WEB INTELLIGENCE

Lecture 1: Introduction

Russa Biswas

September 02, 2025

AALBORG UNIVERSITY

Organisational Details

Lecturer: Russa Biswas

Schedule:

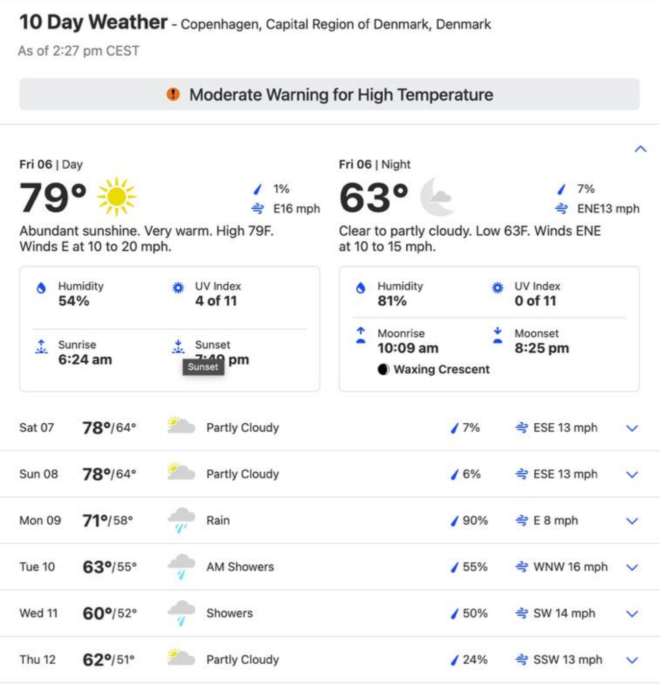
- Every Tuesday at 12:30 – 14:00 in Room 2.2.120
- Exercises: Every Monday at 14:15 – 16:15 in Room 2.2.120/ Group Room

Exercises:

Classroom Exercises: smaller problems solved on pen-paper (or white board)

Extended Exercises: Exercises similar to classroom exercises + Practical programming tasks with Python and Jupyter notebooks

Introduction



10 Day Weather - Copenhagen, Capital Region of Denmark, Denmark

As of 2:27 pm CEST

Moderate Warning for High Temperature

Fri 06 | Day

79° 1% E16 mph

Abundant sunshine. Very warm. High 79F. Winds E at 10 to 20 mph.

Humidity **54%**

UV Index **4 of 11**

Sunrise **6:24 am**

Sunset **7:40 pm**

Fri 06 | Night

63° 7% ENE13 mph

Clear to partly cloudy. Low 63F. Winds ENE at 10 to 15 mph.

Humidity **81%**

UV Index **0 of 11**

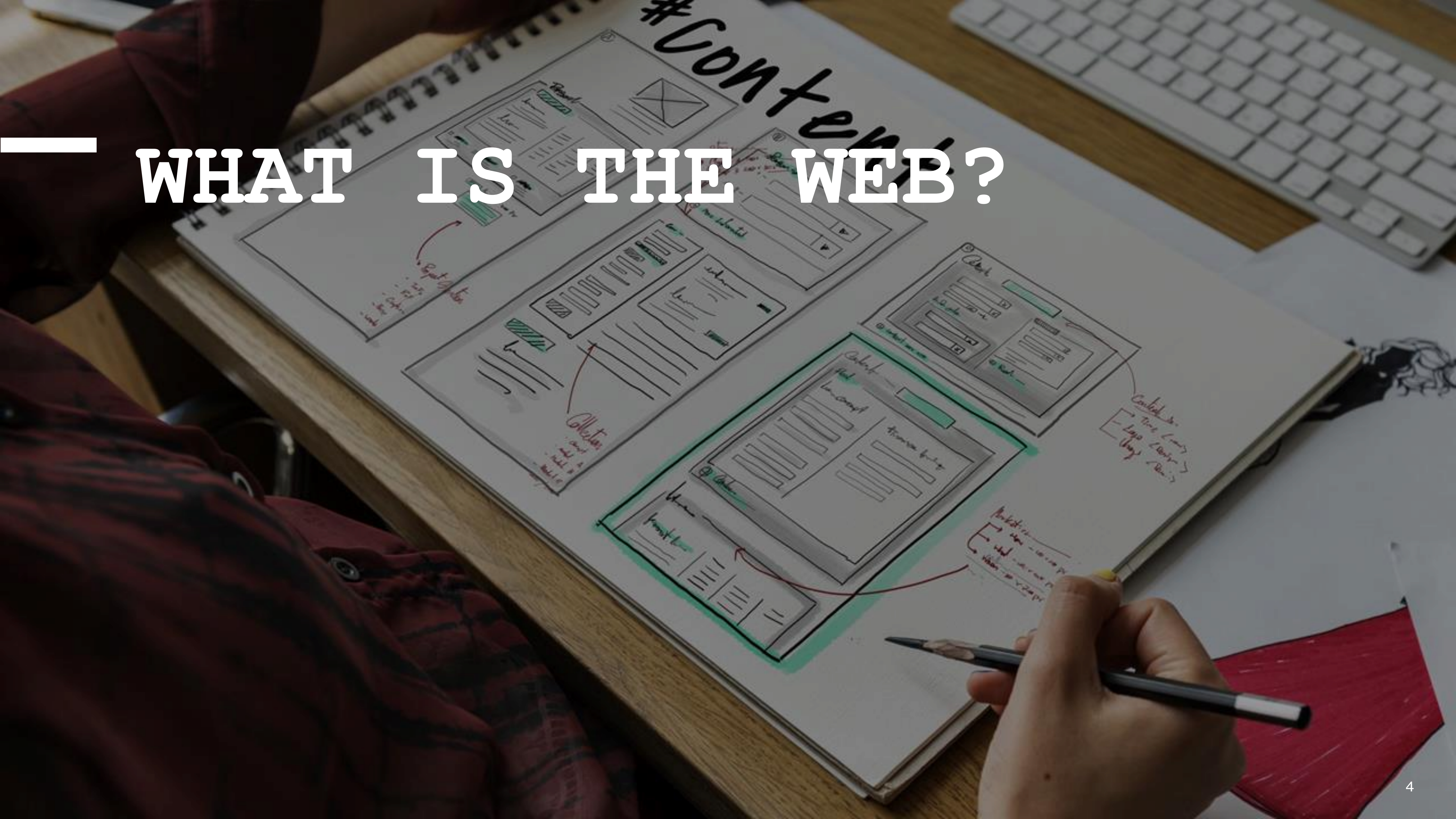
Moonrise **10:09 am**

Moonset **8:25 pm**

● **Waxing Crescent**

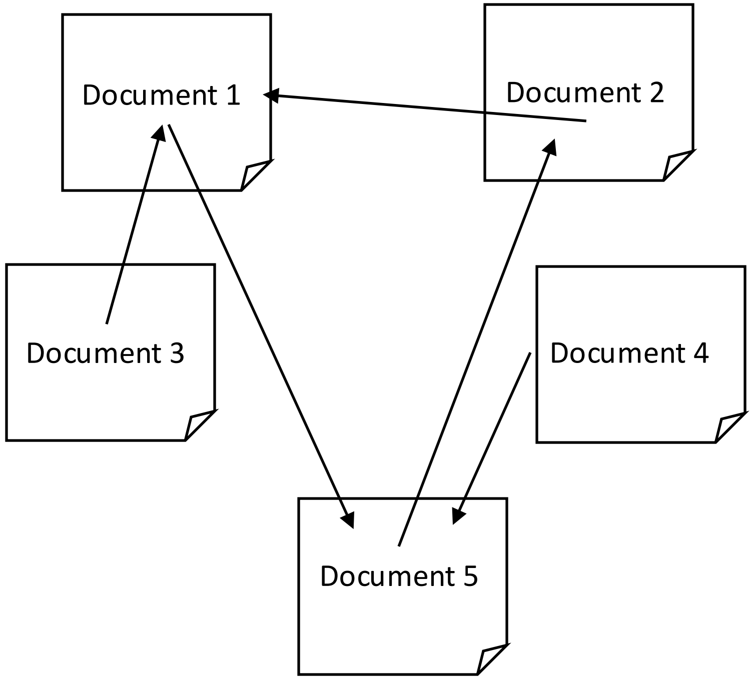
Sat 07	**78°**/64°	Partly Cloudy	7%	ESE 13 mph
Sun 08	**78°**/64°	Partly Cloudy	6%	ESE 13 mph
Mon 09	**71°**/58°	Rain	90%	E 8 mph
Tue 10	**63°**/55°	AM Showers	55%	WNW 16 mph
Wed 11	**60°**/52°	Showers	50%	SW 14 mph
Thu 12	**62°**/51°	Partly Cloudy	24%	SSW 13 mph

WHAT IS THE WEB?



What is the Web?

“…The dream behind the Web is of a common information space in which we communicate by sharing information… The web of human-readable document is being merged with a web of machine-understandable data…” – Tim Berners-Lee, https://w3.org/People/Berners-Lee/ShortHistory.html



- The Web is a
  - global **decentralized hypertext-based** information system.
  - **Hyperlinks** are used to navigate from one document to another.
  - This information space build on a set of technical standards for the identification, retrieval and representation of content.

