THE UNIVERSITY
of EDINBURGH

Text Technologies for Data Science

INFR11145

Web Search (2)

Instructor:
Walid Magdy

14-Nov-2017

Lecture Objectives

- Learn about:
  - Basics of Web search
  - Brief History of web search
  - SEOs
  - Web Crawling (intro)

The Reality of Web Search

- **PageRank** is used in Google, but is hardly the full story of ranking
  - Many sophisticated features are used
  - Some address specific query classes
  - Machine-learned ranking heavily used
    - Learning to Rank (L2R)
    - Many features are used, including PR
  - Still counted as a very useful feature

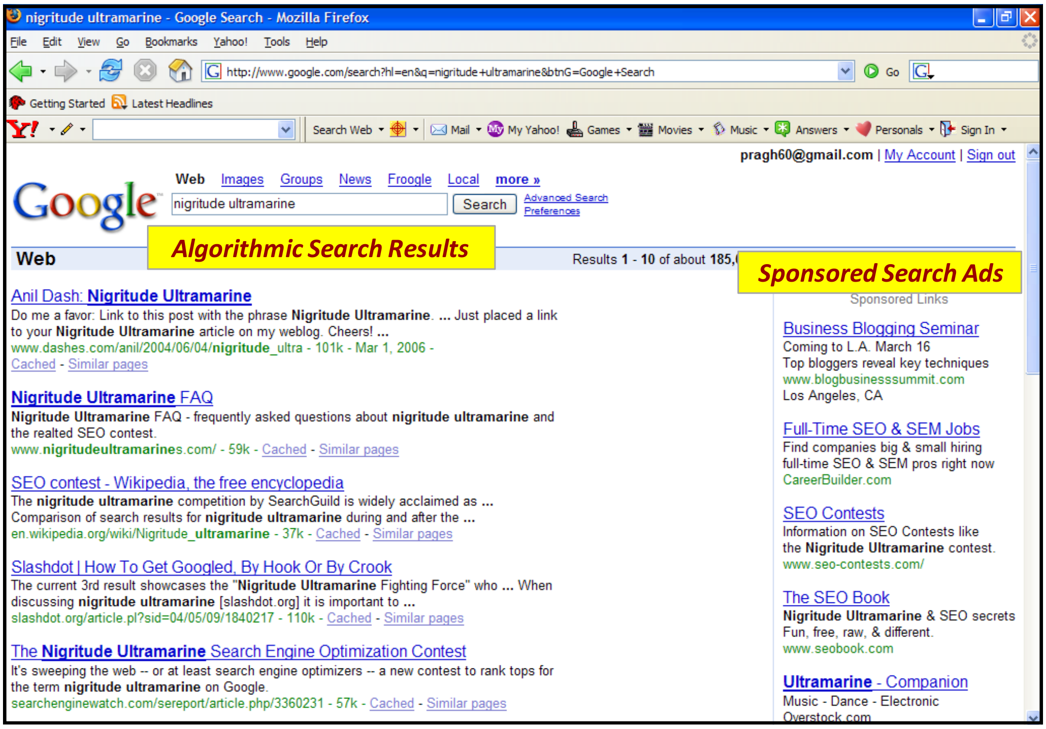
Brief History

- Early keyword-based engines (1995-1997)
  - Altavista, Excite, Infoseek, Lycos
  - Traditional IR techniques
  - Scalability is an issue

- Paid search ranking: Goto (morphed into Overture.com → Yahoo!)
  - Your search ranking depended on how much you paid
  - Auction for keywords
  - Called “sponsored search”
    - CPC (Cost Per Click)
    - CPM (Cost Per Thousand Impressions)

Brief (non-technical) History

- 1998+: Link-based ranking pioneered by Google
  - Blew away all early engines
  - Great user experience in search of a business model
  - Meanwhile Goto/Overture's annual revenues: ~ $1 billion
- Result: Google added paid search "ads" to the side, independent of search results
  - Yahoo followed, acquiring Overture (for paid placement) and Inktomi (for search)
- 2005+: Google gains search share, dominating in Europe and very strong in North America
  - 2009: Yahoo! and Microsoft combined paid search offering


nigritude ultramarine - Google Search - Mozilla Firefox
File Edit View Go Bookmarks Yahoo! Tools Help
http://www.google.com/search?hl=en&q=nigritude+ultramarine&btnG=Google+Search
Go
Getting Started
Latest Headlines
Y!
Search Web
Mail
My Yahoo!
Games
Movies
Music
Answers
Personals
Sign In
pragh60@gmail.com | My Account | Sign out
Google
Web Images Groups News Froogle Local more »
nigritude ultramarine
Search
Advanced Search
Preferences
Web
Results 1 - 10 of about 185,
Sponsored Links

