

# Information Retrieval

Venkatesh Vinayakarao

Term: Aug – Dec, 2018

Indian Institute of Information Technology, Sri City



My whole life, I've been a seeker, searching for something.  
- Mike White.



# Musings from the Real World

# Disclaimer

Most examples and discussions in this talk revolve around google and bing. This is just to share with you, my industry experiences. Please keep in mind that ***IR is beyond search engines.***

# Our Agenda: The Beauty of IR!

## Offline Horror!

Crawling



Content Processing

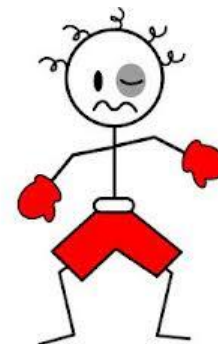


Indexing



How to process Korean queries for local listings?

Me!



## Online Terror!

Query (Intent) Understanding



Ranking



User Interface



# Crawling

- How frequently should we crawl?
  - Fresh & Super-Fresh! How to crawl cricket scores? Are we even crawling here?

Google search for "scorecard india vs new zealand live" showing search results for a cricket match.

India in New Zealand, 2 Test Series, 2014  
Wednesday, February 5, 5:00 PM  
Eden Park, Auckland

New Zealand	vs.	India
503/10 (121.4)	1st Innings	202/10 (60)
105/10 (41.2)	2nd Innings	366/10 (96.3)

New Zealand won

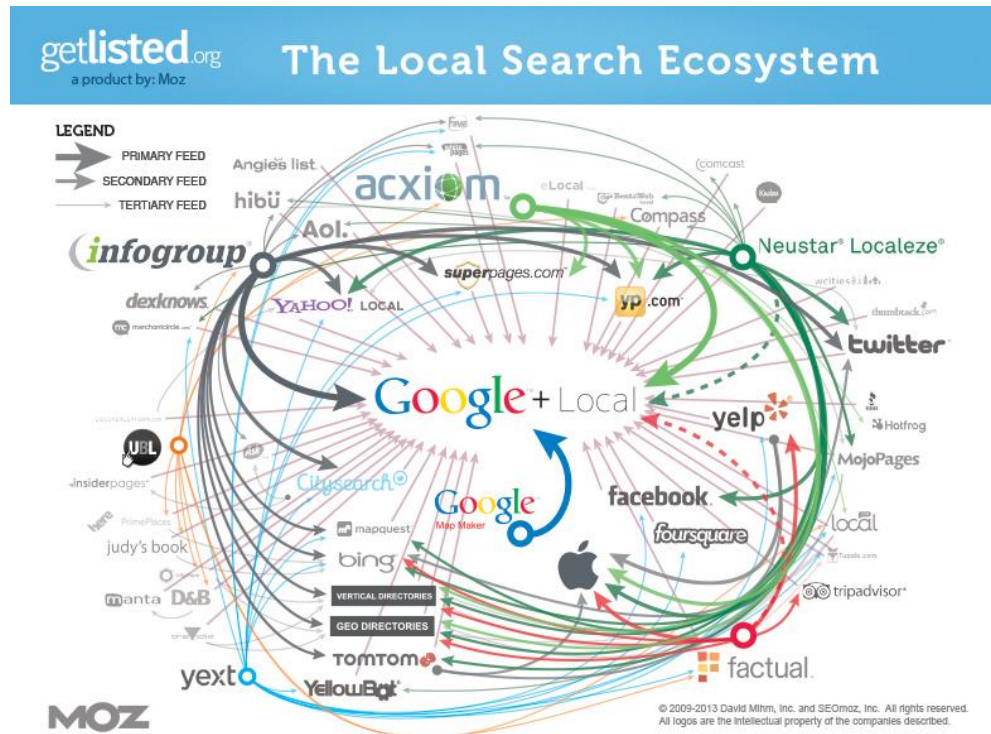
Thu, Feb 13 New Zealand India 5:00 PM

All times are in Eastern Time

- How to avoid 404 - Page not found?
- How much time did it take google to show your first personal page?

