

Web Search and Data Mining

Computer Science MSc Course

João Magalhães

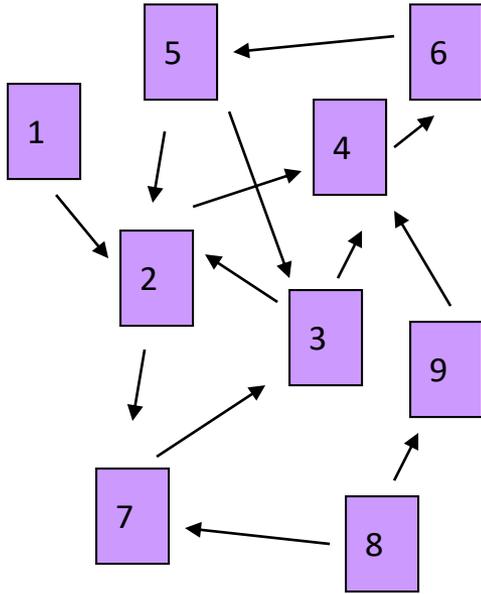
How to search Web information?

- Textual and visual data can communicate a wide variety of information that are critical for several decision processes.
- Temporal and spatial structure adds organization and usability to information.
- Non-structured data (language and vision) puts a heavy complexity burden on standard data structures.

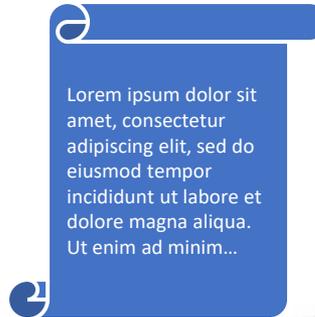
Course plan

Web Mining and Search		
Week #	Lecture	In-class labs
05/mar/20	1 Introduction	1 Lab setup
12/mar/20	2 Web data representation	2 Data representations
19/mar/20	3 Web-graph analysis	3 Project
26/mar/20	4 Recommendation algorithms	4 Project
02/abr/20	5 Learning vision data representations	5 Project checkpoint
09/abr/20	6 Natural language representations	6 Project
16/abr/20	7 Case studies ECIR	7 Case studies ECIR
23/abr/20	8 Named Entities and Knowledge Graph	8 Project
30/abr/20	9 Multimodal representations	9 Project
07/mai/20	10 Locality sensitive hashing	10 Project checkpoint
14/mai/20	11 Visual Question Answering	11 Project
21/mai/20	12 Paper summary	12 Project
28/mai/20	13 Paper summary	13 Project
04/jun/20	14 Revisions	14 Project
11/jun/20	Test	Project submission