World Wide Web (WWW)

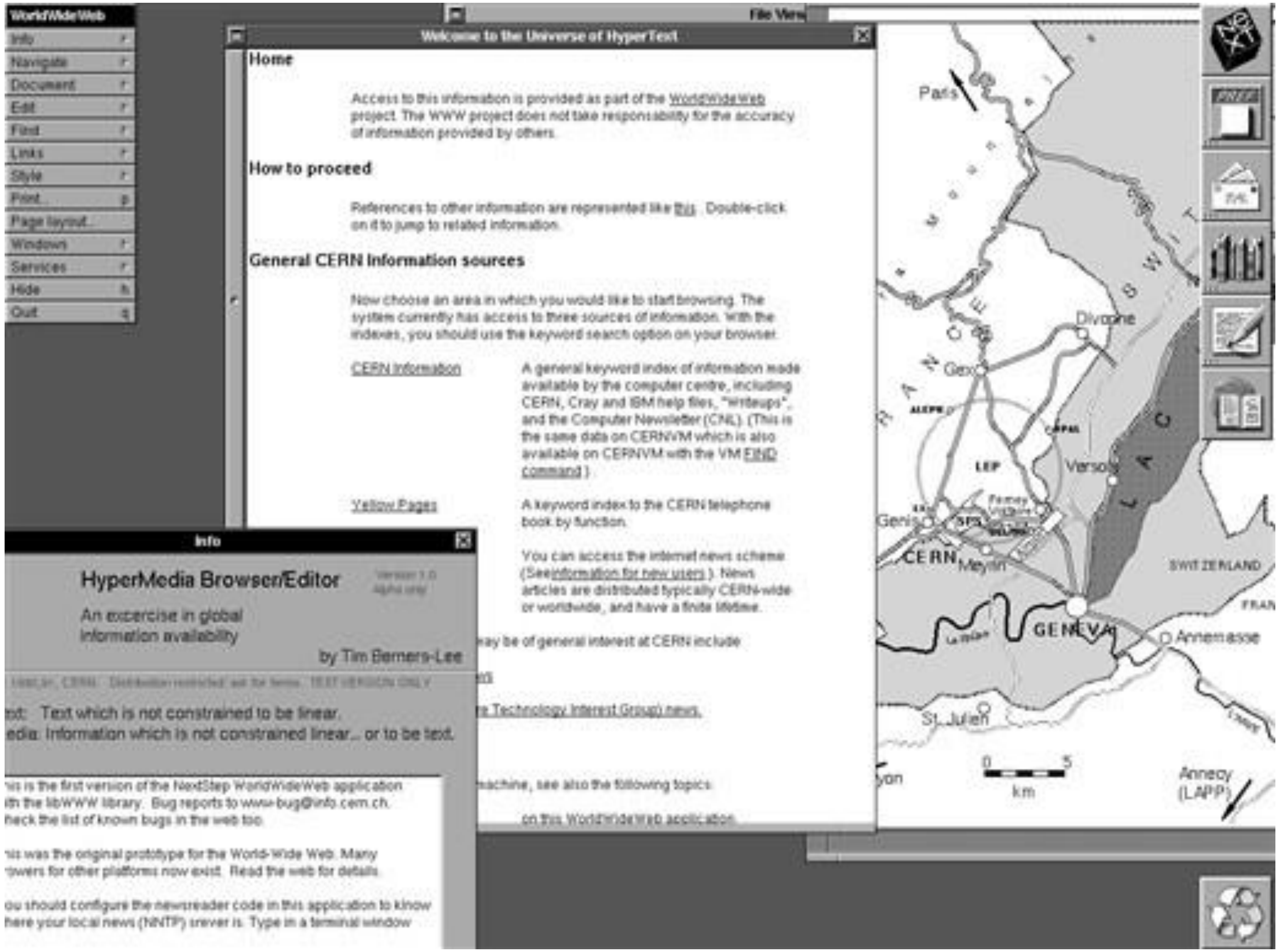
- The **World Wide Web** (referred to as WWW or W3 or simply Web), developed by **Tim Berners-Lee** in 1989 at CERN, Switzerland
  - The project document described a "**hypertext project**" called "**WorldWideWeb**" in which a "**web**" of "**hypertext documents**" could be viewed by “**browsers**”.[1]
- First Website in 1991: info.cern.ch
- First Webpage address: http://info.cern.ch/hypertext/WWW/TheProject.html

Tim Berners-Lee, pictured at CERN (Image: CERN)

[1] https://home.cern/science/computing/birth-web/short-history-web

World Wide Web (WWW)

- First Browser
- In 1993: Graphical Web Browsers such as Mosiac and Netscape Navigator were made accessible outside of academia.
- In 1994: W3C (World Wide Web Consortium) was founded by Tim Berners Lee. W3C publishes recommendations, that are considered web standards.
    - Web standards are blueprints –or building blocks– of a consistent and harmonious digitally connected world. They are implemented in browsers, blogs, search engines, and other software that power our experience on the web.



Home

Access to this information is provided as part of the WorldWideWeb project. The WWW project does not take responsability for the accuracy of information provided by others.

How to proceed

References to other information are represented like this . Double-click on it to jump to related information.

General CERN Information sources

Now choose an area in which you would like to start browsing. The system currently has access to three sources of information. With the indexes, you should use the keyword search option on your browser.

CERN Information — A general keyword index of information made available by the computer centre, including CERN, Cray and IBM help files, "writeups", and the Computer Newsletter (CNL). (This is the same data on CERNVM which is also available on CERNVM with the VM FIND command).

Yellow Pages — A keyword index to the CERN telephone book by function.

You can access the internet news scheme (See information for new users). News articles are distributed typically CERN-wide or worldwide, and have a finite lifetime.

ay be of general interest at CERN include

e Technology Interest Group) news.

machine, see also the following topics

on this WorldWideWeb application

Info

HyperMedia Browser/Editor

Version 1.0 Alpha only

An excercise in global information availability

by Tim Berners-Lee

CERN Distribution restricted: ask for terms. TEXT VERSION ONLY

xt: Text which is not constrained to be linear.

edia: Information which is not constrained linear... or to be text.

is is the first version of the NextStep WorldWideWeb application ith the libWWW library. Bug reports to www-bug@info.cern.ch. heck the list of known bugs in the web too.

is was the original prototype for the World-Wide Web. Many owers for other platforms now exist. Read the web for details.

ou should configure the newsreader code in this application to know here your local news (NNTP) srever is. Type in a terminal window

A screenshot showing the NeXT world wide web browser created by Tim Berners-Lee (Image: CERN)

World Wide Web – Data Model

W3 data model enables:

- Information need only be represented once, as a reference may be made instead of making a copy.
- Links allow the topology of the information to evolve, so modelling the state of human knowledge at any time is without constraint.
- The web stretches seamlessly from small personal notes on the local workstation to large databases on other continents.
- Indexes such as phone books are presented as documents, and so may themselves be found by searches and/or following links.
- The documents in the web do not have to exist as files; they can be "virtual" documents generated by a server in response to a query or document name. They can therefore represent views of databases, or snapshots of changing data (such as the weather forecasts, financial information, etc.).

Advantages:

- Information access doesn’t require expert knowledge
- Information Retrieval via search engines

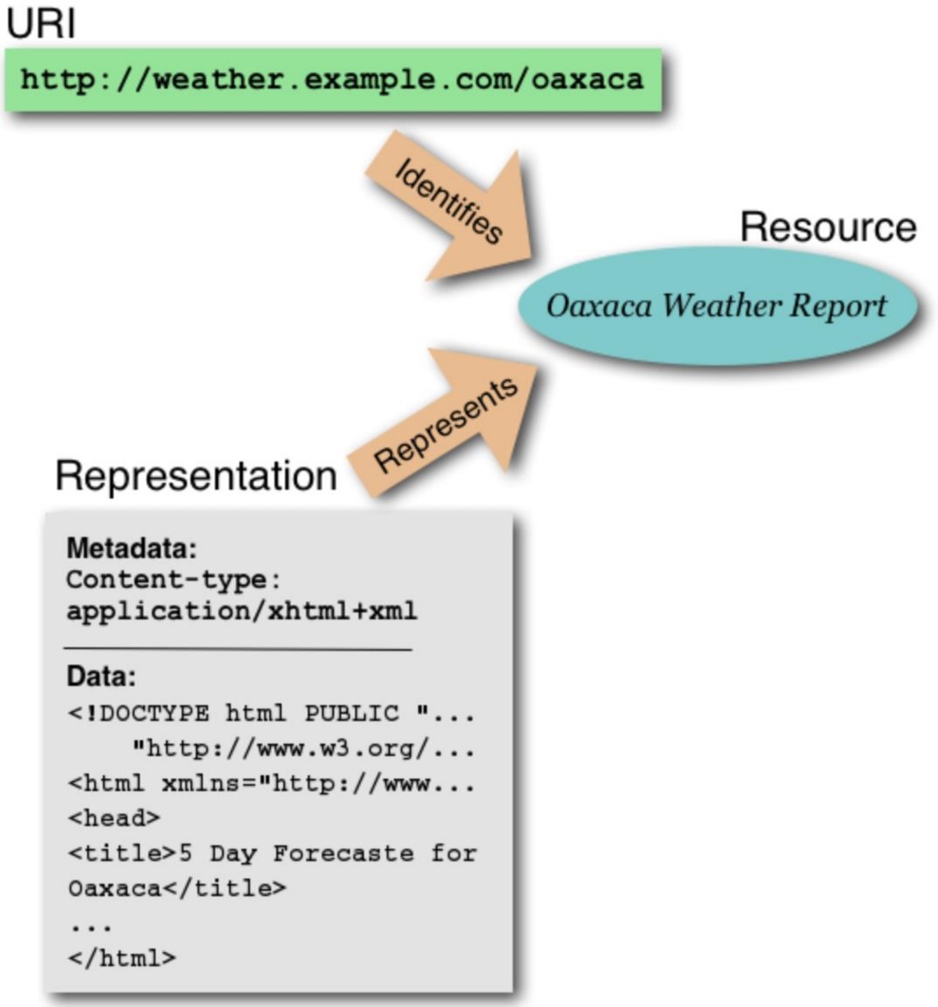
World Wide Web - Components

The World Wide Web (WWW) comprise of different components:

- **Identification:** Universal Resource Identifiers (URIs)- Address system; globally unique identification of the web resources.

