Inspire…Educate…Transform.

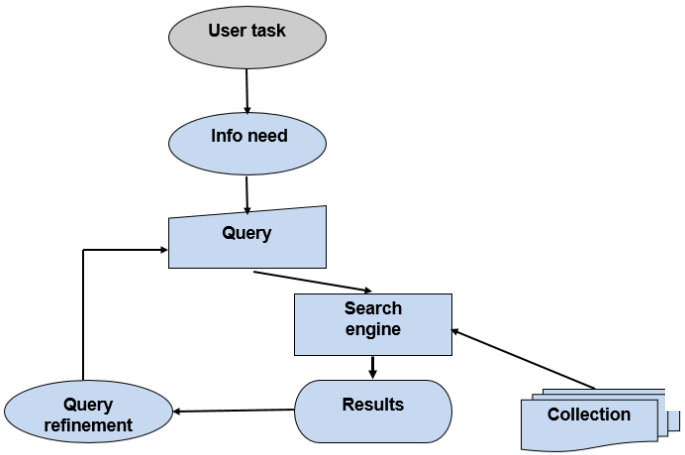
Summary

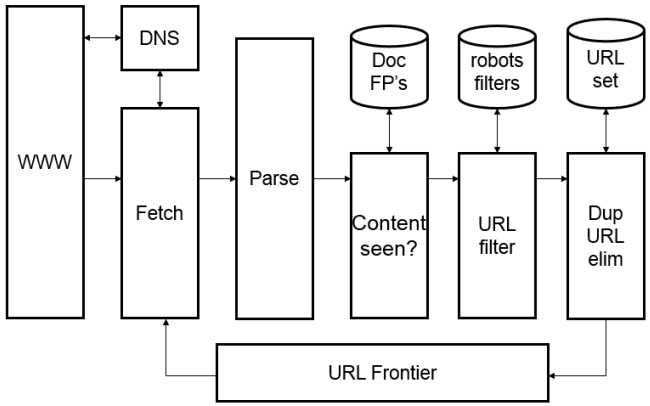
Dr. Manish Gupta (manishg.iitb@gmail.com)
Sr. Mentor – Academics, INSOFE

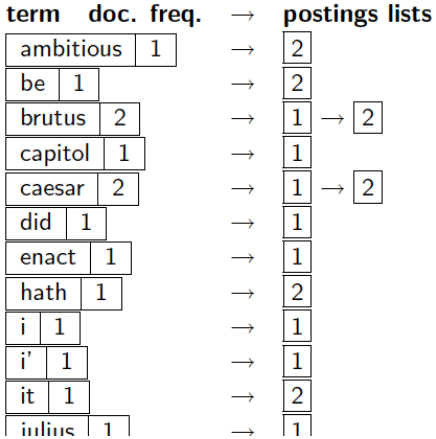
Course Content

- Collection of three main topics of high recent interest.
  - Search engines (Crawling, Indexing, Ranking)
    - Language Modeling
    - Text Indexing and Crawling
    - Relevance Ranking
    - Link Analysis Algorithms
  - Text Processing (NLP, NER, Sentiments)
    - Named Entity Recognition
    - Natural Language Processing
    - Sentiment Analysis
    - Summarization
  - Social networks (Properties, Influence Propagation)
    - Social Network Analysis
    - Influence Propagation in Social Networks

Search Engine Pipeline

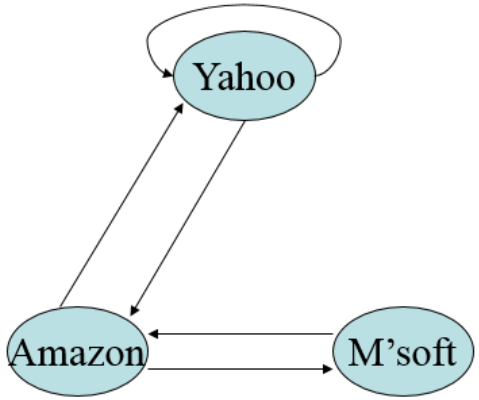






$$w_{t,d} = \log(1+\mathrm{tf}_{t,d}) \times \log_{10}(N/\mathrm{df}_t)$$

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\bullet\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\bullet\frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$



$$\mathrm{NDCG}_q = Z_q \sum_{j=1}^{L} \frac{2^{r_q(j)} - 1}{\log(1+j)}$$

Search Engine Pipeline

- Tools
  - Crawler: Nutch, Heritrix
  - Indexing mechanisms: Lucene, BerkeleyDB, MG4J
  - Search: Lucene
- Typically useful if you have a new web portal which needs search support
  - Job portal
  - Products portal
- Design choices
  - What (which fields) to index
  - What factors are important for showing relevant listings