# Content Processing

- Good Read: <https://getlisted.org/static/resources/local-search-data-providers.html>



# Content Processing

- Query: “Schools in Delhi”
  - Answer: “Delhi Public School”
  - Good or Bad?
- Query: “Schools in Hyderabad”
  - Answer: “Delhi Public School”
  - Good or Bad?
- Query: “Hotels in Bombay”
  - Answer: “Grand Hyatt, Mumbai”
  - Good or Bad?
  - How to get same results for both Mumbai and Bombay?

# Content Processing & Indexing

- A real example:

[http://www.yelp.com/dataset\\_challenge/](http://www.yelp.com/dataset_challenge/)

## business

```
{
  'type': 'business',
  'business_id': (encrypted business id),
  'name': (business name),
  'neighborhoods': [(hood names)],
  'full_address': (localized address),
  'city': (city),
  'state': (state),
  'latitude': latitude,
  'longitude': longitude,
  'stars': (star rating, rounded to half-stars),
  'review_count': review count,
  'categories': [(localized category names)]
  'open': True / False (corresponds to closed, not business hours),
}
```

## Enriched Business

- Category Synonyms (for eg., auto service & car service are replaceable at times)
- User's query forms (for eg., McDonalds is commonly queried as McD)



# Derived Values & Indexing

- Given a location, how will you find all businesses within 1km radius?
  - Query: schools near govindpuri delhi

[Shri Satya Sai Vidya Vihar School](#)  
[plus.google.com](#)  
4 Google reviews

[Udgam Play School](#)  
[www.udgamonline.com](#)  
Google+ page

[Indo Westn Dance Claes - Rock Strng](#)  
[plus.google.com](#)  
1 Google review

A Kalkaji Extension,  
Govindpuri  
New Delhi, Delhi (state)

B Mor Pocket - 104,  
Kalkaji, Near Balaji  
Estate  
Govindpuri Extension,  
Govindpuri, New Delhi,  
DL  
011 3296 6675

C D 24, 2nd Floor, Kalkaji  
Govindpuri Extension,  
Govindpuri, New Delhi,  
NI

# Query Understanding Challenge

# Rules

- I will give an entity name.
- You will have to frame at least three different (dissimilar) queries (and as many as you can) that give same document as the correct result at first place.
- At the end, you should submit:
  - Query, Max. no. of top n correct results that you maintained to be same.
- You will have 5 minutes.

# Questions

- Tom Cruise
- Aishwarya Rai
- Tom Hanks
- Venkatesh Vinayakarao
- Amir Khan
- Andre Agassi
- Manmohan Singh

# Query Understanding

- Query: Michael Jordon
  - Which MJ to return? The basketball player or actor?
- Factors
  - User profile
  - Query context (session details, browser data, links, etc)
  - ...
- Query: Delhi School
  - What does user want? “Delhi Public School” or “Schools in Delhi” or “some Indian school in US”?
- Query: “IR”
  - Predict top three results

# Query = IR

- Results

- Indian Railways
- IR – Wikipedia
- International Relations – DoT
- What is Infrared Radiation? Definition from WhatIS.com

# Ok! I give up!!

- A frustrated search user: “please show me some t-shirt brands”

[I want to make new brand for my shirt factory please suggest me a ...](#)

[in.answers.yahoo.com](#) > ... > Fashion & Accessories ▼ Yahoo! ▼

Jan 25, 2013 - Make Yahoo Your Homepage ... Show me another ». I want to make new brand for my shirt factory please suggest me a unique name for that?

<a href="#">Please suggest me Best brand in jeans and shirt?</a>	7 answers	8 Sep 2013
<a href="#">Please suggest brand name boys t.shirt?</a>	4 answers	12 Mar 2012
<a href="#">Brand Name suggestion : Can you please suggest me ...</a>	4 answers	8 Jun 2011
<a href="#">Please name the best brands of T Shirts available in ...</a>	7 answers	20 Nov 2007

[More results from in.answers.yahoo.com](#)

[Make Money Blogging in 31 Days: Selling Brand T-Shirts | Your ...](#)

[www.yourdreamblog.com/selling-brand-t-shirts-blogging/](#) ▼

Oct 23, 2013 - Selling brand t-shirts from your blog is not only a great way to earn income ... Please show me your thankful for writing this article for you by ...

