



THE UNIVERSITY
of EDINBURGH

Text Technologies for Data Science

INFR11145

Web Search (2)

Instructor:
Walid Magdy

05-Nov-2025

1

Lecture Objectives

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



THE UNIVERSITY
of EDINBURGH

Walid Magdy, TTDS 2025/2026

2

Brief History

- Early keyword-based engines (1995-1997)
 - Altavista, Excite, Infoseek, Lycos, AOL
 - 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)

Walid Magdy, TTDS 2025/2026



3

CPC / CPM / RPM

- With new services on the web → RPM
- RPM: Revenue per 1000 video views
- Read more:
Understand ad revenue analytics
<https://support.google.com/youtube/answer/9314357>

Walid Magdy, TTDS 2025/2026



4

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
- 2024+: AI + RAG systems (*no clear winner yet!*)

Walid Magdy, TTDS 2025/2026



5

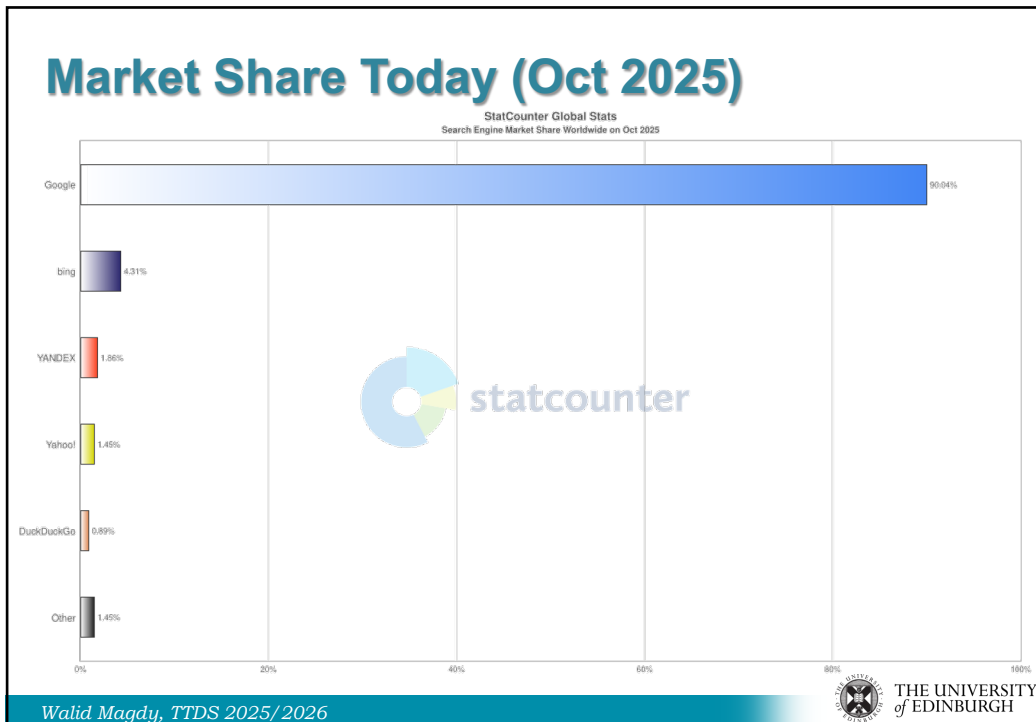
Brief (non-technical) History



Walid Magdy, TTDS 2025/2026



6



7

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

Getting Started Latest Headlines

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

Algorithmic Search Results

Results 1 - 10 of about 185,000

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 SearchGuld 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](#)

Sponsored Search Ads

Sponsored Links

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

8

How is it going?

Google

All News Images Videos Web Maps Shopping More Tools

Sponsored

sky.com
https://www.sky.com › superfast › broadband-deals

Sky Broadband Full Fibre
Super Reliable Fibre Broadband — Full Fibre Speeds For £25/m. Prices May Vary In 24m Contract. £0 Set-Up. Digital Exclusive. Super Reliable, Superfast Broadband - Keeping You Connected To The Things That Matter...

Sponsored

Hyperoptic
https://www.hyperoptic.com

Hyperoptic Broadband | Hyper Reliable Full Fibre
Uswitch Broadband Provider Of The Year 2023. As Voted For By Customers. Enjoy...

