# CS3245 Information Retrieval

#### Lecture 12: Crawling and Link Analysis





Live Q&A https://pollev.com/jin

#### Last Time



Chapter 11

1. Classic Probabilistic Approaches for IR

Chapter 12

1. Language Models for IR

Information Retrieval



## Today – The Web

#### Chapter 20

Crawling

#### Chapter 21

Link Analysis

Google	how search works	× 🎍 ९			
	Q All ▶ Videos 🖆 Images 🖓 Maps 🖽 News 🗄 More	Settings Tools			
	About 3,610,000,000 results (0.55 seconds)				
	https://www.google.com > search > howsearchworks *				
	Google Search - Discover How Google Search Works				
	Wondering how Google <b>search works</b> ? Learn how Google looks through and organizes all the information on the internet to give you the most useful and				
	Algorithm · Crawling · Our Mission · Maximize access				
	https://developers.google.com > search > docs > beginner 💌				

#### How Google Search Works | Search Central | Google ...

Googlebot uses an algorithmic process to determine which sites to crawl, how often, and how many pages to fetch from each site. Google's crawl process begins ...

https://moz.com > Introduction \*

#### How Search Engines Work: Crawling, Indexing, and Ranking ...

When someone performs a **search**, **search** engines scour their index for highly relevant content and then orders that content in the hopes of solving the searcher's ...

https://web.dev > how-search-works \*

#### How search works - web.dev

5 Nov 2018 — When a user **searches** for something, **search** engines determine the most useful results and then show them to the user. Ranking, or ordering, the ...





# How hard can crawling be?

- Web search engines must crawl their documents.
- Getting the content of the documents is easier for many other IR systems.
  - E.g., indexing all files on your hard disk: just do a recursive descent on your file system
- Bandwidth, latency...
  - Not just on crawler side, **but also on the web server side**.



# How hard can crawling be?

Google Search

#### How Google searches 30 trillion web pages, 100 billion times a month

I'm Feeling

JOHN KOETSIER, TUNE MARCH 1, 2013 12:43 PM TAGS: FEATURED, GOOGLE, INDEX, SEARCH ENGINE, WEB SEARCH To fetch 30T pages in one month, we need to fetch almost 11M pages per second!

Actually many more since many of the pages we attempt to crawl will be duplicates, unfetchable, spam etc.

the droids we're looking for

http://venturebeat.com/2013/03/01/how-google-searches-30-trillion-web-pages-100-billion-times-a-month/

#### NUS National University of Singapore

# **Basic crawler operation**

- Initialize queue with URLs of known seed pages
- Repeat
  - Take URL from queue
  - Fetch and parse page
  - Extract URLs from page
  - Add URLs to queue
  - Call the indexer to index the page
- What's wrong with this Crawler?

{https://en.wikipedia.org/,
https://nlp.stanford.edu/IR-book/, ...}

https://en.wikipedia.org/

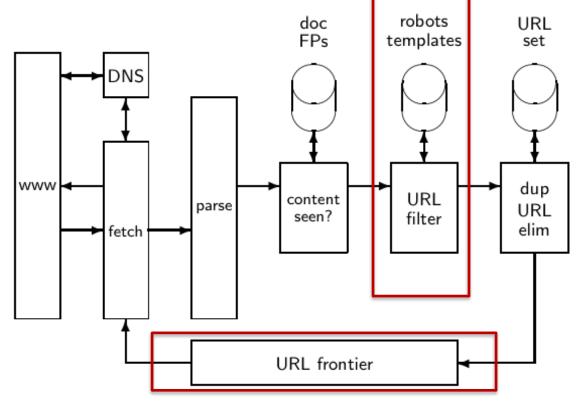
https://en.wikipedia.org/wiki/Main\_Page, ...