[T-Shirts in Durban City | Gumtree South Africa](#)

[www.gumtree.co.za](#) > ... > Men's Clothing ▼

# More Fun with Auto Completion

The image shows a screenshot of a Google search interface with four search bars, each displaying auto-completion suggestions. The first search bar contains the text "never put a" and suggests "never put a **woman on a pedestal**", "never put a **comma before and**", "never put a **sock in a toaster meaning**", and "never put a **preposition at the end of a sentence**". The second search bar contains "what would happen" and suggests "what would happen **if there was no moon**", "what would happen **if yellowstone erupts**", "what would happen **without enzymes**", and "what would happen **if the moon was destroyed**". The third search bar contains "Lectures are" and suggests "lectures are **useless**", "lectures are **boring**", "lectures are **ineffective**", and "lectures are **a waste of time**". The fourth search bar contains "University" and suggests "university **of phoenix**", "university **of washington**", "university **of michigan**", and "university **of alabama**". Each search bar includes a microphone icon and a search button.

Google

never put a  
never put a **woman on a pedestal**  
never put a **comma before and**  
never put a **sock in a toaster meaning**  
never put a **preposition at the end of a sentence**

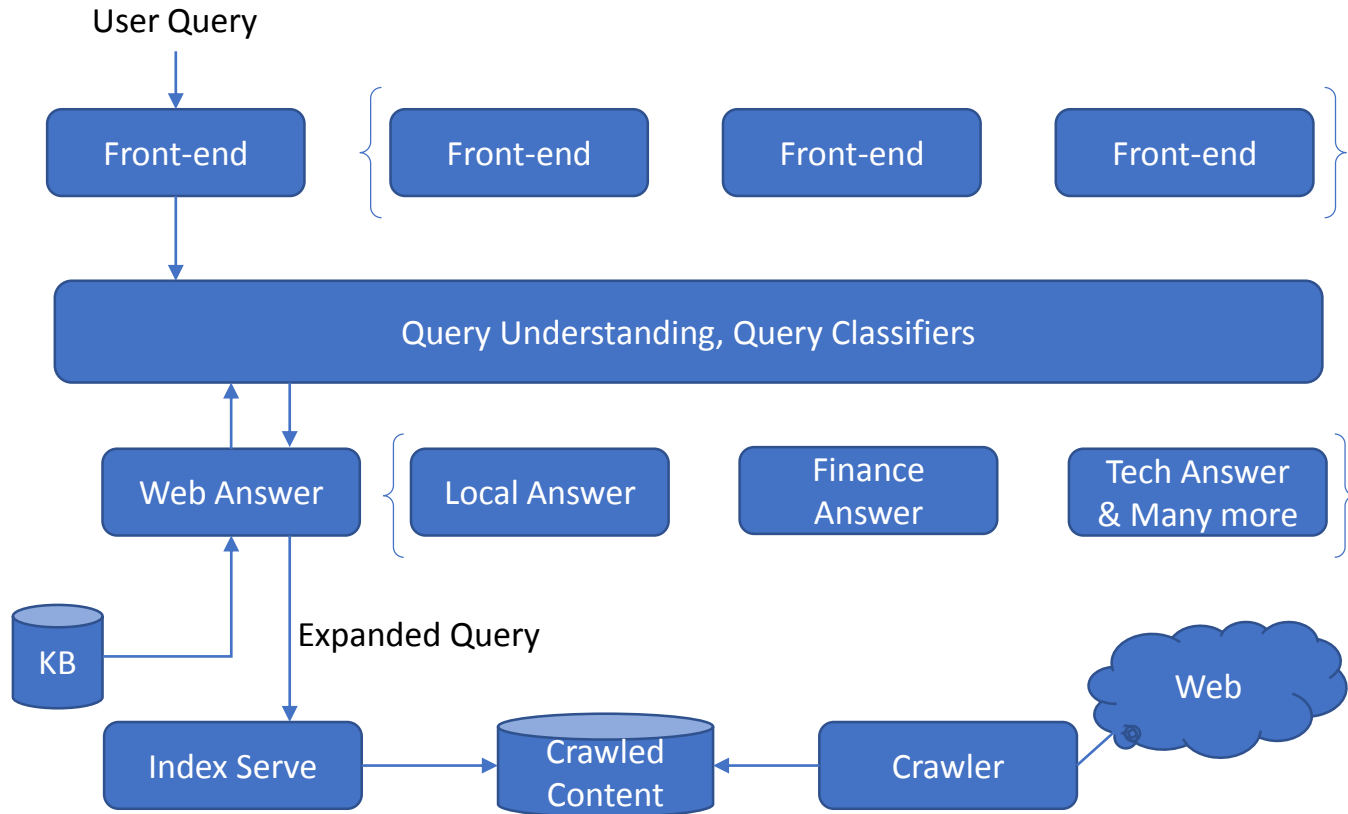
what would happen  
what would happen **if there was no moon**  
what would happen **if yellowstone erupts**  
what would happen **without enzymes**  
what would happen **if the moon was destroyed**