Sponsored

Uswitch
https://www.uswitch.com

Compare Broadband Deals
100s of Deals Available — Avoid Overpaying On Your Broadband Contract. Super Fast Fibre Broadband Deals Available

Sponsored

Virgin Media
https://www.virginmedia.com

Virgin Media Broadband Deals
Hurry, Offer Ends Soon — Stay Connected With Fast Speed: 132Mbps & Free Installation | Check Our Lowest WiFi Prices

Compare the Market
https://www.comparethemarket.com › broadband

Compare Broadband Deals – October 2024
Compare broadband deals in minutes with Compare the Market. Choose your speed and check what's available in your area. Search for cheap broadband deals.

Sponsored results are prioritized

Walid Magdy, TTDS 2025/2026

THE UNIVERSITY of EDINBURGH

9

How is it going?

Google

All Shopping Images Videos News Product sites Maps More Tools

Sponsored

DUSK Ascot Ottoman Stora... £319.00 DUSK +£19.99 delivery ★★★★★ (551) Bed · Double · Ottoman · Velvet By Crowd Sho...	Collect today IKEA NEIDEN Standard Doubl... £89.00 IKEA ★★★★★ (2k+) Bed · Double · Wood By Swoop	SALE Collect today IKEA MALM Standard Doubl... £195.00 £239 IKEA ★★★★★ (3k+) Bed · Double · Wood By Swoop	DUSK Ascot Ottoman Stora... £319.00 DUSK +£19.99 delivery ★★★★★ (551) Bed · Double · Ottoman · Linen By Crowd Sho...	Double Bed Frame Aurora... £66.99 Wayfair.co.uk +£4.99 delivery ★★★★★ (164) Bed · Double By Productcaster	BEDZONLINE Double... £99.95 Amazon.co.uk Free delivery Double · Mattress · Orthopaedic · ... By Kelkoo	Small Single Divan Bed with... £99.00 Beds.co.uk Free delivery Free returns Bed · Mattress · Headboard · ... By Feedoptimise	DUSK Ascot Ottoman Stora... £319.00 DUSK +£19.99 delivery ★★★★★ (551) Bed · Double · Ottoman By Crowd Sho...
---	--	---	--	---	--	--	--

Shopping posts (only sponsored)

Walid Magdy, TTDS 2025/2026

THE UNIVERSITY of EDINBURGH

10

How is it going?

Google

what is nigritude ultramarine

All Images Videos News Web Maps Books More Tools

Did you mean: what is **nigritude** ultramarine

AI Overview

The nigritude ultramarine competition was an SEO contest created by DarkBlue.com and run by SearchGuild. It is considered the first major SEO contest in the English-speaking world. The contest began on May 7, 2004, and Anil Dash won it two months later. The challenge was for webmasters to rank number one on Google for the search phrase "seraphim proudleduck" within three months.

SEO contest - Wikipedia

The first recorded SEO contest was Schnitzelkartoffelsalat by German webmasters, started on November 15, 2002, in th...

Wikipedia

Show more

In the English-language world, the nigritude ultramarine competition created by DarkBlue.com and run by SearchGuild is widely acclaimed as the mother of all SEO contests.