Algorithmic Search Results

Sponsored Search Ads

Anil Dash: Nigritude Ultramarine
Do me a favor: Link to this post with the phrase **Nigritude Ultramarine**. ... Just placed a link to your **Nigritude Ultramarine** article on my weblog. Cheers! ...
www.dashes.com/anil/2004/06/04/**nigritude**_ultra - 101k - Mar 1, 2006 -
Cached - Similar pages

Nigritude Ultramarine FAQ
Nigritude Ultramarine FAQ - frequently asked questions about **nigritude ultramarine** and the realted SEO contest.
www.**nigritudeultramarines**.com/ - 59k - Cached - Similar pages

SEO contest - Wikipedia, the free encyclopedia
The **nigritude ultramarine** competition by SearchGuild is widely acclaimed as ...
Comparison of search results for **nigritude ultramarine** during and after the ...
en.wikipedia.org/wiki/Nigritude_**ultramarine** - 37k - Cached - Similar pages

Slashdot | How To Get Googled, By Hook Or By Crook
The current 3rd result showcases the "**Nigritude Ultramarine** Fighting Force" who ... When discussing **nigritude ultramarine** [slashdot.org] it is important to ...
slashdot.org/article.pl?sid=04/05/09/1840217 - 110k - Cached - Similar pages

The **Nigritude Ultramarine** Search Engine Optimization Contest
It's sweeping the web -- or at least search engine optimizers -- a new contest to rank tops for the term **nigritude ultramarine** on Google.
searchenginewatch.com/sereport/article.php/3360231 - 57k - Cached - Similar pages

Business Blogging Seminar
Coming to L.A. March 16
Top bloggers reveal key techniques
www.blogbusinesssummit.com
Los Angeles, CA

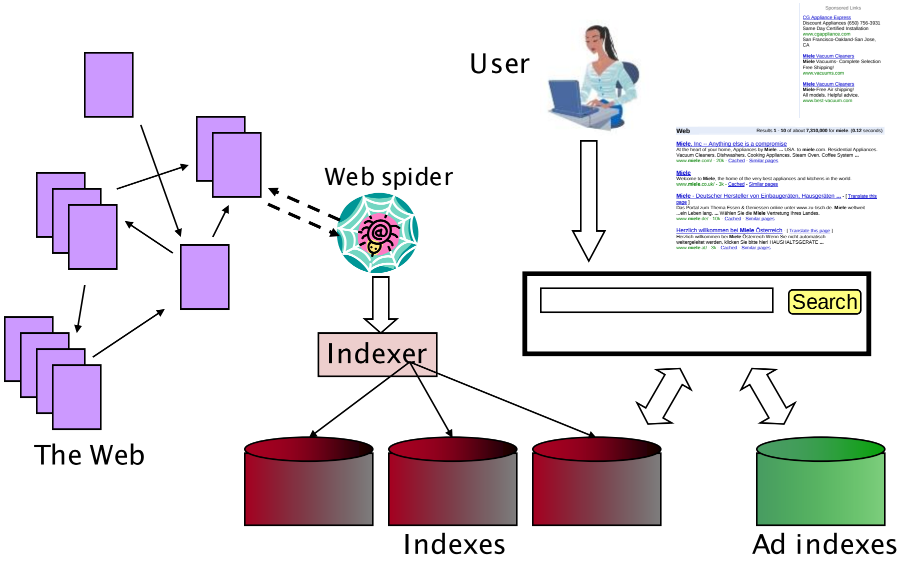
Full-Time SEO & SEM Jobs
Find companies big & small hiring
full-time SEO & SEM pros right now
CareerBuilder.com

SEO Contests
Information on SEO Contests like
the **Nigritude Ultramarine** contest.
www.seo-contests.com/