Lectures are  
lectures are **useless**  
lectures are **boring**  
lectures are **ineffective**  
lectures are **a waste of time**

University  
university **of phoenix**  
university **of washington**  
university **of michigan**  
university **of alabama**



# System Overview



# Ranking & Relevance

- How do we know if the document is relevant (in web search context)?
  - Popularity of url
  - Domain score (is it ac.in or .edu?)
  - Entity, Chain entity?
  - Trust Factor (Wikipedia?)
  - Inlinks/Outlinks
  - Position of query terms
  - Sequence of query terms
  - ... and 1000 of such things

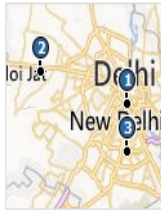
# Are Current Search Engines Good at Relevance & Ranking?

Query1: Vegetarian hotels in south delhi

## Bing

### [Vegetarian Hotels South near Delhi, Delhi](#)

Bing Local

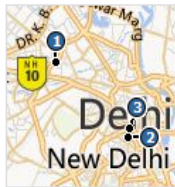


1. [Saravana Bhavan](#) · +91-1123316060  
P 13, Connaught Circus, Near Nalli Silks, Connaug...  
[Directions](#)
2. [The Royal Residency Hotel](#) · +91-1126833913  
Hotel-20,C-669,New Friends Colony, South, Delhi...  
[Directions](#)
3. [Melrose Guest House](#) · +91-1126257318  
A-27, South Extension Part Ii, Delhi, Delhi 110049 · [Directions](#)

Query2: South Indian hotels in south delhi

### [South Indian Hotels South near Delhi, Delhi](#)

Bing Local



1. [South Indian Hotel](#)  
Maharshi Balaram Marg, Shakurpur, Delhi, Delhi 1...  
[Directions](#)
  2. [The Spice Route](#) · +91-1141116605  
The Imperial Hotel, Lobby Level, Janpath Road, In ...  
[Directions](#)
  3. [Saravana Bhavan](#) · +91-1123316060  
P 13, Connaught Circus, Near Nalli Silks, Connaug...  
[Directions](#)
- [More listings](#)

## Google

### [Naveen Veg Hotel](#)

<https://plus.google.com/113812899091404930078/about?gclid=...>  
[Google+ page](#) · [Be the first to review](#)

Block E, Katwaria Sarai, New Delhi, DL 110016  
095 60 837948

### [Kerala Food Channel](#)

[www.keralafoodchannel.com](http://www.keralafoodchannel.com)  
3.6 ★★★★★ 6 Google reviews