For e.g., the URI of the main page for the first WWW project is http://info.cern.ch/hypertext/WWW/TheProject.html

- **Interaction:** Hypertext Transfer Protocol (HTTP) - network protocol used for transferring information/interacting between the web resources. The data transferred can be plain text, hypertext, images, etc.
- **Content Format:** Hypertext Markup Language (HTML) - a markup language, used to define the structure and the content of the webpage. HTML supports various content types, including text, images, video, audio, scripts, and hyperlinks for easy web resource access.



URI

`http://weather.example.com/oaxaca`

Identifies

Resource

Oaxaca Weather Report

Represents

Representation

Metadata:

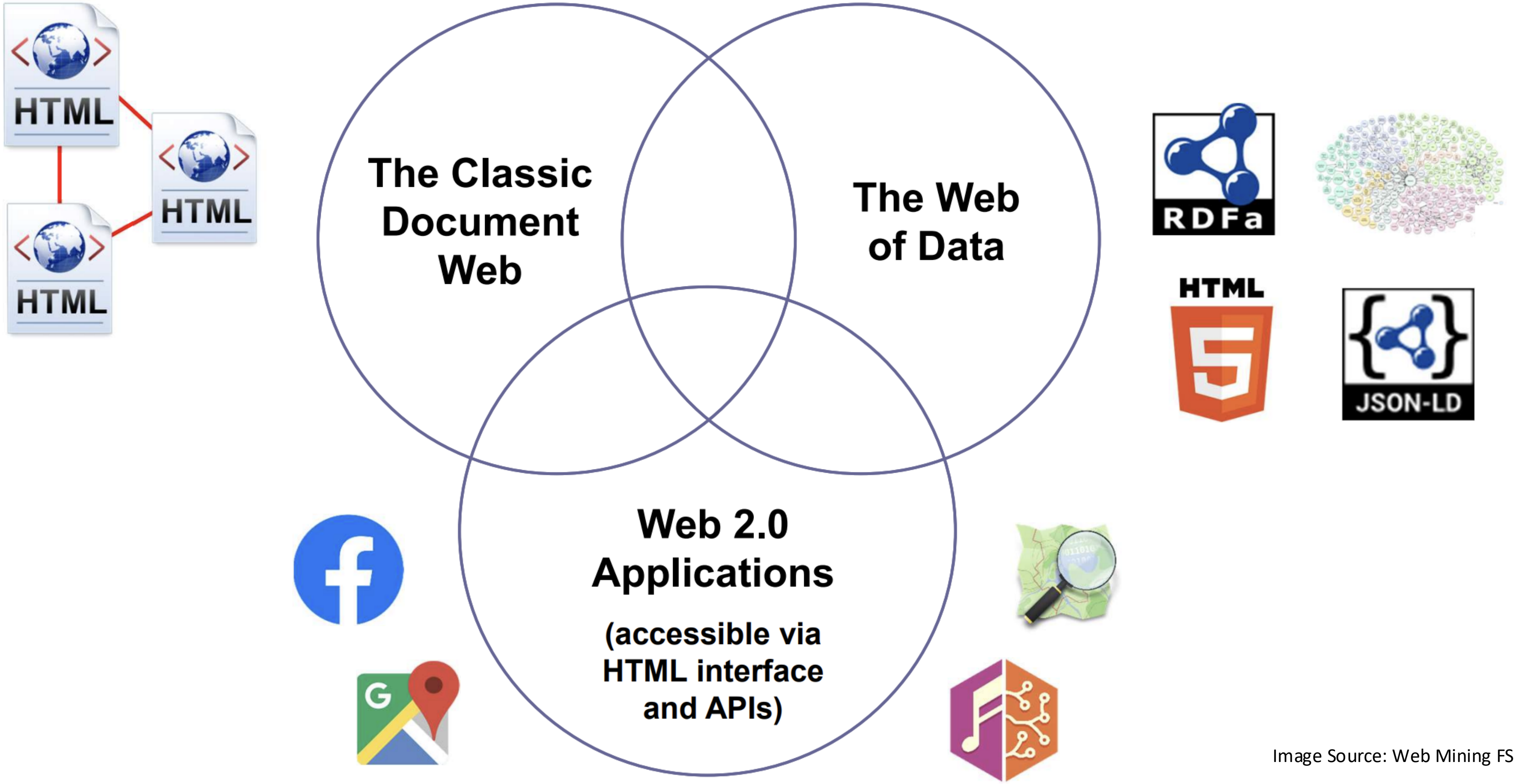
```
Content-type:
application/xhtml+xml
```

Data:

```
<!DOCTYPE html PUBLIC "...
    "http://www.w3.org/...
<html xmlns="http://www...
<head>
<title>5 Day Forecaste for
Oaxaca</title>
...
</html>
```

Image source: https://www.w3.org/TR/webarch/

Topology of the Web



The Classic Document Web

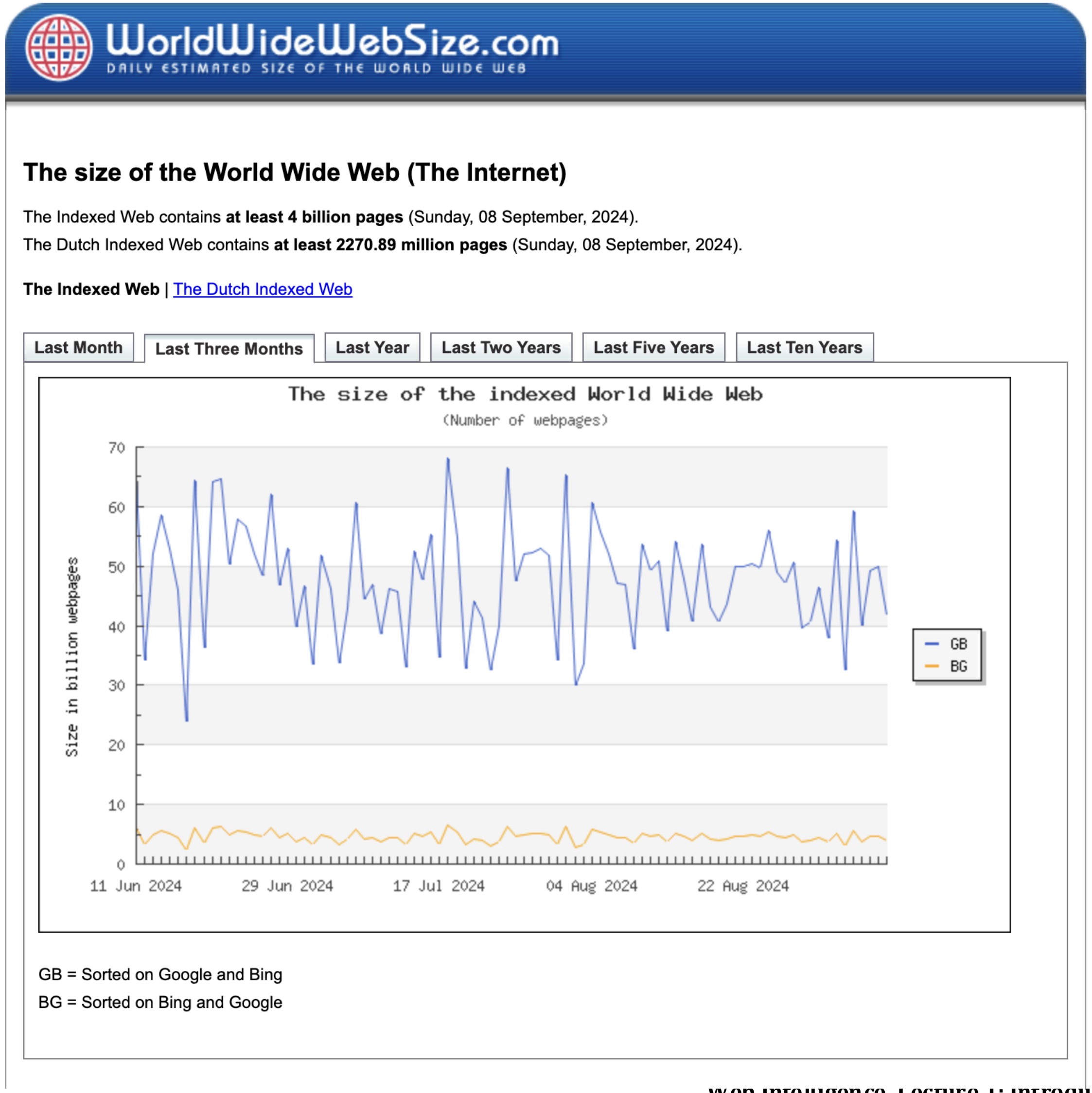
The Web of Data

Web 2.0 Applications

(accessible via HTML interface and APIs)