The SEO Book
Nigritude Ultramarine & SEO secrets
Fun, free, raw, & different.
www.seobook.com

Ultramarine - Companion
Music - Dance - Electronic
Overstock.com

Web Search Basics



Sponsored Links

CG Appliance Express
Discount Appliances (650) 756-3931
Same Day Certified Installation
www.cgappliance.com
San Francisco-Oakland-San Jose, CA

Miele Vacuum Cleaners
Miele Vacuums- Complete Selection
Free Shipping!
www.vacuums.com

Miele Vacuum Cleaners
Miele-Free Air shipping!
All models. Helpful advice.
www.best-vacuum.com

Web Results **1** - **10** of about **7,310,000** for **miele**. (**0.12** seconds)

Miele, Inc -- Anything else is a compromise
At the heart of your home, Appliances by **Miele**. **...** USA. to **miele**.com. Residential Appliances. Vacuum Cleaners. Dishwashers. Cooking Appliances. Steam Oven. Coffee System **...**
www.**miele**.com/ - 20k - Cached - Similar pages

Miele
Welcome to **Miele**, the home of the very best appliances and kitchens in the world.
www.**miele**.co.uk/ - 3k - Cached - Similar pages

Miele - Deutscher Hersteller von Einbaugeräten, Hausgeräten **...** - [Translate this page]
Das Portal zum Thema Essen & Geniessen online unter www.zu-tisch.de. **Miele** weltweit ...ein Leben lang. **...** Wählen Sie die **Miele** Vertretung Ihres Landes.
www.**miele**.de/ - 10k - Cached - Similar pages

Herzlich willkommen bei **Miele** Österreich - [Translate this page]
Herzlich willkommen bei **Miele** Österreich Wenn Sie nicht automatisch weitergeleitet werden, klicken Sie bitte hier! HAUSHALTSGERÄTE **...**
www.**miele**.at/ - 3k - Cached - Similar pages

User Need on Web Search

- **Informational** – want to learn about something (~40% / 65%)
  `Information Retrieval`
- **Navigational** – want to go to that page (~25% / 15%)
  `United Airlines`
- **Transactional** – want to do something (web-mediated) (~35% / 20%)
  - Access a service `Seattle weather`
  - Downloads `Mars surface images`
  - Shop `Canon S410`
- **Gray areas**
  - Exploratory search “see what’s there”

Search Engine Optimization (SEO)

- The Trouble with Paid Search Ads:
  It costs money. What's the alternative?
- ***Search Engine Optimization (SEO):***
  - "Tuning" your web page to rank highly in the algorithmic search results for selected keywords
  - Alternative to paying for placement
  - Thus, intrinsically a marketing function
- Performed by companies, webmasters and consultants ("Search engine optimizers") for their clients
- Some perfectly legitimate, some very shady

Walid Magdy, TTDS 2017/2018

SEO: Simplest Form

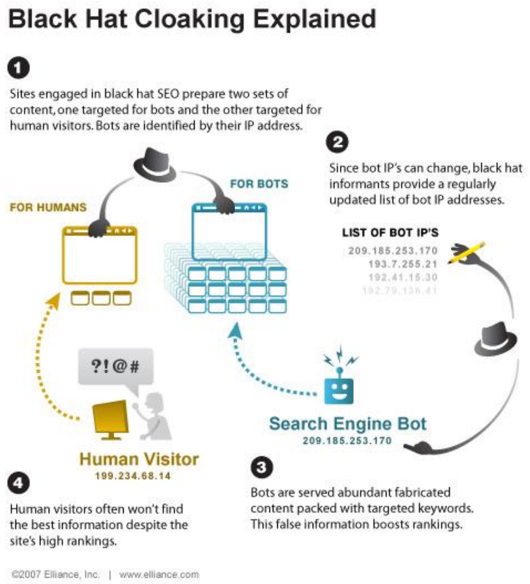
- First generation engines relied heavily on *tf/idf*
  - The top-ranked pages for the query **`maui resort`** were the ones containing the most **`maui`**'s and **`resort`**'s
- SEOs responded with dense repetitions of chosen terms
  - e.g., **`maui resort maui resort maui resort`**
  - Misleading meta-tags, excessive repetition
  - Often, the repetitions would be in the same color as the background of the web page
    - Repeated terms got indexed by crawlers
    - But not visible to humans on browsers

Pure word density cannot be trusted as an IR signal

Walid Magdy, TTDS 2017/2018

SEO: Cloaking

- Serve fake content to search engine spider
- Famous technique: **Black Hat**
- Kind of a spam!