Wikipedia
https://en.wikipedia.org/wiki/SEO_contest

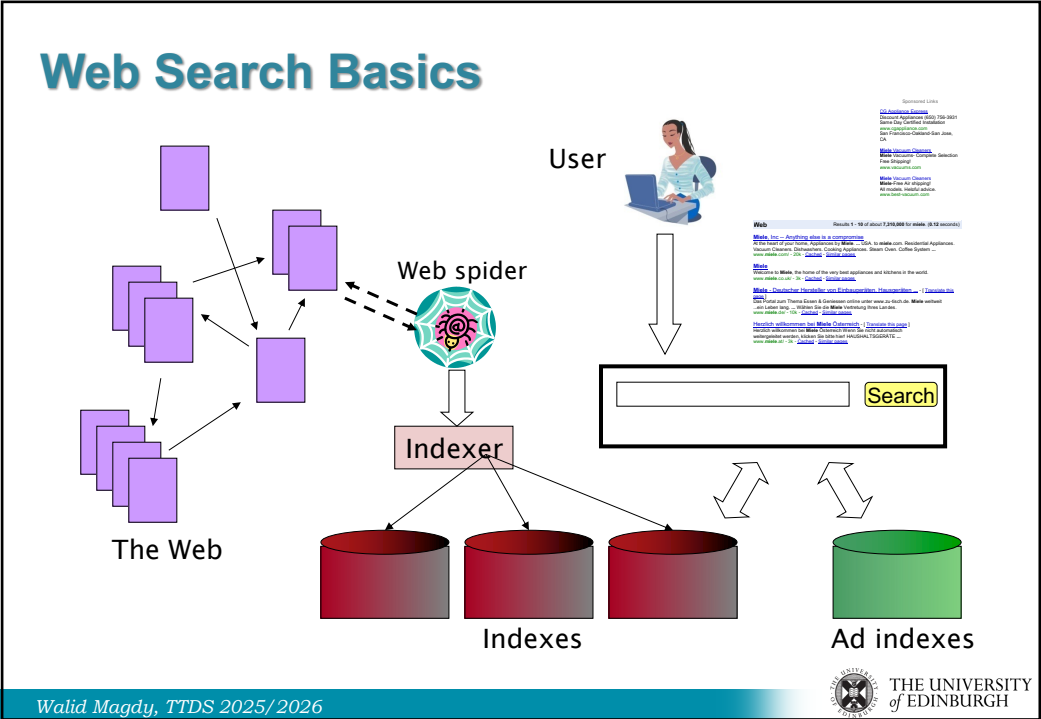
SEO contest - Wikipedia

About featured snippets Feedback

THE UNIVERSITY of EDINBURGH

Walid Magdy, TTDS 2025/2026

11



12

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”

Walid Magdy, TTDS 2025/2026



13

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
- Performed by companies, webmasters and consultants (“Search engine optimizers”) for their clients
- Some perfectly legitimate, some very shady

Walid Magdy, TTDS 2025/2026

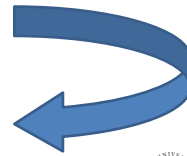


14

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 2025/2026



THE UNIVERSITY
of EDINBURGH

15

SEO word manipulating examples

- XYZ Hotel in ABC city
 - Accommodation, hotel, room, flat, travel, sights, attractions, vacation, holiday, in ABC ABC ABC
- XYZ for family advices
 - Family, couples, parents, spouse, wife, husband, fights, relationship, cheating, communication, kids, children
- XYZ Umbrellas
 - Raining, rainy, wet, weather, day

Walid Magdy, TTDS 2025/2026



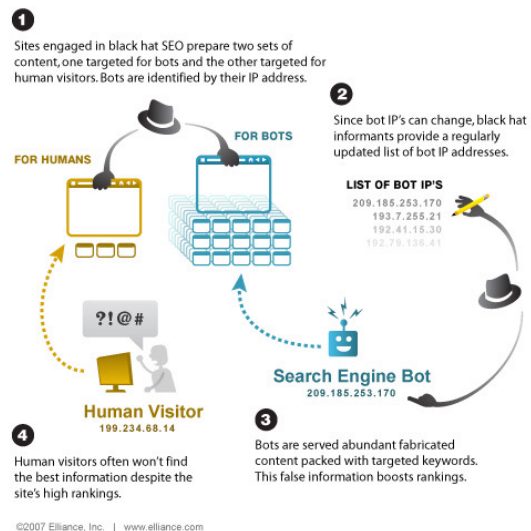
THE UNIVERSITY
of EDINBURGH

16

SEO: Cloaking

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

Black Hat Cloaking Explained



Walid Magdy, TTDS 2025/2026



17

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
 - $A \approx B$ & $B \approx C \rightarrow$ doesn't have to mean $A \approx C$

Walid Magdy, TTDS 2025/2026



18

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)