{https://nlp.stanford.edu/IR-book/, ...,
https://en.wikipedia.org/wiki/Main\_Page, ...}



#### Politeness

- Space out the requests for a site
- Crawl only the allowed pages



URLs crawled

and parsed

#### **URL Frontier**

URL frontier:

found, but

not yet crawled

unseen URLs



- The URL frontier is the data structure that holds and manages URLs we've seen, but not crawled yet.
  - May include multiple pages from the same host but must avoid trying to fetch them all at the same time



#### Robots.txt

- Protocol for giving crawlers ("robots") limited access to a website, originally from 1994
- Example:
  - # Observed spamming large amounts of https://en.wikipedia.org/?curid=NNNNNN

User-agent: MJ12bot

Disallow: /

# Wikipedia work bots:

User-agent: IsraBot

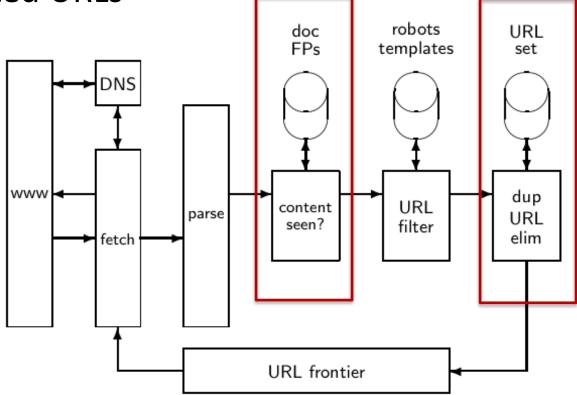
Disallow:

**Important**: cache the **robots.txt** file of each site we are crawling



### Duplicate detection

- Duplicated contents
- Duplicated URLs



#### Content seen



- For each page fetched: check if the content is already in the index
- Check this using document fingerprints or shingles
- Skip documents whose content has already been indexed



#### URL seen

- Duplicate elimination
  - Ignore the URLs which have been seen before.
- Normalization
  - We need to normalize (expand) all relative URLs.
    - E.g., at <u>http://mit.edu</u>, we may have aboutsite.html which is in fact <u>http://mit.edu/aboutsite.html</u>

#### Freshness

Crawl some pages (e.g., news sites) more often than others



#### Scalability

- Run multiple crawl threads
- Use different (geographically distributed) nodes

#### Data center locations

We own and operate data centers around the world to keep our products running 24 hours a day, 7 days a week. Find out more about our data center locations, community involvement, and job opportunities in our locations around the world.



#### Americas

Berkeley County, South Carolina Council Bluffs, Iowa Douglas County, Georgia Jackson County, Alabama Lenoir, North Carolina Mayes County, Oklahoma Montgomery County, Tennessee Quilicura, Chile The Dalles, Oregon

#### Asia

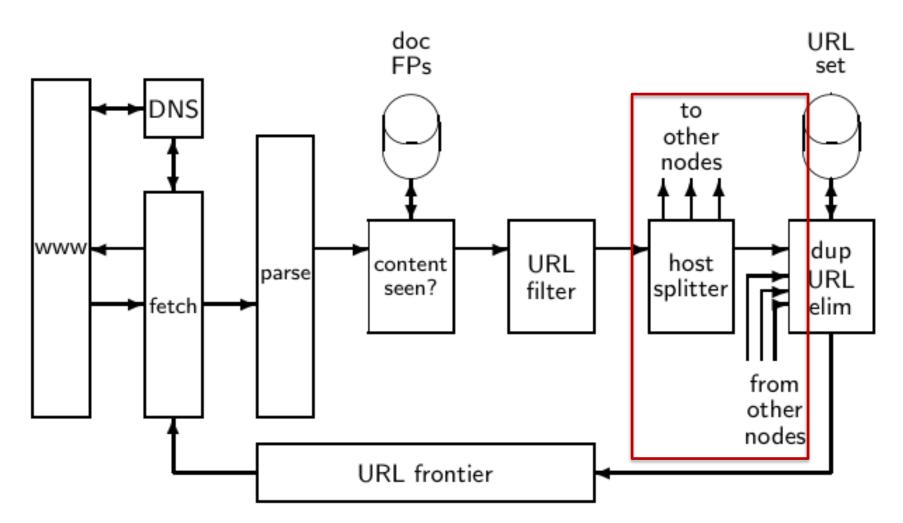
Changhua County, Taiwan Singapore

#### Europe

Dublin, Ireland Eemshaven, Netherlands Hamina, Finland St Ghislain, Belgium



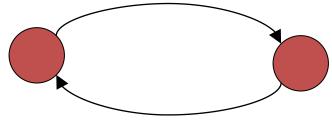
# Distributed crawling architecture



### Spider traps



- A set of webpages that cause the crawler to go into an infinite loop or crash
  - A simple loop back