Black Hat Cloaking Explained

1 Sites engaged in black hat SEO prepare two sets of content, one targeted for bots and the other targeted for human visitors. Bots are identified by their IP address.

2 Since bot IP's can change, black hat informants provide a regularly updated list of bot IP addresses.

3 Bots are served abundant fabricated content packed with targeted keywords. This false information boosts rankings.

4 Human visitors often won't find the best information despite the site's high rankings.

©2007 Elliance, Inc. | www.elliance.com

Duplicate Detection

- The web is full of duplicated content
- Strict duplicate detection = exact match
  - Not as common
  - can be detected with fingerprints
- But many, many cases of **near duplicates**
  - e.g., last modified date the only difference between two copies of a page
- *Near-Duplication*: Approximate match
  - Use similarity threshold to detect near-duplicates
    - e.g., Similarity > 80% => Documents are “near duplicates”
    - Not transitive though sometimes used transitively
      - A ≈ B & B ≈ C → doesn’t have to mean A ≈ C

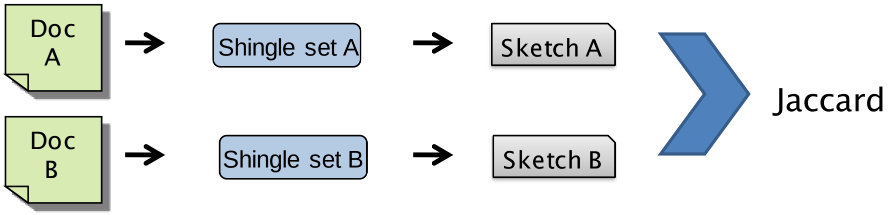
Duplicate Detection: MiniHash

- Features of similarity:
  - Segments of a document (natural or artificial breakpoints)
  - **Shingles** (word n-grams)
  - ***a rose is a rose is a rose*** →

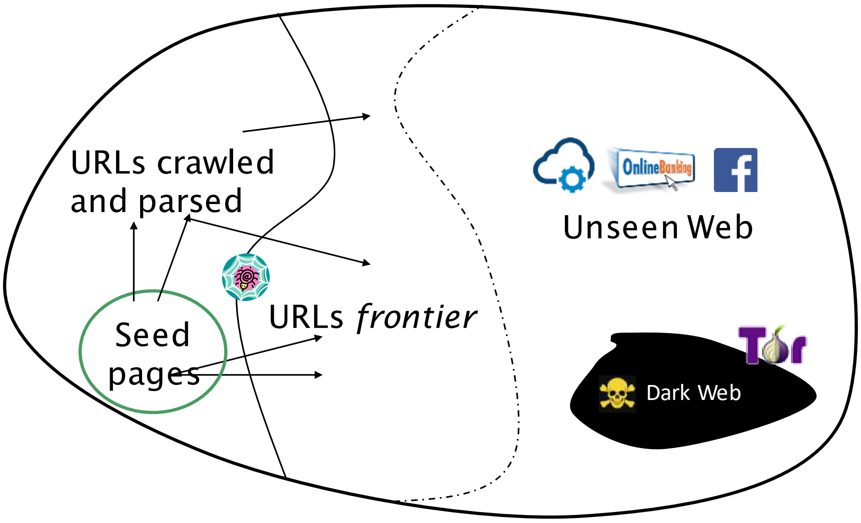
    a_rose_is_a
    rose_is_a_rose
    is_a_rose_is
    a_rose_is_a
- Similarity measure between two docs (= sets of shingles)
  - Set intersection
  - Specifically (Size_of_Intersection / Size_of_Union)

Shingles + Set Intersection

- Computing exact set intersection of shingles between all pairs of documents is expensive/intractable
- Approximate using a cleverly chosen subset of shingles from each (a sketch)
- Estimate $\frac{\textit{size of intersection}}{\textit{size of union}}$ based on a short sketch