Image Source: Web Mining FSS2024, University of Mannheim

Size of the web



WorldWideWebSize.com

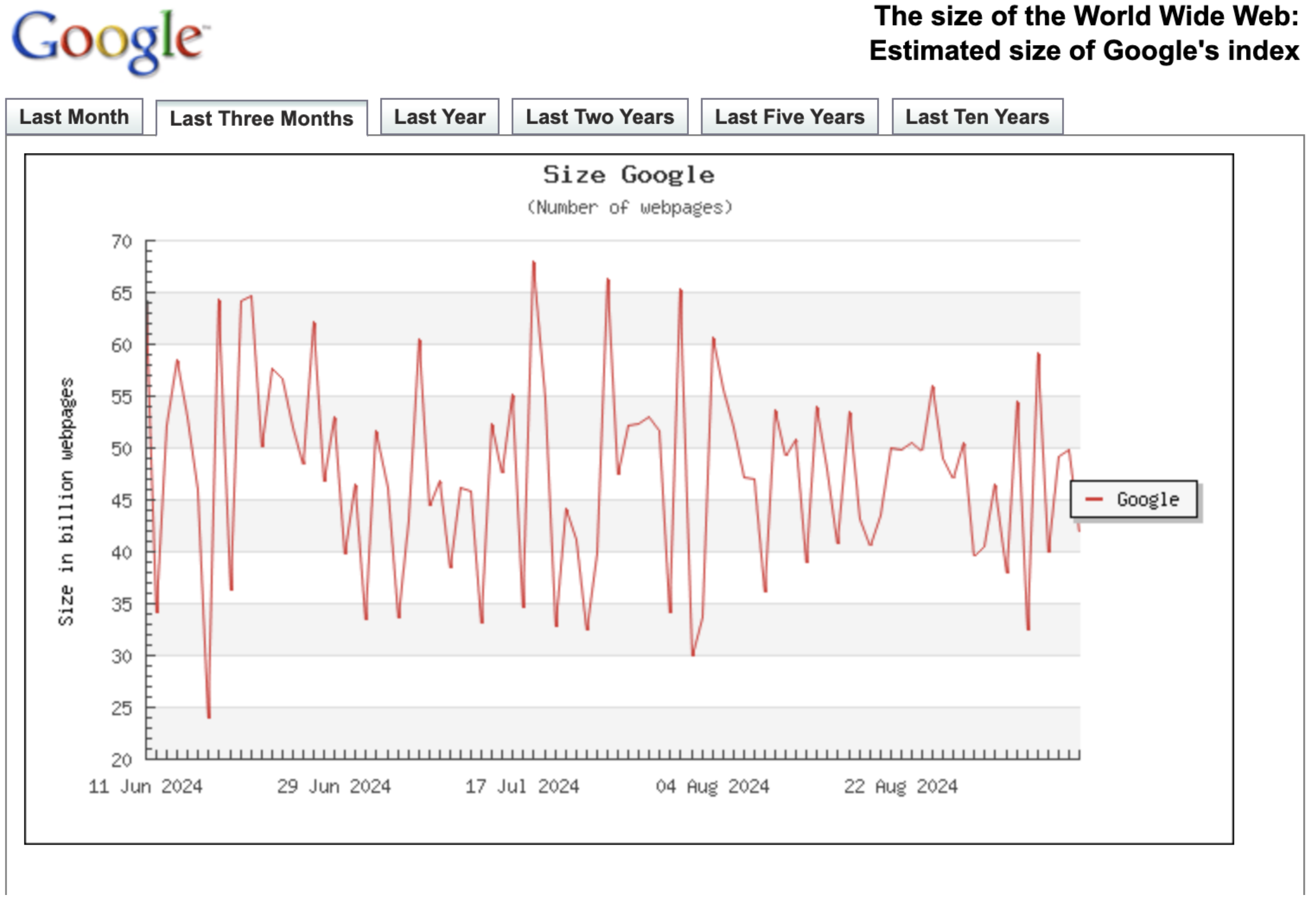
DAILY ESTIMATED SIZE OF THE WORLD WIDE WEB

The size of the World Wide Web (The Internet)

The Indexed Web contains **at least 4 billion pages** (Sunday, 08 September, 2024).
The Dutch Indexed Web contains **at least 2270.89 million pages** (Sunday, 08 September, 2024).

The Indexed Web | The Dutch Indexed Web

GB = Sorted on Google and Bing
BG = Sorted on Bing and Google



The size of the World Wide Web:
Estimated size of Google's index

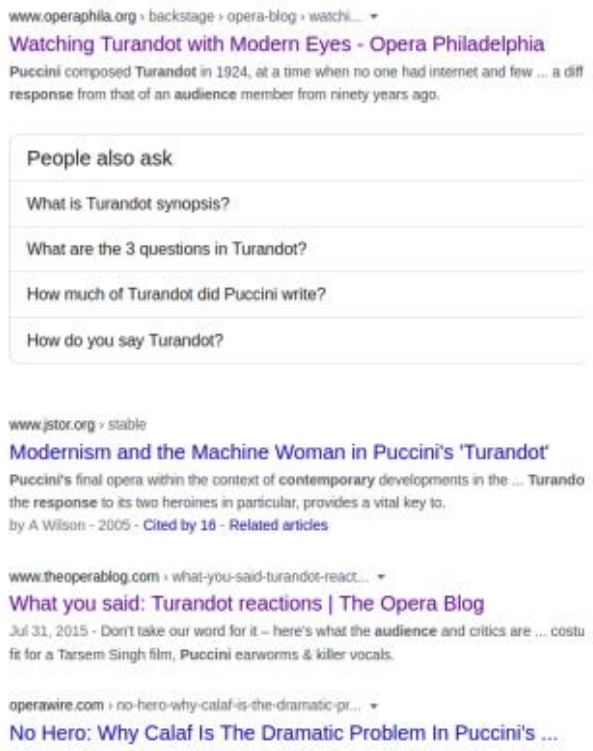
What is Web Intelligence?

- Intelligent ways to extract information and knowledge from the web:
  - finding relevant information available on the web
  - obtaining new knowledge by analyzing web data: the web itself, but also how it evolves, and how users interact on and with the web

- Some applications:
  - Intelligent Search
  - Recommender Systems
  - Business Analytics
  - Crowd Sourcing
  - Not so nice ones: advertising, manipulation, surveillance

Intelligent Search

Query: Contemporary reaction Puccini Turandot



www.operaphila.org › backstage › opera-blog › watchi... ▾

Watching Turandot with Modern Eyes - Opera Philadelphia

Puccini composed **Turandot** in 1924, at a time when no one had internet and few ... a diff **response** from that of an **audience** member from ninety years ago.

People also ask

- What is Turandot synopsis?
- What are the 3 questions in Turandot?
- How much of Turandot did Puccini write?
- How do you say Turandot?

www.jstor.org › stable

Modernism and the Machine Woman in Puccini's 'Turandot'

Puccini's final opera within the context of **contemporary** developments in the ... **Turando** the **response** to its two heroines in particular, provides a vital key to.

by A Wilson - 2005 - Cited by 16 - Related articles

www.theoperablog.com › what-you-said-turandot-react... ▾

What you said: Turandot reactions | The Opera Blog

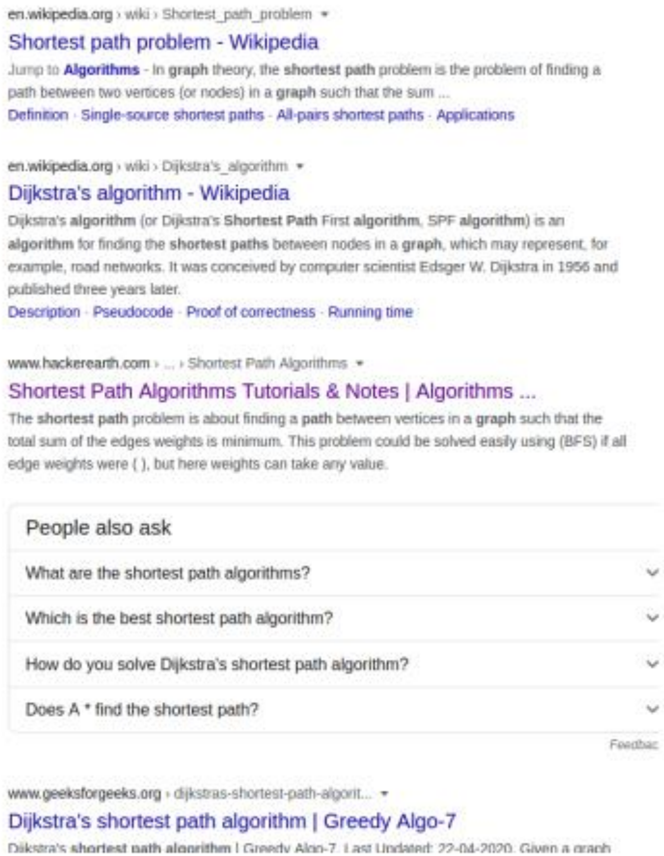
Jul 31, 2015 - Don't take our word for it – here's what the **audience** and critics are ... costu fit for a Tarsem Singh film, **Puccini** earworms & killer vocals.

operawire.com › no-hero-why-calaf-is-the-dramatic-pr... ▾

No Hero: Why Calaf Is The Dramatic Problem In Puccini's ...