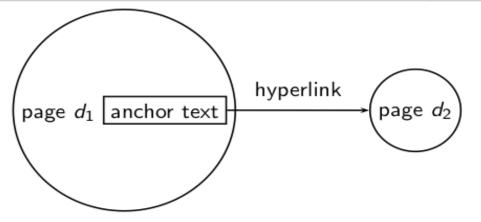
- Calender page with a dynamic link to the next month
- May be created intentionally or unintentionally

## LINK ANALYSIS (ANCHOR TEXT)

#### Sec. 21.1



#### Anchor Text



Example: "You can find cheap cars <a href="<u>http://...</u>">here</a>." Anchor text: "You can find cheap cars here."

- Assumption: The anchor text describes the content of d<sub>2</sub>.
  - Anchor text is loosely defined as the text **surrounding** the hyperlink.

### [text of $d_2$ ] only vs. [text of $d_2$ ] + [anchor text $\rightarrow d_2$ ]

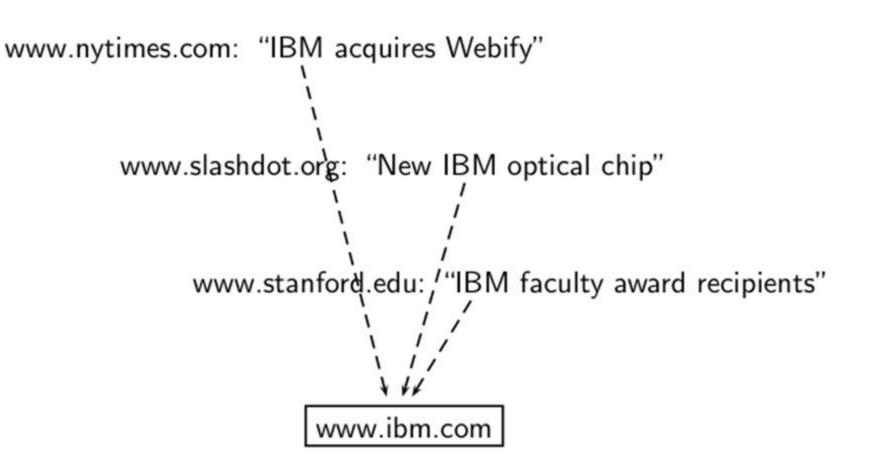


- Searching on [text of  $d_2$ ] + [anchor text  $\rightarrow d_2$ ] is often more effective than searching on [text of  $d_2$ ] only.
- Example: Query *IBM* 
  - IBM Wikipedia article (398 mentions)
  - Many spam pages that include hundreds of mentions of IBM intentionally
  - IBM home page (only 22 mentions with many graphics / video)
- Searching on [anchor text  $\rightarrow d_2$ ] is better for the query IBM
  - In this representation, the page with most occurrences of IBM is www.ibm.com.

CS3245 – Information Retrieval

# Anchor text containing *IBM* pointing to <u>www.ibm.com</u>





#### Indexing anchor text



- Thus: Anchor text is often a better description of a page's content than the page itself.
- Anchor text can be weighted more highly than document text.

# Google bombs

- Is a search with "bad" results due to maliciously manipulated anchor text.
- E.g., [dangerous cult] on Google, Bing, Yahoo
  - Coordinated link creation by those who dislike the Church of Scientology
- Google introduced a new weighting function in January 2007 that fixed many Google bombs.
- Defused Google bombs: [who is a failure?], [evil empire]



Web	Images Maps News Shopping Gmail more  BloomSEO OOgle dangerous cult Search			
Web	Results 1 - 10 of about 252,000 for dangerous cult. (0.06 se			
Scientology - Church of Scientology Official Site Living in a Dangerous Environment · Drug and Alcohol Problems · Personalities, Emot and How to Deal with Others www.scientology.org/ - 73k · <u>Cached</u> - <u>Similar pages</u> - <u>Note this</u> <u>The Most Dangerous Cult in The World by Laura Knight-Jadczyk</u> There's a new religious cult in America. It's not composed of so-called "crazies" so mu mainstream, middle to upper-middle class Americans www.cassiopaea.org/cass/Laura-Knight-Jadczyk/fastest_growing_cult.htm - 144k - <u>Cached</u> - <u>Similar pages</u> - <u>Note this</u> <u>Dangerous Cult Warning Signs</u> If you, or a loved one, are in a dangerous cult, as determined by the above checklist, must do everything you possibly can to remove the potential www.vistech.net/users/rsturge/cults.html - 4k - <u>Cached</u> - <u>Similar pages</u> - <u>Note this</u> <u>The Watchman Expositor: The Most Dangerous Cult in America</u> However, when the world's final chapter is written, which will prove to be "THE most dangerous cult in America?" One of the cults mentioned above? www.watchman.org/reltop/budcomp.htm - 10k - <u>Cached</u> - <u>Similar pages</u> - <u>Note this</u>				