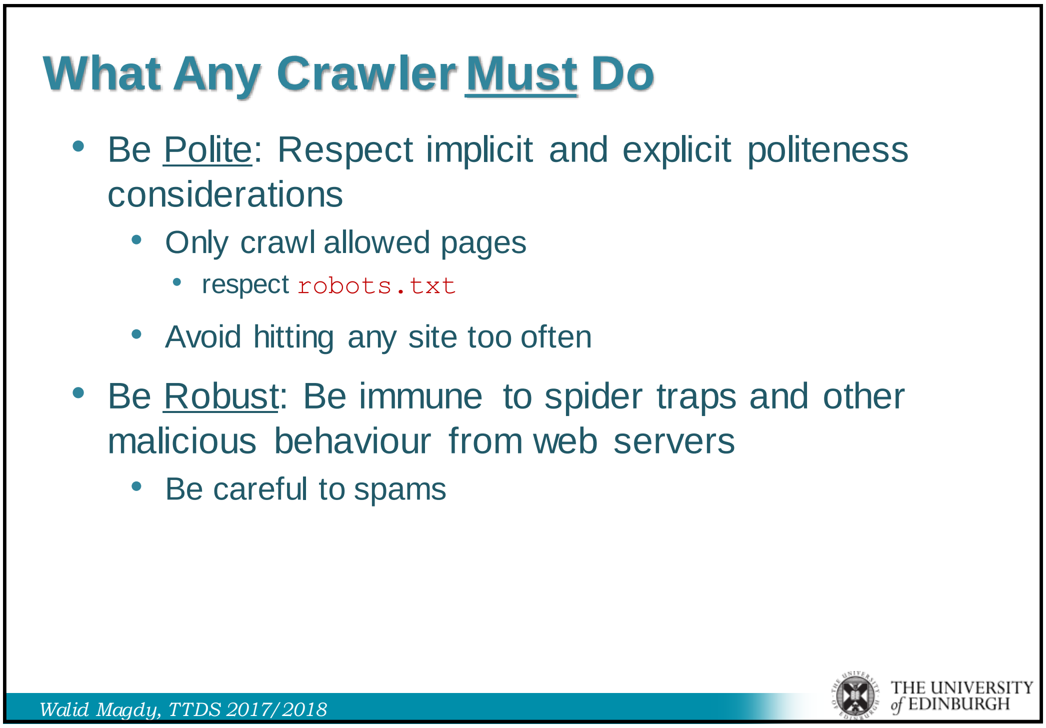


Web Crawling



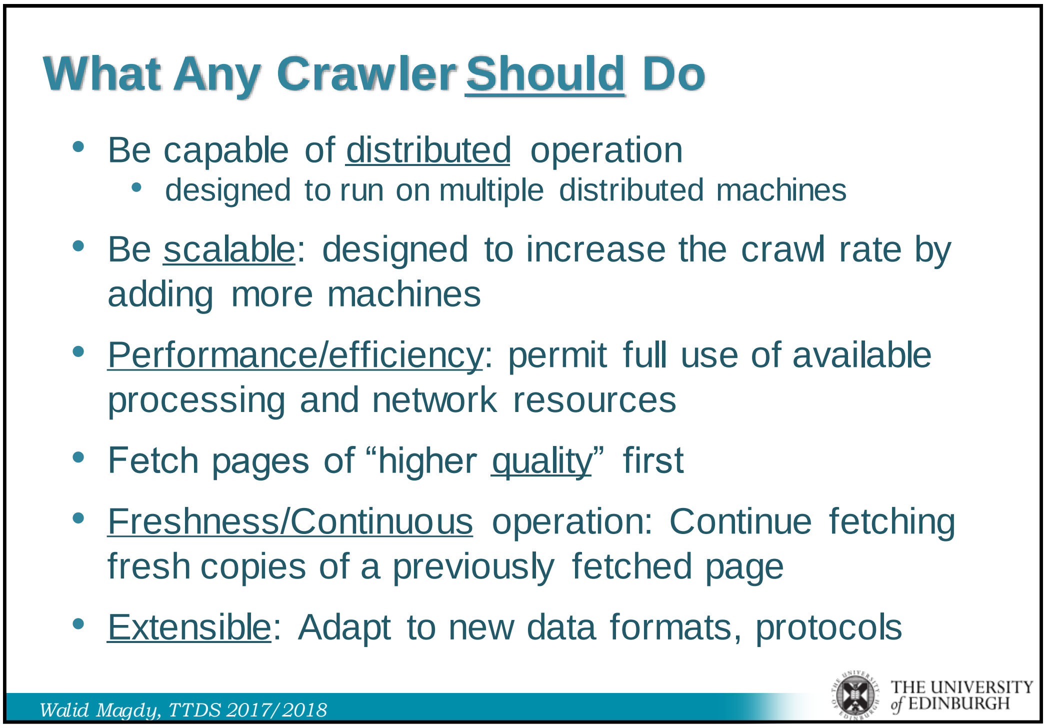
Basic Crawler Operation

- Begin with known “seed” URLs
- Fetch and parse them
  - Extract URLs they point to
  - Place the extracted URLs on a queue
- Fetch one URL from the queue
- Repeat



What Any Crawler Must Do