➡Close, but not quite the information I was looking for

Query: shortest path algorithms



en.wikipedia.org › wiki › Shortest_path_problem ▾

Shortest path problem - Wikipedia

Jump to **Algorithms** - In **graph** theory, the **shortest path** problem is the problem of finding a path between two vertices (or nodes) in a **graph** such that the sum ...

Definition · Single-source shortest paths · All-pairs shortest paths · Applications

en.wikipedia.org › wiki › Dijkstra's_algorithm ▾

Dijkstra's algorithm - Wikipedia

Dijkstra's **algorithm** (or Dijkstra's **Shortest Path** First **algorithm**, SPF **algorithm**) is an **algorithm** for finding the **shortest paths** between nodes in a **graph**, which may represent, for example, road networks. It was conceived by computer scientist Edsger W. Dijkstra in 1956 and published three years later.

Description · Pseudocode · Proof of correctness · Running time

www.hackerearth.com › ... › Shortest Path Algorithms ▾

Shortest Path Algorithms Tutorials & Notes | Algorithms ...

The **shortest path** problem is about finding a **path** between vertices in a **graph** such that the total sum of the edges weights is minimum. This problem could be solved easily using (BFS) if all edge weights were ( ), but here weights can take any value.

People also ask

- What are the shortest path algorithms?
- Which is the best shortest path algorithm?
- How do you solve Dijkstra's shortest path algorithm?
- Does A * find the shortest path?

www.geeksforgeeks.org › dijkstras-shortest-path-algorit... ▾

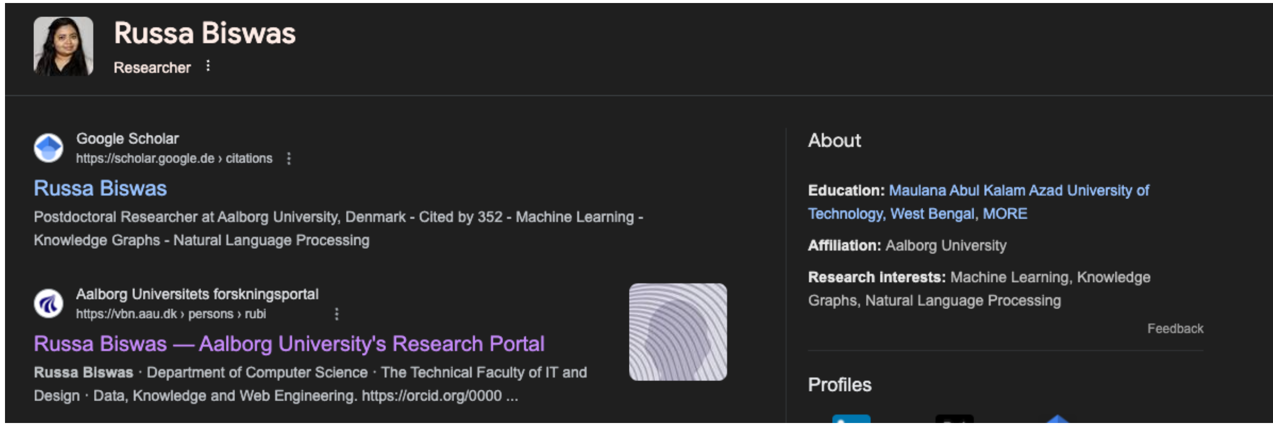
Dijkstra's shortest path algorithm | Greedy Algo-7

Dijkstra's **shortest path algorithm** | Greedy Algo-7. Last Updated: 22-04-2020. Given a graph

➡Relevant and competent resources

Intelligent Search

- Keyword Search: the words in the query appear frequently in the document, in any order (*bag of words*).
- Disadvantages:
  - May not retrieve relevant documents that include synonymous terms (e.g., cannot distinguish between “restaurant” and “café”)
  - May retrieve irrelevant documents that include ambiguous terms (e.g., cannot distinguish between “bat” mammal and “bat” baseball)
- Beyond Keywords:
  - Considering the meaning of the words used
  - Adapting to user feedback (direct or indirect)
  - Considering the authority of the source



Russa Biswas

Researcher

Google Scholar

https://scholar.google.de › citations

Russa Biswas

Postdoctoral Researcher at Aalborg University, Denmark - Cited by 352 - Machine Learning - Knowledge Graphs - Natural Language Processing

Aalborg Universitets forskningsportal

https://vbn.aau.dk › persons › rubi

Russa Biswas — Aalborg University's Research Portal

Russa Biswas · Department of Computer Science · The Technical Faculty of IT and Design · Data, Knowledge and Web Engineering. https://orcid.org/0000 ...

About

Education: Maulana Abul Kalam Azad University of Technology, West Bengal, MORE

Affiliation: Aalborg University

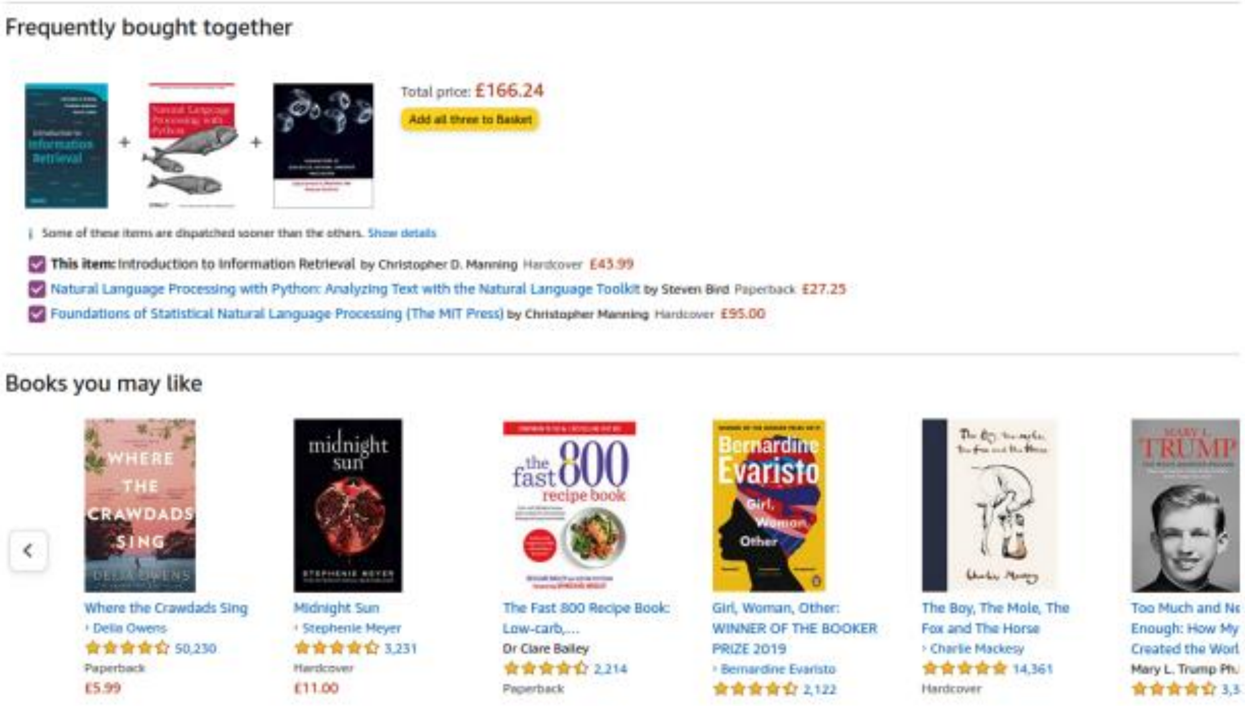
Research Interests: Machine Learning, Knowledge Graphs, Natural Language Processing

Feedback

Profiles

Recommender Systems

Looking on Amazon for the Manning IR book:



Frequently bought together

Total price: £166.24

Add all three to Basket

Some of these items are dispatched sooner than the others. Show details

☑ **This item:** Introduction to Information Retrieval by Christopher D. Manning Hardcover £43.99

☑ Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit by Steven Bird Paperback £27.25

☑ Foundations of Statistical Natural Language Processing (The MIT Press) by Christopher Manning Hardcover £95.00

Books you may like

Where the Crawdads Sing
› Delia Owens
★★★★☆ 50,230
Paperback
£5.99

Midnight Sun
› Stephenie Meyer
★★★★☆ 3,231
Hardcover
£11.00

The Fast 800 Recipe Book: Low-carb,...
Dr Clare Bailey
★★★★☆ 2,214
Paperback

Girl, Woman, Other: WINNER OF THE BOOKER PRIZE 2019
› Bernardine Evaristo
★★★★☆ 2,122