### [NEW MADRAS HOTEL](#)

[plus.google.com](http://plus.google.com)  
3.8 ★★★★★ 12 Google reviews

### [Naivedyam](#)

[plus.google.com](http://plus.google.com)  
4.1 ★★★★★ 28 Google reviews

### [Ever Green Sweet House](#)

[www.evergreensweethouse.com](http://www.evergreensweethouse.com)  
3.3 ★★★★★ 22 Google reviews

### [Naivedyam](#)

[plus.google.com](http://plus.google.com)  
4.1 ★★★★★ 24 Google reviews

A DDA Flats, DDA Flats  
Kalkaji, Block L 2,  
Govindpuri  
New Delhi, DL  
097 17 789571

B Shop No. 7, Lodhi Road  
Khanna Market, Lodi  
Colony, New Delhi, DL  
011 2461 1726

C 1, Hauz Khas Village  
Hauz Khas Tank, Deer  
Park, Hauz Khas, New  
Delhi, Delhi  
011 2696 0426

D S/30 Green Park Market,  
Hauz Khas, Hauz Khas  
New Delhi, DL  
011 2651 4646

E F-12, Kalkaji Main Road,  
Near Deshbandhu  
College, Kalkaji, New  
Delhi, Delhi, 110016

# ...More Examples

Query3: South Indian restaurants in south delhi

## South Indian Restaurants South near Delhi, Delhi

Bing Local



1. **South Indian** · +91-8471029965  
Service Road, Laxmi Nagar, Delhi, Delhi · [Dir](#)
2. **South Indian Restaurant** · +91-1128751019  
Daulatram Ahuja Marg, Karol Bagh, Delhi, De  
[Directions](#)
3. **The Spice Route** · +91-1141116605  
The Imperial Hotel, Lobby Level, Janpath Roa  
[Directions](#)

[Ever Green Sweet House](#)  
www.evergreensweethouse.com  
3.3 ★★★★★ 22 Google reviews