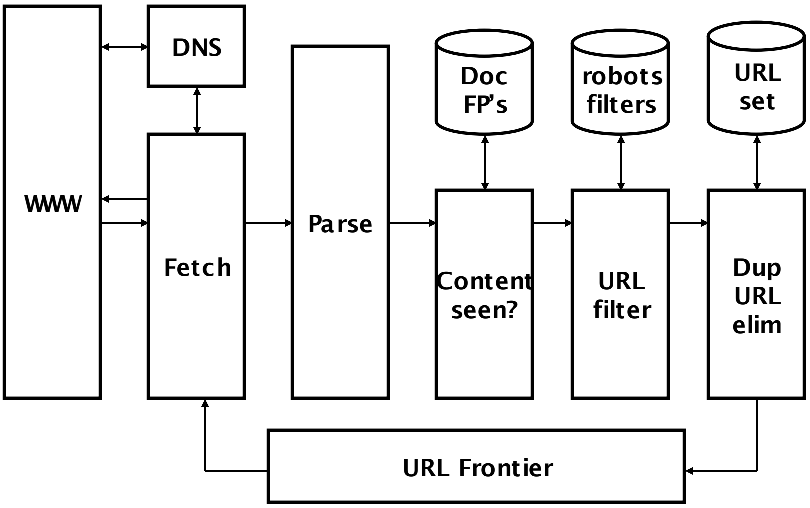
- Be Polite: Respect implicit and explicit politeness considerations
  - Only crawl allowed pages
    - respect `robots.txt`
  - Avoid hitting any site too often
- Be Robust: Be immune to spider traps and other malicious behaviour from web servers
  - Be careful to spams