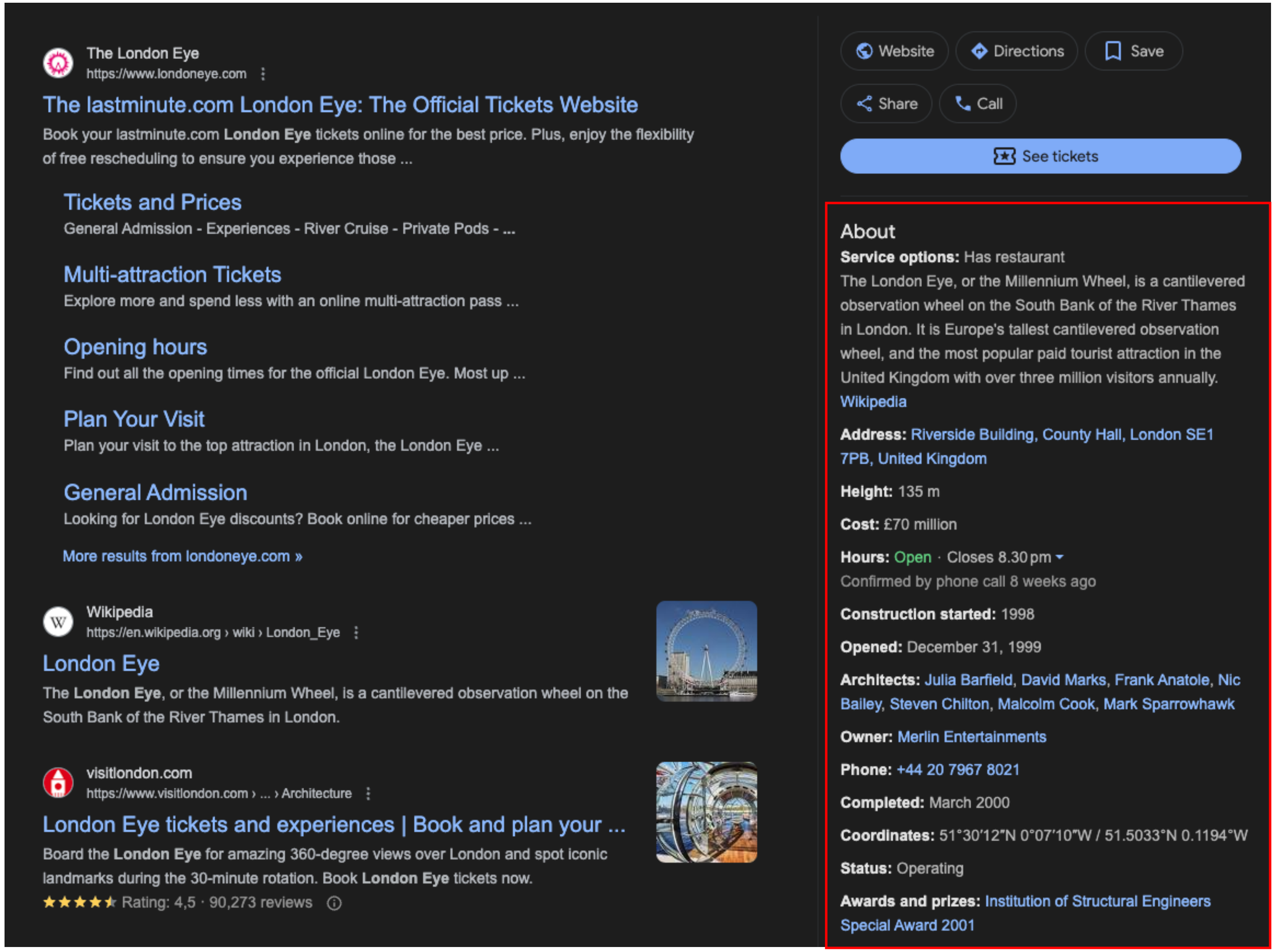
The Boy, The Mole, The Fox and The Horse
› Charlie Mackesy
★★★★★ 14,361
Hardcover

Too Much and Ne
Enough: How My
Created the Worl
Mary L. Trump Ph.
★★★★☆ 3,3

Kinds of recommendations:

- Product Based (collaborative filtering): similar books
- User- Based (content based filtering): based on search history
- Hybrid

Knowledge Graphs



The London Eye
https://www.londoneye.com

The lastminute.com London Eye: The Official Tickets Website

Book your lastminute.com **London Eye** tickets online for the best price. Plus, enjoy the flexibility of free rescheduling to ensure you experience those ...

Tickets and Prices
General Admission - Experiences - River Cruise - Private Pods - ...

Multi-attraction Tickets
Explore more and spend less with an online multi-attraction pass ...

Opening hours
Find out all the opening times for the official London Eye. Most up ...

Plan Your Visit
Plan your visit to the top attraction in London, the London Eye ...

General Admission
Looking for London Eye discounts? Book online for cheaper prices ...

More results from londoneye.com »

Wikipedia
https://en.wikipedia.org › wiki › London_Eye

London Eye

The **London Eye**, or the Millennium Wheel, is a cantilevered observation wheel on the South Bank of the River Thames in London.

visitlondon.com
https://www.visitlondon.com › ... › Architecture

London Eye tickets and experiences | Book and plan your ...

Board the **London Eye** for amazing 360-degree views over London and spot iconic landmarks during the 30-minute rotation. Book **London Eye** tickets now.

★★★★★ Rating: 4,5 · 90,273 reviews

About

Service options: Has restaurant

The London Eye, or the Millennium Wheel, is a cantilevered observation wheel on the South Bank of the River Thames in London. It is Europe's tallest cantilevered observation wheel, and the most popular paid tourist attraction in the United Kingdom with over three million visitors annually. Wikipedia

Address: Riverside Building, County Hall, London SE1 7PB, United Kingdom

Height: 135 m

Cost: £70 million

Hours: Open · Closes 8.30 pm
Confirmed by phone call 8 weeks ago

Construction started: 1998

Opened: December 31, 1999

Architects: Julia Barfield, David Marks, Frank Anatole, Nic Bailey, Steven Chilton, Malcolm Cook, Mark Sparrowhawk

Owner: Merlin Entertainments

Phone: +44 20 7967 8021

Completed: March 2000

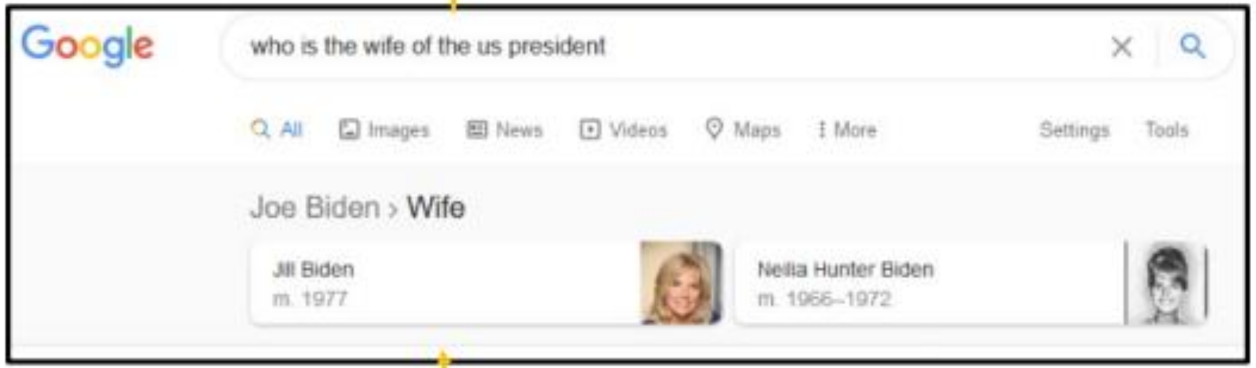
Coordinates: 51°30′12″N 0°07′10″W / 51.5033°N 0.1194°W

Status: Operating

Awards and prizes: Institution of Structural Engineers Special Award 2001

Search Engines store information about entities in Knowledge Graphs and use them to summarise search results