#### Sec. 21.2

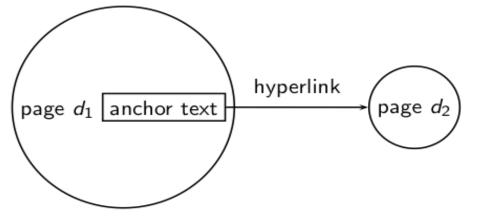


# LINK ANALYSIS (PAGERANK)

#### Sec. 21.1



# Quality Signal



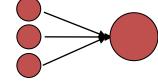
Example: "You can find cheap cars <a href="<u>http://...</u>">here</a>." Anchor text: "You can find cheap cars here."

- Assumption: A hyperlink is a quality signal.
  - The hyperlink d<sub>1</sub>→d<sub>2</sub> indicates that d<sub>1</sub>'s author deems d<sub>2</sub> high-quality and relevant.

### Quality Signal

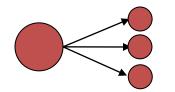


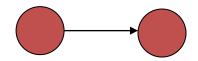
It is good to have more endorsements.



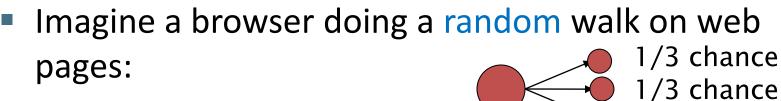
It is good for the endorsement to come from an important source.



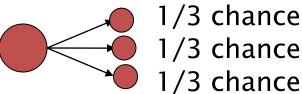




#### PageRank



Start at a random page

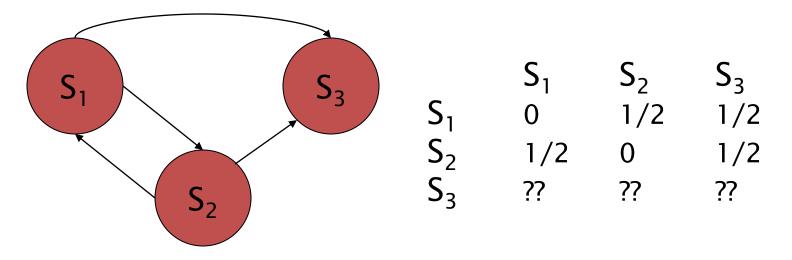


- At each step, follow one of the *n* links on that page, each with 1/n probability
- Do this repeatedly. Use the "long-term visit rate" as the page's score
  - This is a global score for the page, based on the topology of the network
  - Think of it as g(d) from Chapter 7

#### Random walks



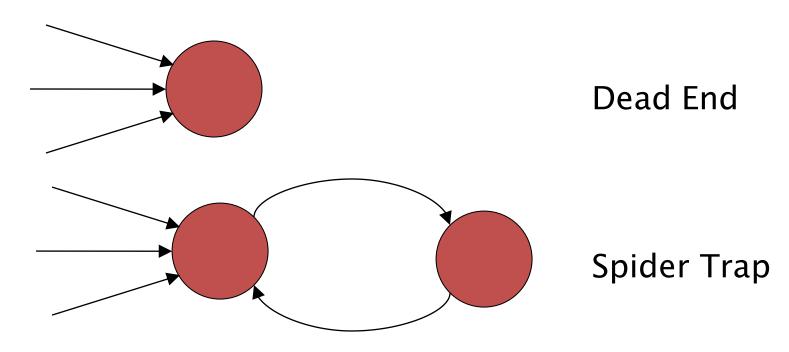
- A first order Markov chain consisting of n states and an nxn transition probability matrix A.
  - Each state correspond to a web page.
  - $A_{ik}$  is the probability of going from state *i* to state *k*.
  - The next state depends only on the current state.