What Any Crawler Should Do

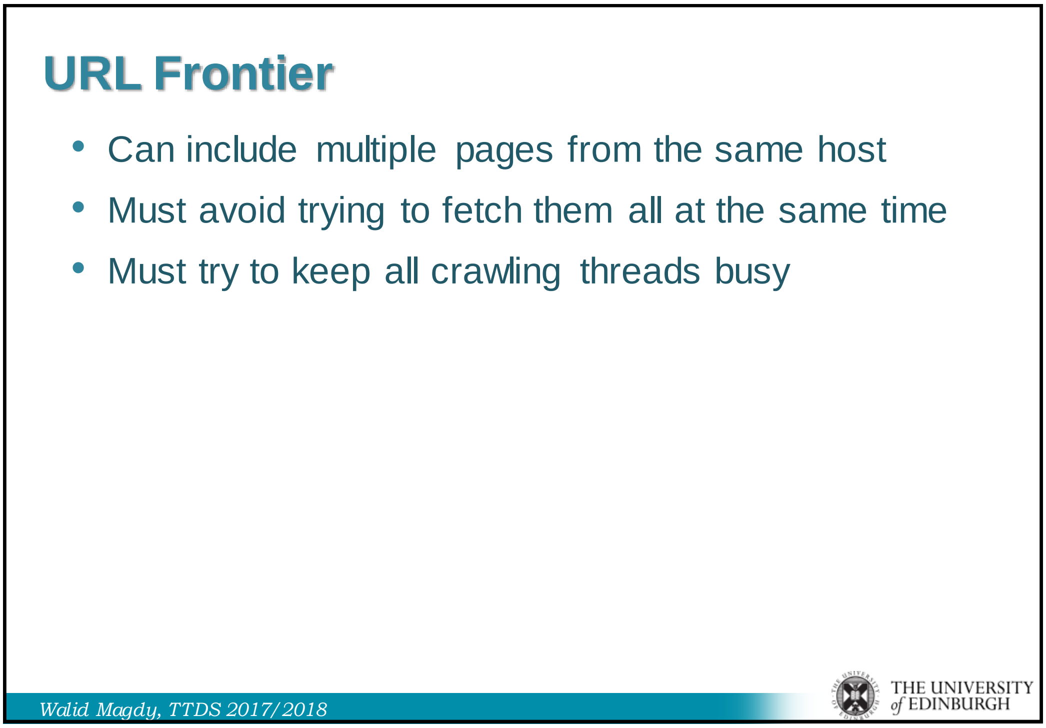
- Be capable of distributed operation
  - designed to run on multiple distributed machines
- Be scalable: designed to increase the crawl rate by adding more machines
- Performance/efficiency: permit full use of available processing and network resources
- Fetch pages of “higher quality” first
- Freshness/Continuous operation: Continue fetching fresh copies of a previously fetched page
- Extensible: Adapt to new data formats, protocols

Basic Crawler Architecture



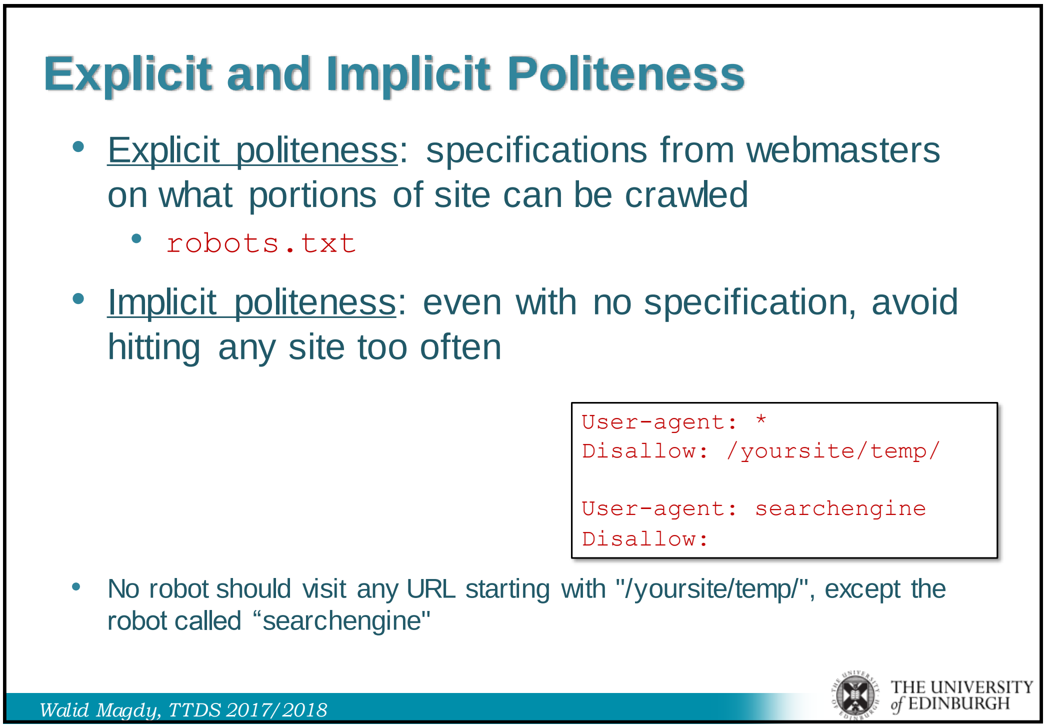
Processing Steps in Crawling

1. Pick a URL from the frontier
2. Fetch the document at the URL
3. Parse the document
    1. Extract links from it to other docs (URLs)
4. Check if document has content already seen
    1. If not, add to indexes
5. For each extracted URL
    1. Ensure it passes certain URL filter tests
    2. Check if it is already in the frontier (duplicate URL elimination)