Question Answering

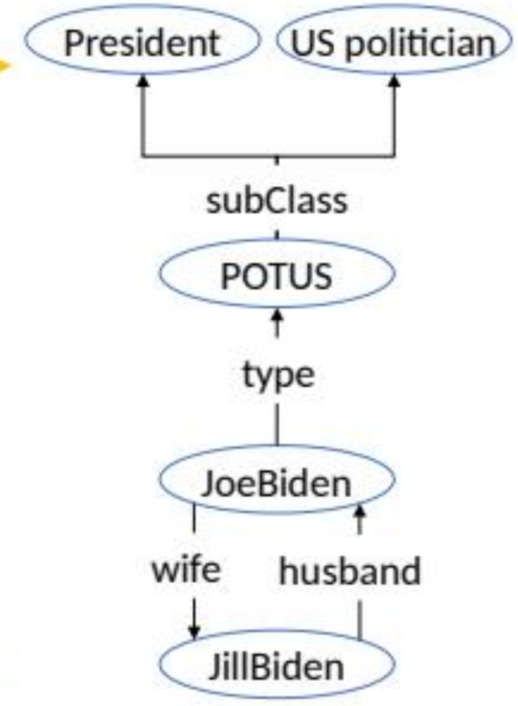


who is the wife of the us president

Joe Biden › Wife

Jill Biden
m. 1977

Neilia Hunter Biden
m. 1966–1972

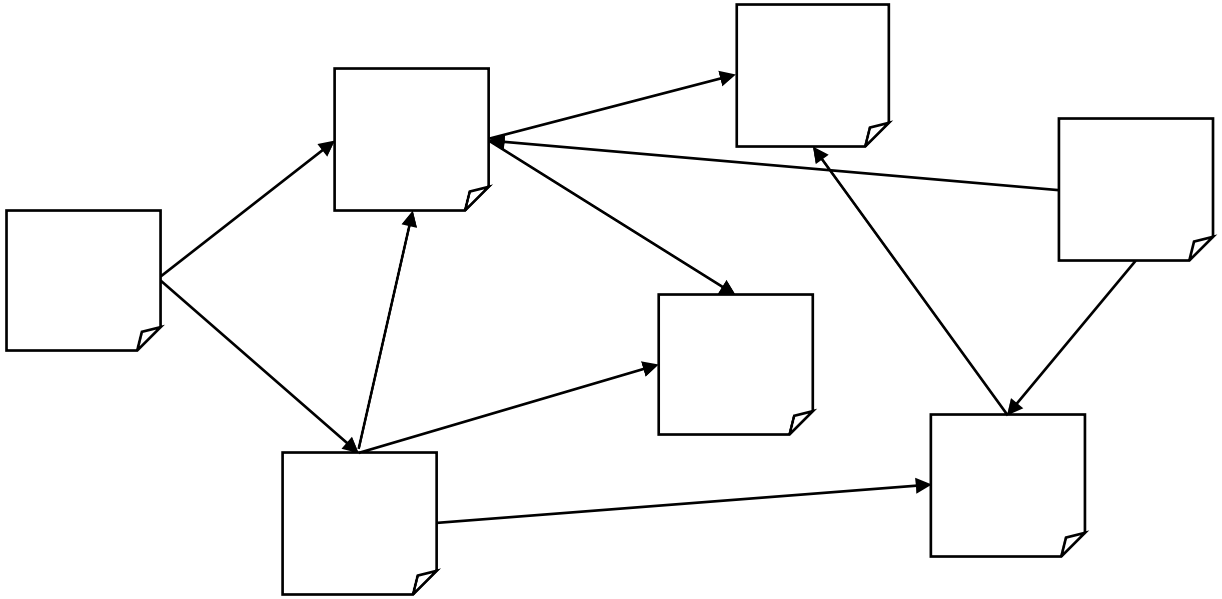


The same knowledge graphs can be used to find answers to user's questions (instead of finding the documents related to the search keywords)

[Google - https://www.google.com/search?q=who+is+the+wife+of+the+us+president]

Node Classification

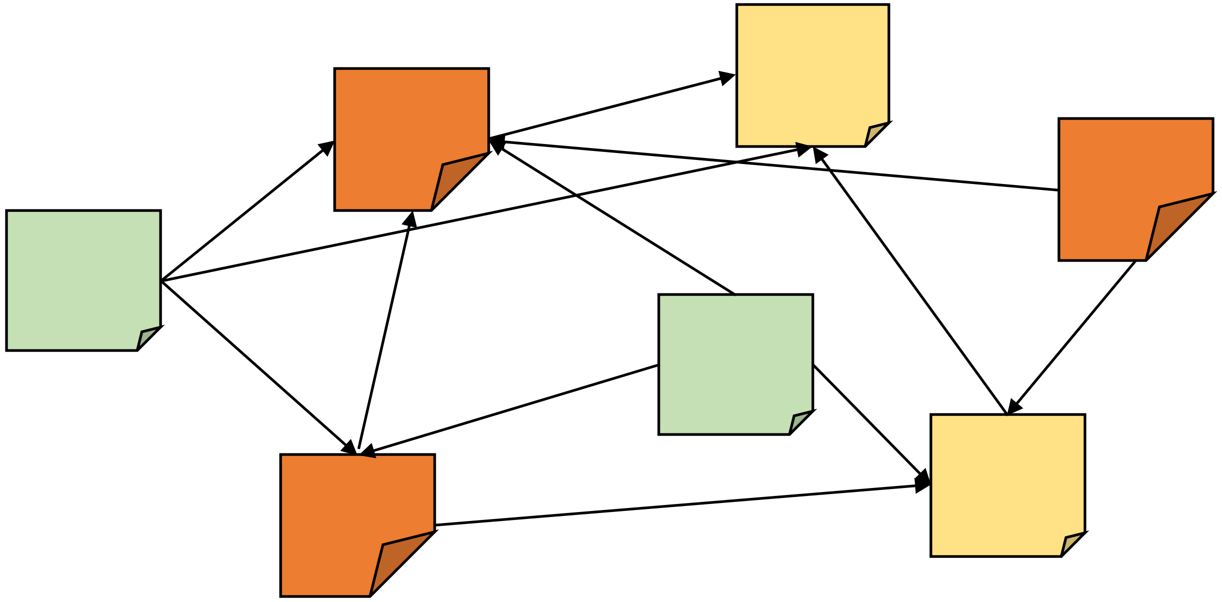
- Suppose this is aau.dk domain
- Identify the personal homepages, educational programs, research projects etc.



Node Classification

- Suppose this is aau.dk domain

- Identify the personal homepages, educational programs, research projects etc.



🟧 Educational programs

🟨 Research Projects

🟩 Personal Homepages

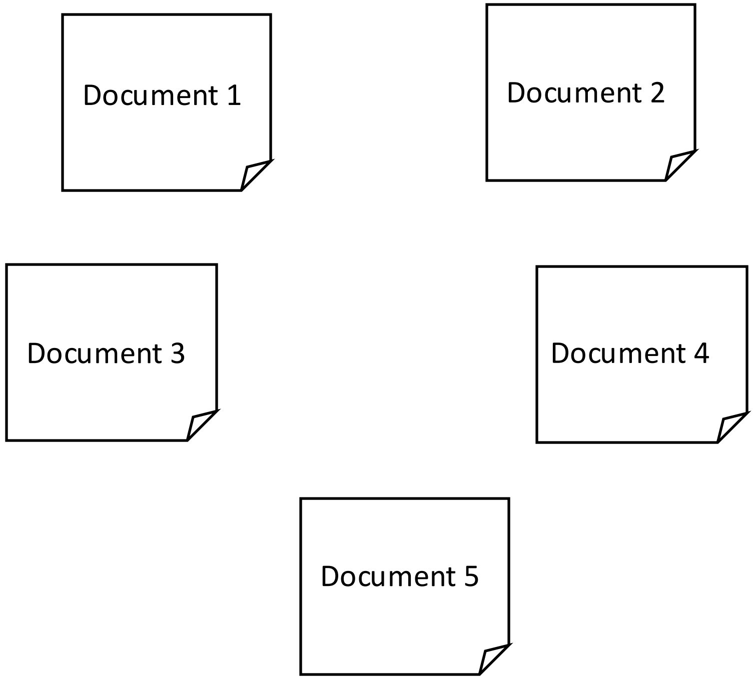
Prediction models are based on node attributes and link attributes

Information diffusion

- Information diffusion: process by which a piece of information or knowledge is spread and reaches individuals through interactions
  - Studied in Sociology, epidemiology (the study of the determinants, occurrence, and distribution of health and disease in a defined population)

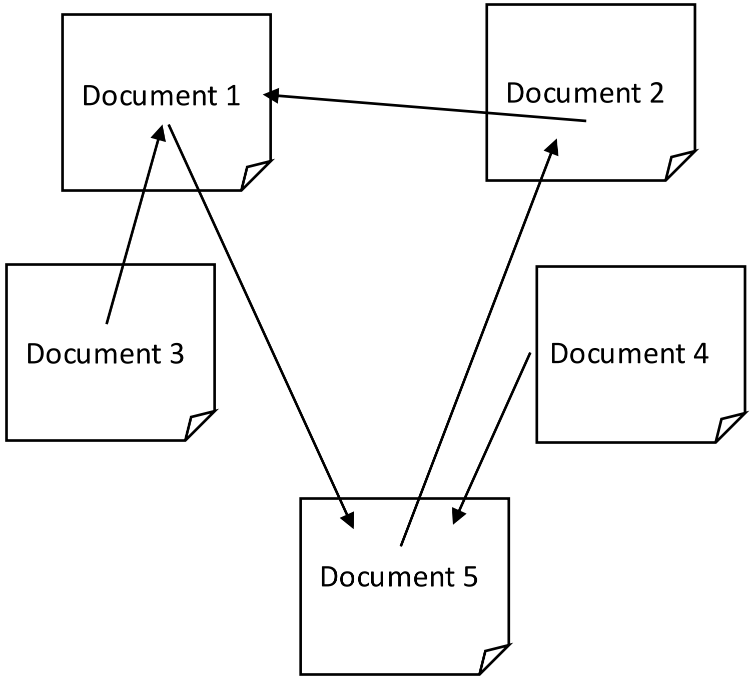
Different levels of Web

- We can distinguish 3 levels (perspectives) of modeling and analytics
- Level 1: Web Content – The web is a collection of documents