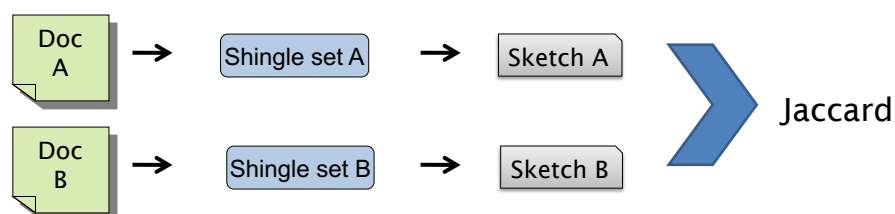
Walid Magdy, TTDS 2025/2026



19

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 $\frac{\text{size of intersection}}{\text{size of union}}$ based on a short sketch

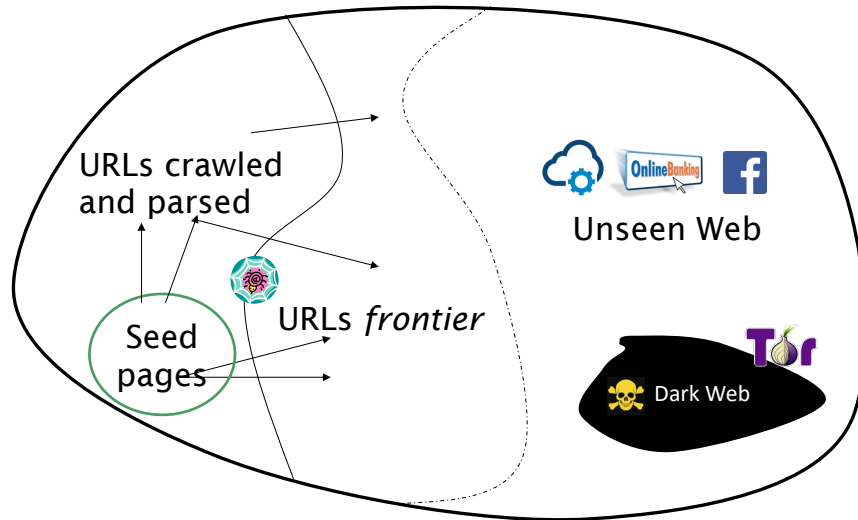


Walid Magdy, TTDS 2025/2026



20

Web Crawling



Walid Magdy, TTDS 2025/2026



21

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

Walid Magdy, TTDS 2025/2026



22

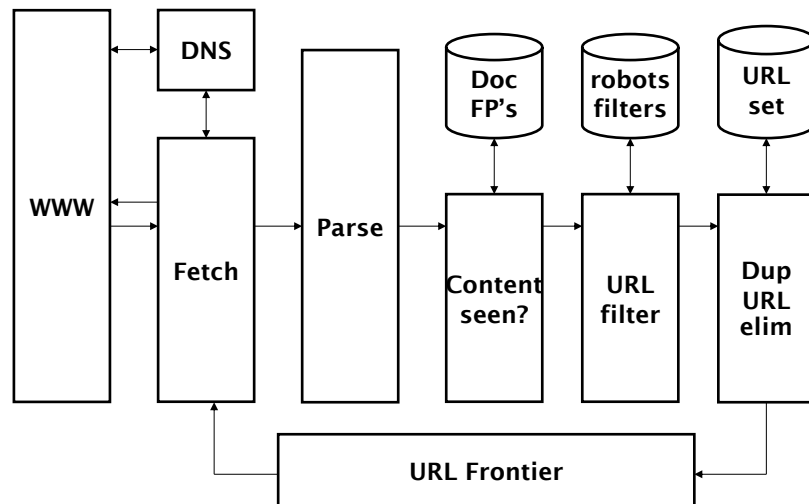
What Any Crawler Must Do

- 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 (link farms)

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



Walid Magdy, TTDS 2025/2026



25

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)

Walid Magdy, TTDS 2025/2026



26

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 $\gg$ time taken in most recent fetch from that host

Walid Magdy, TTDS 2025/2026



29

Summary

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

Walid Magdy, TTDS 2025/2026



30

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>
 - Most popular websites (search engines) over time
<https://www.youtube.com/watch?v=MirrGCbslp4>
 - This is How Much YouTube Pays Me
<https://www.youtube.com/watch?v=I3MeCEwVxB0>

Walid Magdy, TTDS 2025/2026



THE UNIVERSITY
of EDINBURGH