4-share-of-the-market-followed-by-mobikwik-with-5-meanwhile-peer-to-peer-payment...)
https://www.financialexpress.com › INDUSTRY ▾
Jun 20, 2017 - Alibaba-backed payment wallet Paytm grew market share in the six months to April at 67.9% while rival Freecharge had a share of 11.4%, data from research firm KalaGato Pte. ... AirtelMoney managed a 5.4% share of the market followed by Mobikwik with 5%. Meanwhile, peer-to-peer payment ...

Words and Knowledge Bases

❖ Make your judgments on relevance

Paytm registers 600% growth in UPI transactions in 6 months ...

<https://www.hindustantimes.com/.../paytm.../story-iLTPwMRjKqj55fNCVnR5QK.html>

Nov 2, 2018 - Paytm on Friday said it witnessed 600 % growth in the Unified Payments Interface (UPI) payments with over 33 per cent of the overall market share. ... With this, Paytm has over 80 % share of all offline merchant ...

Paytm hogs market share at 68% vs rival Freecharge at 11.5%

<https://www.financialexpress.com/.../INDUSTRY>

PhonePe becomes largest UPI player, leaps ahead

<https://www.digit.in/.../Apps>

Aug 2, 2018 - Of these PhonePe had the largest market share of 40%. ... network, ahead of other payment apps like Paytm and Tez.

Berkshire Hathaway Takes Stake In India's Paytm -

<https://www.forbes.com/sites/.../08/.../berkshire-hathaway-takes-stake-in-paytm>

Aug 27, 2018 - Berkshire Hathaway has taken a stake in Paytm, India's largest digital payments company. It has invested \$1 billion and now holds the lion's share of the market with 9.9%, followed by PayPal ...



Type of business	Private
Type of site	E-commerce
Founded	2010
Headquarters	Noida, Uttar Pradesh, India
Area served	India, Canada
Founder(s)	Vijay Shekhar Sharma
Key people	Vijay Shekhar Sharma (CEO)
Industry	Internet
Products	Paytm Mall Paytm Payments Bank Paytm Money Paytm Gamepind
Services	Online shopping, payment systems, digital wallets
Revenue	₹814 crore (US\$110 million) (FY 2017) ^[1]
Parent	One97 Communications Ltd
Website	paytm.com

Text Collections

❖ Structured data

- ❖ Information stored DB
- ❖ Strict format
- ❖ Limitation
- ❖ Not all data collected is structured

❖ Semi-structured data

- ❖ Data may have certain structure but not all information collected has identical structure
- ❖ Some attributes may exist in some of the entities of a particular type but not in others

❖ Unstructured data

- ❖ Very limited indication of data type
- ❖ For example, look into a simple text document



Words and their occurrences

- ❖ Consider any city:
- ❖ Check with Wikipedia
 - ❖ For e.g, search for **Darjeeling** (English Version)
- ❖ **<https://en.wikipedia.org/wiki/Darjeeling>**
- ❖ How many times Darjeeling comes up in the document?
- ❖ Does it mean the following?
 - ❖ more it occurs ... more it is important?

Look at Text Segment

❖ Darjeeling is a city and a municipality in the Indian state of West Bengal. It is located in the Lesser Himalayas at an elevation of 6,700 ft. It is noted for its tea industry, its views of Kangchenjunga, the world's third-highest mountain, and the Darjeeling Himalayan Railway, a UNESCO World Heritage Site. Darjeeling is the headquarters of the Darjeeling District which has a partially autonomous status within the state of West Bengal. It is also a popular tourist destination in India.



Word Count – How do you get it?

❖ **Darjeeling** is a city and a municipality in the Indian state of West Bengal. It is located in the Lesser Himalayas at an elevation of 6,700 ft. It is noted for its tea industry, its views of Kangchenjunga, the world's third-highest mountain, and the **Darjeeling** Himalayan Railway, a UNESCO World Heritage Site. **Darjeeling** is the headquarters of the **Darjeeling** District which has a partially autonomous status within the state of West Bengal. It is also a popular tourist destination in India.



Word Count - Darjeeling

❖ **Darjeeling** is a city and a municipality in the Indian state of West Bengal. It is located in the Lesser Himalayas at an elevation of 6,700 ft. It is noted for its tea industry, its views of Kangchenjunga, the world's third-highest mountain, and the **Darjeeling** Himalayan Railway, a UNESCO World Heritage Site. **Darjeeling** is the headquarters of the **Darjeeling** District which has a partially autonomous status within the state of West Bengal. It is also a popular tourist destination in India.

Word Count - Darjeeling

❖ Darjeeling is a city and a municipality in the Indian state of West Bengal. It is located in the Lesser Himalayas at an elevation of 6,700 ft. It is noted for its tea industry, its views of Kangchenjunga, the world's third-highest mountain, and the Darjeeling Himalayan Railway, a UNESCO World Heritage Site.

Darjeeling is the headquarters of the Darjeeling District which has a partially autonomous status within the state of West Bengal. It is also a popular tourist destination in India.

Term Statistics: Text Data

- How do we extract the term statistics?

Tokenization

Darjeeling is a city and a municipality

in the Indian state of West Bengal.

All TERMS (WORDS)

Darjeeling – 1	in – 1
is – 1	the – 1
a – 1	Indian – 1
city – 1	state – 1
and – 1	of – 1
a – 1	west – 1
municipality – 1	Bengal – 1

Unique TERMS (WORDS)

Darjeeling – 1	the – 1
is – 1	Indian – 1
a – 2	state – 1
city – 1	of – 1
and – 1	West – 1
municipality – 1	Bengal – 1
in – 1	

Words – Counts

❖ A subset of words and their counts are given below:

7 the

5 of

5 is

5 a

4 Darjeeling

3 in

3 It

2 state

2 its

2 and

2 West

2 Bengal

1 world's

1 within

1 which

1 views

1 tourist

... .



Look at 3 documents

- d₁** - **Darjeeling** is a city and a municipality in the Indian state of West Bengal. It is located in the Lesser Himalayas at an elevation of 6,700 feet
- d₂** - **Darjeeling** is noted for its tea industry, its views of Kangchenjunga, the world's third-highest mountain, and the **Darjeeling** Himalayan Railway, a UNESCO World Heritage Site
- d₃** - **Darjeeling** is the headquarters of the **Darjeeling** District which has a partially autonomous status within the state of West Bengal. It is also a tourist destination in India

Summary

Focused on exploring the following:

- Fundamentals of an Information Retrieval Systems
- Handling Structured and Unstructured Data
- Collecting Term Statistics
 - From Unstructured Text Data
- How to Create an Index / Inverted Index?
 - Many more to come up ... stay tuned in ...



Help among Yourselves?

- **Perspective Students** (having CGPA above 8.5 and above)
- **Promising Students** (having CGPA above 6.5 and less than 8.5)
- **Needy Students** (having CGPA less than 6.5)
 - Can the above group help these students? (Your work will also be rewarded)
- You may grow a culture of **collaborative learning** by helping the needy students



How to reach me?

→ Please leave me an email:

rajendra [DOT] prasath [AT] iiits [DOT] in

→ Visit my homepage @

→ <https://www.iiits.ac.in/people/regular-faculty/dr-rajendra-prasath/>

(OR)

→ <http://rajendra.2power3.com>



Assistance

- You may post your questions to me at any time
- You may meet me in person on available time or with an appointment
- You may ask for one-to-one meeting

Best Approach

- You may leave me an email any time
(email is the best way to reach me faster)





Questions

It's Your Time



THANKS