Text Processing Pipeline

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA … ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON	ORG	LOC

Einstein met with UN officials in Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕…

The 13th Shanghai International Film Festival…

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

→ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

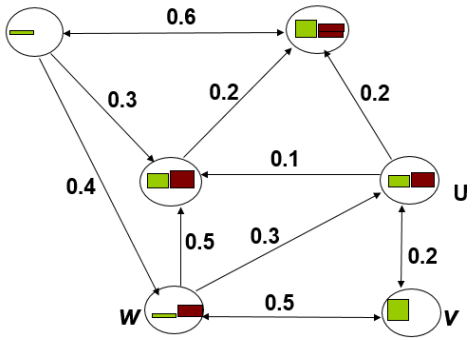
Text Processing Pipeline

- Typical steps
  - Sentence Splitting, Tokenization, Normalization, Stemming, Phrasing, POS/NER, Parsing: Constituency Parse, Dependency Parse
- Useful for many applications like
  - Sentiment analysis, Word sense disambiguation, Summarization
  - Text categorization, Question Answering, Information Extraction
  - Machine Translation, etc.
- Tools
  - Stanford CoreNLP
  - Various packages in R
  - NLTK
  - GATE

Social Network Analysis Pipeline

- Degree distributions
  - Heavy tail
- Clustering
  - High clustering coefficient
- Communities and dense subgraphs
  - Abundance; locally dense, globally dense; spectrum
- Connected components
  - Distribution; bow-tie structure
- Connectivity
  - Low diameter; small-world properties

Influence Maximization



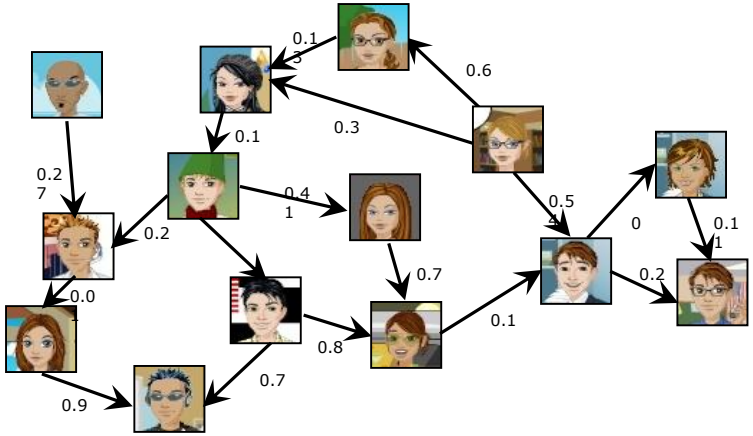
Erdös-Renyi Model

Small World Model and Kleinberg's Model

Power Laws

Preferential Attachment Model

Copying Model, Forest Fire Model



Influence Propagation

Social Network Analysis Pipeline

- Steps
  - Obtain/Store the network, Perform various network analysis tasks
- Useful for
  - Generating synthetic networks
  - Using Twitter data to show recent tweets for a product/event
  - Using Twitter data for extracting sentiment about products or their features
  - Influence Maximization for social advertising, opinion leader finding, influential people in a network.
- Tools: http://www.kdnuggets.com/2015/06/top-30-social-network-analysis-visualization-tools.html

Thanks

manishg.iitb@gmail.com

International School of Engineering

Plot 63/A, 1st Floor, Road # 13, Film Nagar, Jubilee Hills, Hyderabad - 500 033

For Individuals: +91-9502334561/63 or 040-65743991

For Corporates: +91-9618483483

Web: http://www.insofe.edu.in

Facebook: https://www.facebook.com/insofe

Twitter: https://twitter.com/Insofeedu

YouTube: http://www.youtube.com/InsofeVideos

SlideShare: http://www.slideshare.net/INSOFE

LinkedIn: http://www.linkedin.com/company/international-school-of-engineering