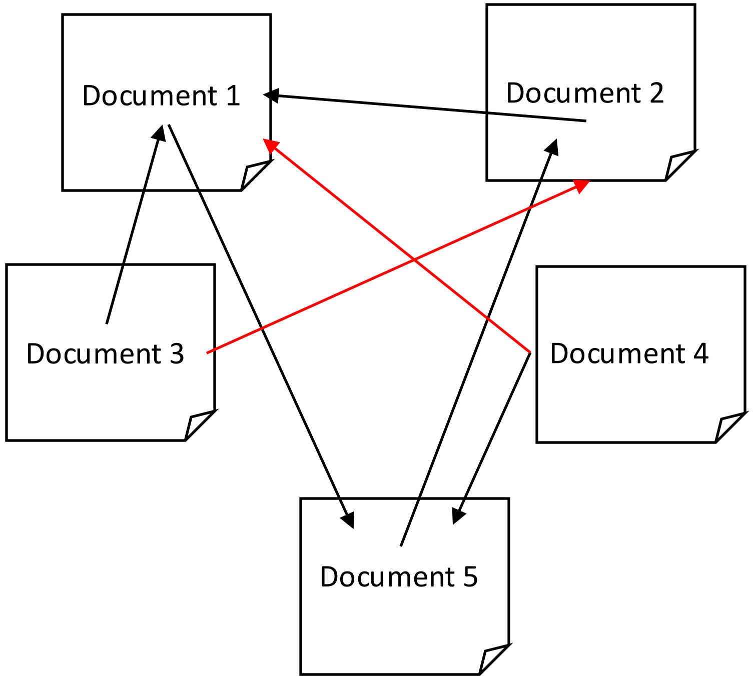
Different levels of Web

Level 2: Web Structure – The web as a network of documents:



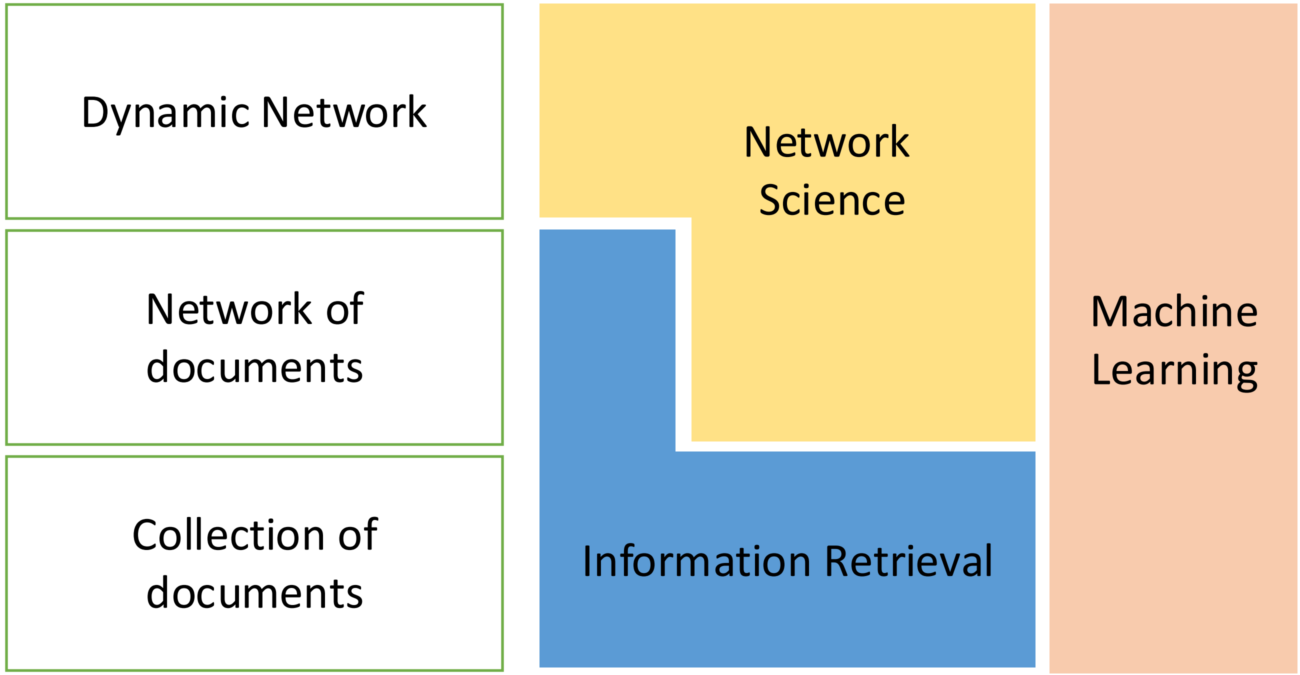
Different levels of Web

Level 3: Web Dynamics – The web as a dynamic network



- Network evolution
- Dynamic processes on the web

Relationship to Scientific disciplines



Dynamic Network

Network of documents

Collection of documents

Network Science

Information Retrieval

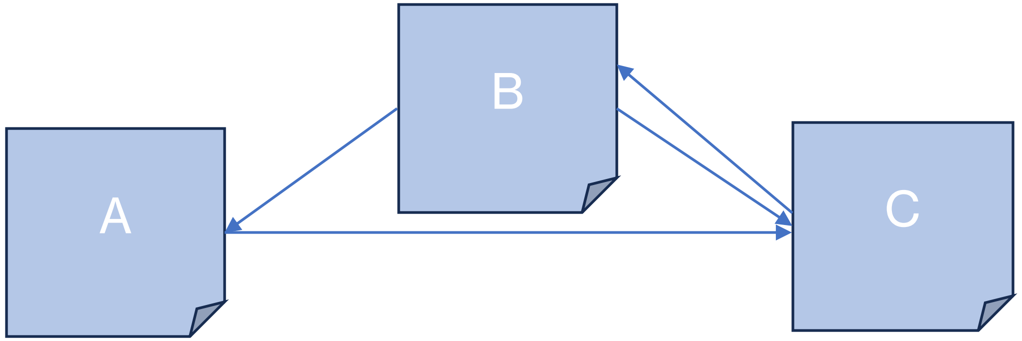
Machine Learning

Open research Area: Hetereogeneous networks

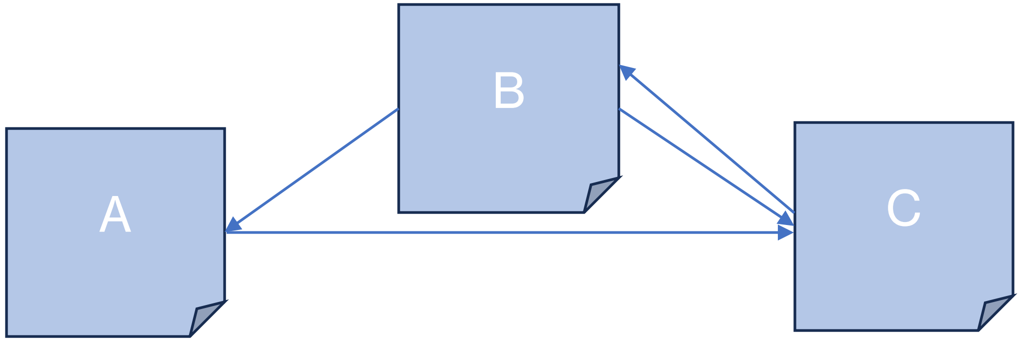
Types of Links in Webpages

Types of links:

- Inbound links: The links into the webpage from the outside
- Outbound links: Links from a webpage to other pages in a site or other websites
- Dangling links: Dangling links that point to any page with no outgoing links



Types of Links in Webpages



Inbound A =

Inbound B =

Inbound C =

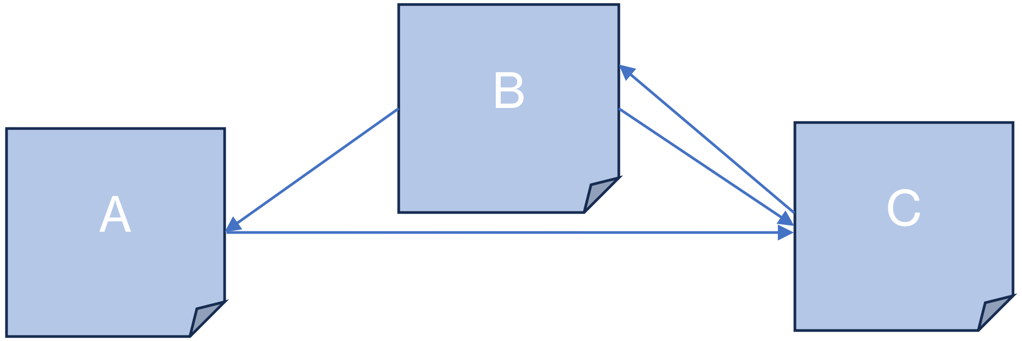
Outbound A =

Outbound B =

Outbound C =

Dangling =

Types of Links in Webpages



Inbound A = 1

Inbound B = 1

Inbound C = 2

Outbound A = 1

Outbound B = 2

Outbound C = 1

Dangling = 0

Course Coverage

- Why do we need this data?
- How do we extract the knowledge available on the Web?
- What are the challenges?
- What are the different forms of data available?
- How do we use the data effectively?
- How do we model the different data formats efficiently?
- How can we use this data to make our life easier?

Literature

1. https://www.w3.org/TR/webarch/
2. Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H.F. and Secret, A., 1994. The world-wide web. *Communications of the ACM*, *37*(8), pp.76-82.
3. Berners-Lee, T., Cailliau, R., Groff, J.F. and Pollermann, B., 1992. World-Wide Web: the information universe. *Internet Research*, *2*(1), pp.52-58.
4. https://www.w3.org/standards/
5. Berners-Lee, T.J., 1992. The world-wide web. *Computer networks and ISDN systems*, *25*(4-5), pp.454-459
6. https://home.cern/science/computing/birth-web/short-history-web