URL Frontier

- Can include multiple pages from the same host
- Must avoid trying to fetch them all at the same time
- Must try to keep all crawling threads busy



Explicit and Implicit Politeness

- Explicit politeness: specifications from webmasters on what portions of site can be crawled
  - `robots.txt`
- Implicit politeness: even with no specification, avoid hitting any site too often

```
User-agent: *
Disallow: /yoursite/temp/

User-agent: searchengine
Disallow:
```

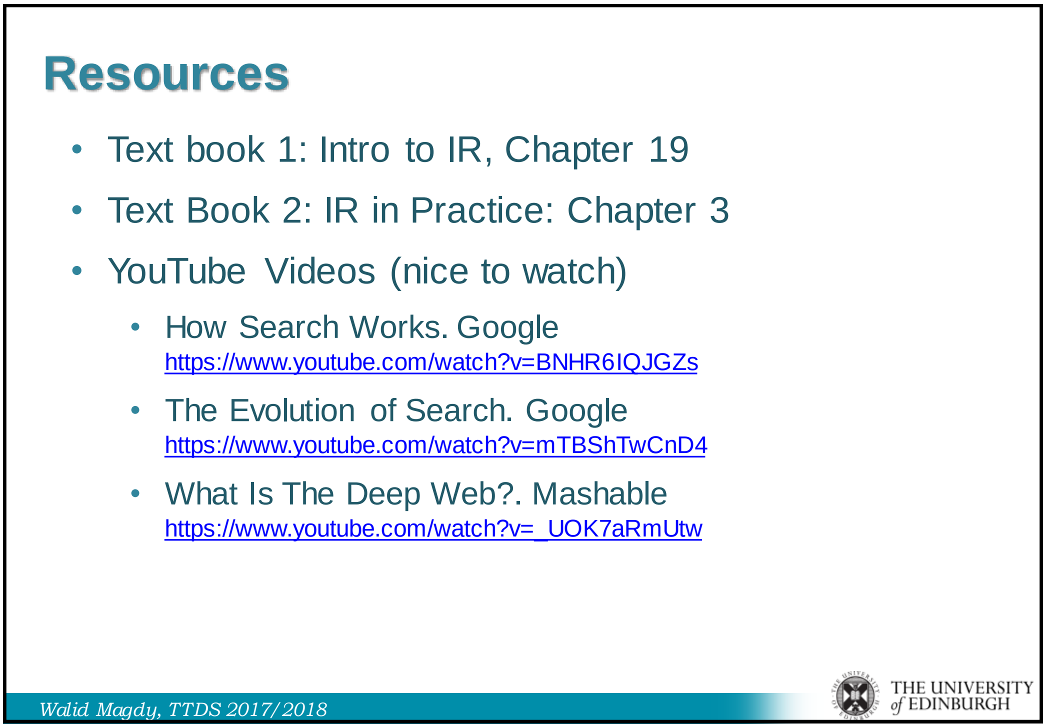
- No robot should visit any URL starting with "/yoursite/temp/", except the robot called “searchengine"

URL Frontier: 2 Main Considerations

- Politeness: do not hit a web server too frequently
- Priority/Freshness: crawl some pages more often than others
  - Pages whose content changes often (e.g. News sites)
- These goals may conflict each other.
  - e.g., simple priority queue fails – many links out of a page go to its own site, creating a burst of accesses to that site.
- Even if we restrict only one thread to fetch from a host, can hit it repeatedly
- Common heuristic: insert time gap between successive requests to a host that is >> time taken in most recent fetch from that host

Summary

- History of Web search
- Basics of web search
- Usage of web search
- SEO
- Web crawling



Resources

- Text book 1: Intro to IR, Chapter 19
- Text Book 2: IR in Practice: Chapter 3
- YouTube Videos (nice to watch)
  - How Search Works. Google
    https://www.youtube.com/watch?v=BNHR6IQJGZs
  - The Evolution of Search. Google
    https://www.youtube.com/watch?v=mTBShTwCnD4
  - What Is The Deep Web?. Mashable
    https://www.youtube.com/watch?v=_UOK7aRmUtw