[Naivedyam](#)  
plus.google.com  
4.1 ★★★★★ 28 Google reviews

[Spice Water Trail Restaurant](#)  
www.spicewatertrail.com  
1 Google review

[Naivedyam](#)  
plus.google.com  
4.1 ★★★★★ 24 Google reviews

[Dakshin](#)  
plus.google.com

**A** S/30 Green Park Market,  
Hauz Khas, Hauz Khas  
New Delhi, DL  
011 2651 4646

**B** 1, Hauz Khas Village  
Hauz Khas Tank, Deer  
Park, Hauz Khas, New  
Delhi, Delhi  
011 2696 0426

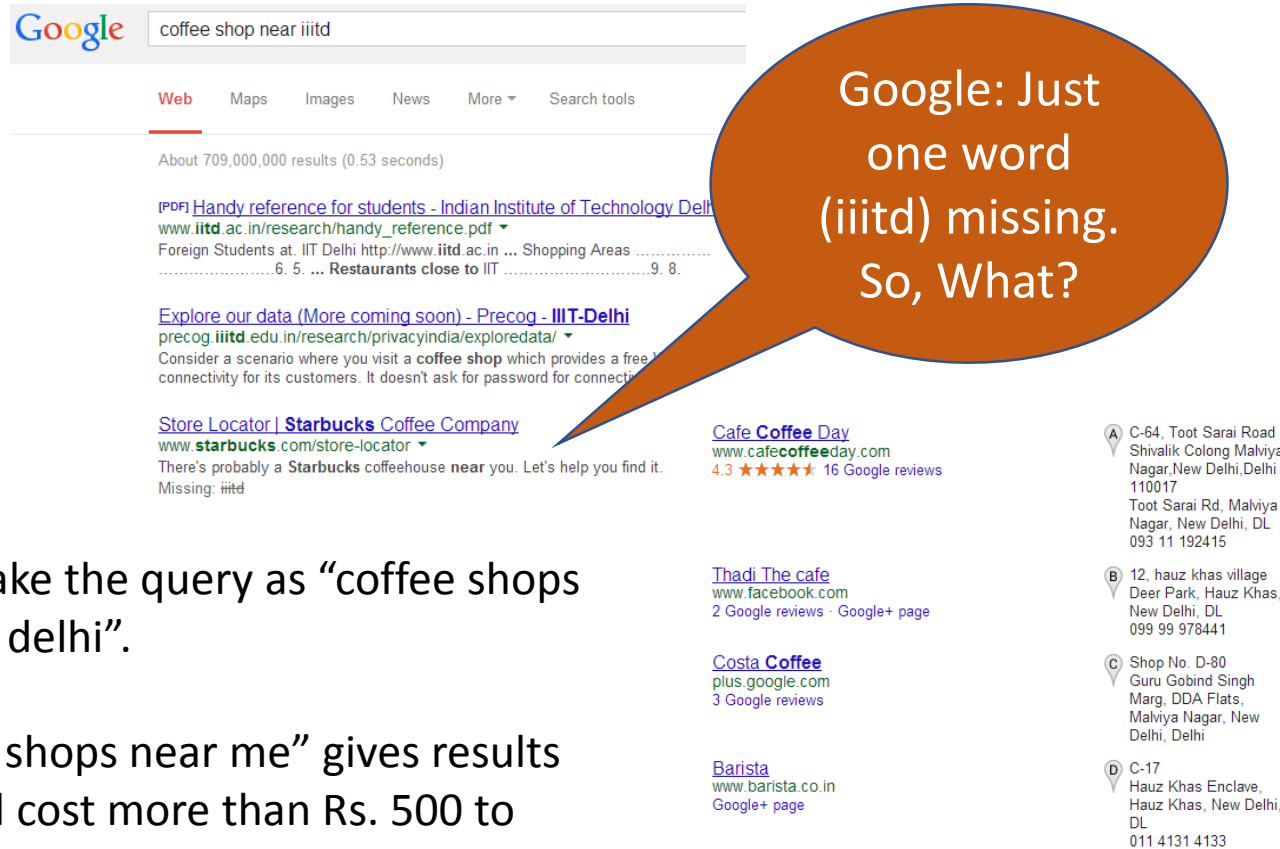
**C** M-24 Greater Kailash  
Part 1, M Block Market  
New Delhi, DL  
011 3089 4731

**D** F-12, Kalkaji Main Road,  
Near Deshbandhu  
College, Kalkaji, New  
Delhi, Delhi, 110019  
Kalkaji Main Rd, Block  
F, Kalkaji, New Delhi, DL  
011 2623 6364

**E** District Centre, Saket  
District Centre, Sector 6

What's the difference between query2 and query3?  
Should search engines give different results?

# How Far for a Coffee?



Google coffee shop near iiitd

Web Maps Images News More Search tools

About 709,000,000 results (0.53 seconds)

[Handy reference for students - Indian Institute of Technology Delhi](#)  
www.iitd.ac.in/research/handy\_reference.pdf  
Foreign Students at IIT Delhi http://www.iitd.ac.in ... Shopping Areas .....6. 5. ... Restaurants close to IIT .....9. 8.

[Explore our data \(More coming soon\) - Precog - IIT-Delhi](#)  
precog.iitd.edu.in/research/privacyindia/exploredata/  
Consider a scenario where you visit a **coffee shop** which provides a free connectivity for its customers. It doesn't ask for password for connecti

[Store Locator | Starbucks Coffee Company](#)  
www.starbucks.com/store-locator  
There's probably a **Starbucks** coffeehouse near you. Let's help you find it.  
Missing: iiitd

[Cafe Coffee Day](#)  
www.cafecoffeeday.com  
4.3 ★★★★★ 16 Google reviews

[Thadi The cafe](#)  
www.facebook.com  
2 Google reviews · Google+ page

[Costa Coffee](#)  
plus.google.com  
3 Google reviews

[Barista](#)  
www.barista.co.in  
Google+ page

A C-64, Toot Sarai Road  
Shivalik Colong Malviya  
Nagar, New Delhi, Delhi  
110017  
Toot Sarai Rd, Malviya  
Nagar, New Delhi, DL  
093 11 192415

B 12, hauz khas village  
Deer Park, Hauz Khas,  
New Delhi, DL  
099 99 978441

C Shop No. D-80  
Guru Gobind Singh  
Marg, DDA Flats,  
Malviya Nagar, New  
Delhi, Delhi

D C-17  
Hauz Khas Enclave,  
Hauz Khas, New Delhi,  
DL  
011 4131 4133

Google: Just one word (iiitd) missing. So, What?