#### Not quite enough

- The web is full of dead ends.
  - What sites have dead ends?
  - Our random walk can get stuck.



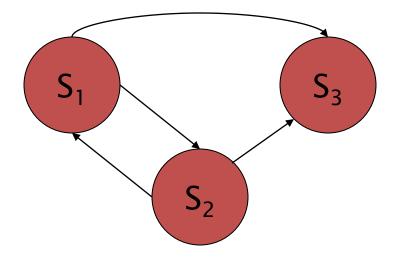


# Teleporting

- When a node has no outlinks
  - Teleport to a random web page
- Otherwise, at each step
  - With probability α (e.g., 10%), teleport to a random web page
  - With remaining probability (e.g., 90%), follow a random link on the page with equal probability



#### Random walks with teleportation



Matrix A with 10% chance of teleportation.

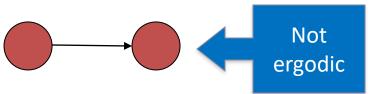
<b>S</b> <sub>1</sub>	S <sub>2</sub>	S <sub>3</sub>
1/30	29/60	29/60
29/60	1/30	29/60
1/3	1/3	1/3

 $S_1$  $S_2$  $S_3$ 



# Ergodic Markov chains

- A Markov chain is ergodic if
  - you have a path from any state to any other
  - you can be in any state at every time step, with non-zero probability



- With teleportation, our Markov chain is ergodic
- Theorem: With an ergodic Markov chain, there is a stable long term visit rate.

#### **Probability vectors**



- A probability (row) vector  $x = (x_1, ..., x_n)$  tells us where the surfer is at a particular point of time
- E.g.,  $(0_1 00 \dots 1_i \dots , 000_n) \rightarrow$  We're in state *i* with 100% probability
- More generally, the vector  $x = (x_1, ..., x_n)$  means the walk is in state *i* with probability  $x_i$ .

$$\sum_{i=1}^{n} x_i = 1$$



### Change in probability vector

- If the probability vector is x = (x<sub>1</sub>, ..., x<sub>n</sub>) at this step, what is it at the next step?
  - Recall at row *i* of the transition prob. Matrix A tells us where we go next from state *i*.
  - So from *x*, our next state is distributed as *xA*.

$$\vec{x_0} = (1\ 0\ 0)$$

$$\vec{x_1} = \vec{x_0}\mathbf{A} = (1/6\ 2/3\ 1/6\ ) \mathbf{A} = (1/3\ 1/3\ 1/3\ )$$

	<b>S</b> 1	<b>S</b> 2	S <sub>3</sub>
<b>S</b> 1	1/6	2/3	1/6
<b>S</b> 2	5/12	1/6	5/12
S <sub>3</sub>	1/6	2/3	1/6

#### **Steady State**



- For any ergodic Markov chain, there is a unique longterm visit rate for each state
  - Over a long period, we'll visit each state in proportion to this rate
  - It doesn't matter where we start

## PageRank algorithm



- Regardless of where we start, we eventually reach the steady state a
  - 1. Start with any distribution (say  $x = (1 0 \dots 0)$ ).
  - 2. After one step, we're at xA.
  - 3. After two steps at  $xA^2$ , then  $xA^3$  and so on.
  - 4. "Eventually" means for "large" k,  $xA^k = a$ .
- Algorithm: multiply x by increasing powers of A until the product looks stable.



#### At the steady state a...

- $a\mathbf{A} = a$
- So the rank vector is an eigenvector of the adjacency matrix
  - In fact, it's the first or principal eigenvector, with corresponding eigenvalue 1.

#### NUS National University of Singapore

## PageRank summary

- Pre-processing:
  - Given a graph of links, build matrix A
  - From it compute a
  - The page rank a<sub>i</sub> is a scaled number between 0 and 1.
- Query processing
  - Retrieve pages meeting query
  - Rank them by their PageRank
  - Order is query-independent



# How important is PageRank?

- Frequent claim: PageRank is the most important component of web ranking.
- The reality:
  - There are several components that are at least as important: e.g., anchor text, phrases, proximity, tiered indexes ...
  - PageRank in its original form (as presented here) has a negligible impact on ranking.
  - However, variants of a page's PageRank are still an essential part of ranking.
  - Addressing link spam is difficult and crucial.