Let's make the query as "coffee shops near iiit delhi".

"Coffee shops near me" gives results that will cost more than Rs. 500 to reach!

# A Lot has Changed over Coffee

## Café Coffee Day

3.9 ★★★★★ (249) · Coffee Shop

3.6 km · Inside Hpcl Petrol Bunk, Guntur - Chennai Highway, Near Tada Railway Stat...

Late-night food · Cosy · Casual



## New City Hot and Cold

4.4 ★★★★★ (10) · Cafe

2.7 km · Central Expy

Casual · Groups



## Cafe Tada

3.6 ★★★★★ (5) · Cafe

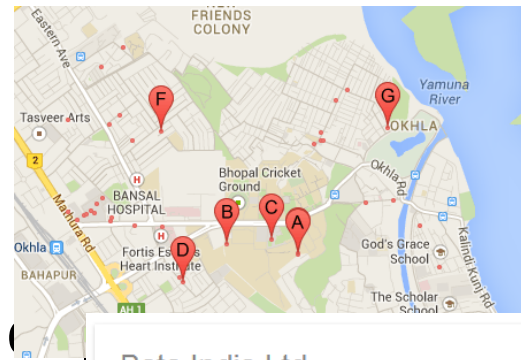
3.3 km

Casual · Good for kids · Groups



# User Interface

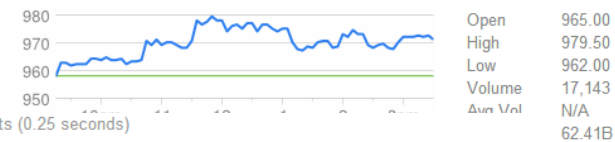
- Is UI important for search (
  - Maps in local results
  - Live sport score cards
  - Finance tickers
- Search Elements
  - Filters
  - Search Operators
  - Entity Infoboxes



## Bata India Ltd.

BSE: 500043 - Feb 7 3:30 PM IST

971.20 +13.15 (1.37%)



About 122,000,000 results (0.25 seconds)

India in New Zealand, 2 Test Series, 2014  
Thursday, 6 February, 3:30 am  
Eden Park, Auckland

 New Zealand	vs.	India 
503/10 (121.4)	1st Innings	202/10 (60)
105/10 (41.2)	2nd Innings	366/10 (96.3)
<b>New Zealand won</b>		

Fri, 14-Feb  New Zealand 3:30 am 14/02  
 India

# Summary: The Beauty of IR is its Challenges!

## Offline Horror!

Crawling



Content Processing

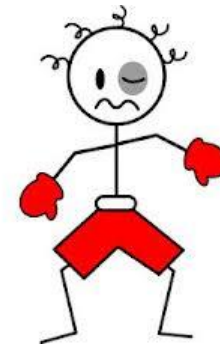


Indexing



How to process Korean queries for local listings?

Me!



## Online Terror!

Query (Intent) Understanding



Ranking



User Interface





# Announcements

# Mid-Term 1 Syllabus

- Following Chapters from Manning's IR Book
  - Boolean Retrieval
  - Term Vocabulary and Posting Lists
  - Dictionaries and Tolerant Retrieval
  - Index Construction
  - Index Compression
- And, anything else discussed in the class.

# Off – Next Two Days

- I am on leave on 29<sup>th</sup> and 30<sup>th</sup>.
- Classes on 30<sup>th</sup> are cancelled. Make-up Classes after Mid-Term 1.

# Mock Mid-Term 1

- On 6<sup>th</sup> Sep.
- You will take it individually.