## Summary



- Crawling Obtaining documents for indexing
  - Need to be polite
- PageRank A g(d) for asymmetrically linked documents
- Chapters 20 and 21 of IIR
- Resources
  - Paper on Mercator Crawler by Heydon et al.
  - Robot Exclusion Standard

PageRank reflects our view of the importance of web pages by considering more than 500 million variables and 2 billion terms. Pages that believe are important pages receive a higher PageRank and are more likely to appear at the top of the search results"





## **General Information**

- Date/Time: 30 Apr (Tue), 5-7pm
- Venue: COM3-MPH
- Format
  - Pen-and-paper
  - Open book (only printed / written materials allowed)
  - Calculators (without pre-stored programs) allowed





- All lectures, tutorials and course materials (including the corresponding sections in the textbook)
- For sample questions, refer to tutorials, and past year exam papers.
- If in doubt, ask on the forum



# Types of Questions

- Q1 is true/false questions on various topics (12 marks). No justifications required.
- Q2 is short questions on various topics (42 marks). No need to show work for calculations.
- Q3-6 are long questions, topic-specific (46 marks in total). Need to show work / give justifications as required.



## Help Session

29 Apr (Mon), 10am-12nn.

## Via Zoom

- Recorded on a best effort basis.
- Agenda
  - Exam Paper 22/23 AY Semester 2
  - Q&A

# WHERE TO GO **FROM HERE**

# Learning Objectives



- You have learnt how to build your own search engine!
- In addition, you have picked up skills that are useful for your future
  - Python one of the easiest and more straightforward programming languages to use.
  - NLTK A good set of routines and data that are useful in dealing with NLP and IR.



#### CS3245 – Information Retrieval



# Opportunities **Service** in IR



Jean Harris William *Trumba Beskobo Ymen & Pageflakes Vimeo
Blogniscient Tin TING & shutterlive and BodDater Feedster Blavoor Parasecont
ZAZZLE Tailrank @TagWorld nu//o @dogear &vokalke @ouper ODDPOST QOOP
iNeds Lutu R Bi blish flogr O settireAnt simply hired tech
gather Rezira exercise alegress Renkoo
YEDDA
Commemeorandum Balesdarius
Suprghi and
riya Wordcast Same and Streedster measure and Same Datasputse. 80000 arra
STREAMAD
nativetext CONCOO PODZINGER. RESEMAD Feed Tier shartere with submuber A
Yub.com
gobbreom Grast achat ach
Chatsum PANDORA LOOKIATEr.
Weblay Smars Noodly 30 windir digo DOX Jots Xmret
vizu digg Residence Grant AlmondRocks Tagyu 302 The Simpy Gtalkr
TRUVEO egoSurf comble pegasus SQUIDOO: picturecloud
newsvine Clipfire
LEXXGaipha
yelp28 Isnetuts Inform magnolia ChoZu Manufacture magnolia
MusicSearch S ClipShack MeetWithApprovation Bourdance SplD spy
Musicsearch enpointer



## **Keep Searching**

### ✓ See what was trending in 2023 - Singapore ≎

#### Singapore News

- 1 Ticketmaster
- 2 Tan Chuan Jin
- 3 Nicole Seah
- 4 CDC Vouchers 2023
- 5 iPhone 15

#### International News

- 1 War in Israel and Gaza
- 2 Abby Choi
- 3 Inter Miami
- 4 Turkey Earthquake
- 5 Al-Nassr

#### Places in Singapore

- 1 Woodleigh Mall
- 2 Sengkang Grand Mall
- 3 Woodlands Checkpoint
- 4 Cedric Grolet Singapore
- 5 Bird Paradise

#### Activities in Singapore

- 1 F1 Singapore 2023
- 2 Chatuchak Singapore
- 3 River Hongbao 2023
- 4 Halloween Horror Nights 2023
- 5 Van Gogh Singapore

#### Movies

- 1 Oppenheimer
- 2 Barbie
- 3 John Wick 4
- 4 Guardians of the Galaxy
- 5 Mission Impossible 7

#### TV Shows

- 1 The Glory
- 2 Physical 100
- 3 King the Land
- 4 Moving
- 5 Mask Girl

#### https://about.google/stories/year-in-search/

#### CS3245 – Information Retrieval