Web data



Links



Text



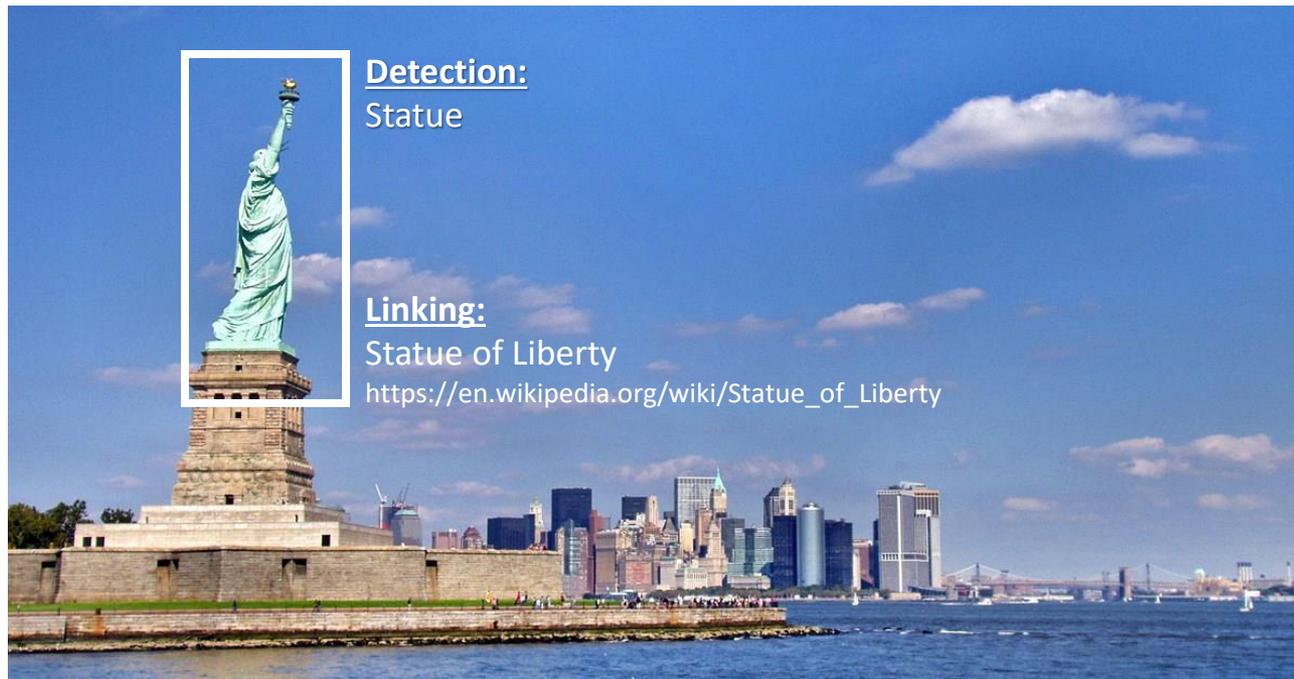
Preferences



Images/videos



Classification, detection, linking

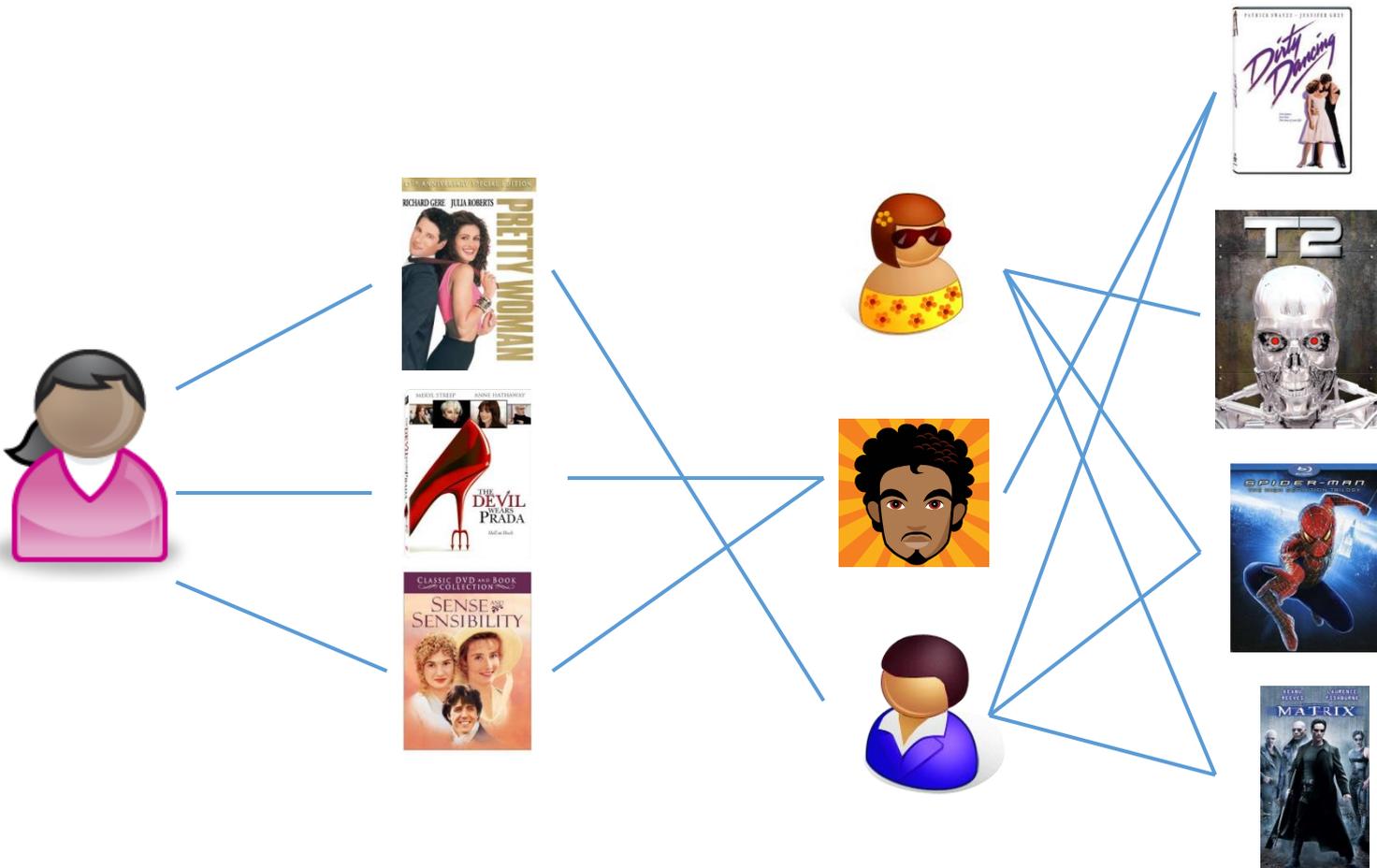


Detection:
Statue

Linking:
Statue of Liberty
https://en.wikipedia.org/wiki/Statue_of_Liberty

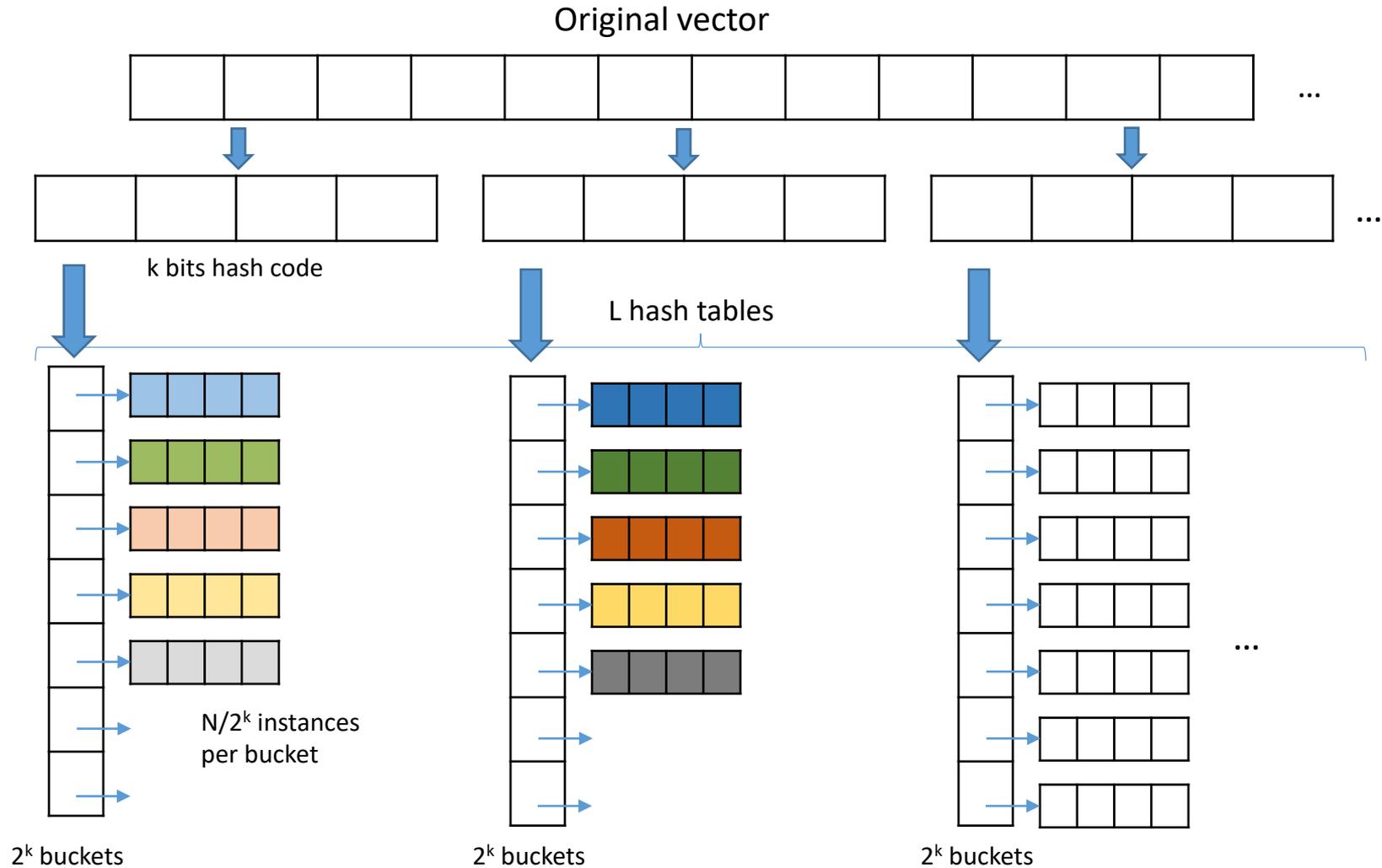
Classification:
Sea side
Statue
City
Sky

Collaborative filtering





Large-scale search data structures



Funny reflex...

Share This

[+](#) [ADD NOTE](#) [SEND TO GROUP](#) [+](#) [ADD TAGS](#) [BLOG THIS](#) [ALL SIZES](#) [ORDER SLIDES](#) [ROTATE](#) [EDIT PHOTO](#) [X](#) [DELETE](#)



Uploaded on August 3, 2006 by [jmc_mag](#)

[+](#)

Best ones (Set)

You are at the first photo,  **11 items**

[browse](#)

Tags

- Arizona [×](#)
- Reflection [×](#)
- Desert [×](#)

[Add a tag](#)

Comments



[alexei_322](#) pro says:

thumbs up! great shot

Posted 31 months ago. ([permalink](#) | [delete](#))

Add your comment

Additional Information

- All rights reserved ([edit](#))
- Anyone can see this photo ([edit](#))
- [Add to your map](#)
- Taken with a [Canon EOS Digital Rebel XT](#). [More properties](#)
- Taken on July 15, 2006 ([edit](#))
- [Photo stats](#)
- Viewed 26 times (Not including you)
- [Edit title, description, and tags](#)

Web data based search

The image shows a collage of three web search engine interfaces. The top interface is Yahoo!, featuring the logo, navigation links like 'NEW COOL RANDOM', and a search bar. The middle interface is AltaVista, with the logo, a search bar, and a red banner for 'AUTOTE! Car Buying & Car Insurance Pain Relief'. The bottom interface is Google! BETA, with the logo, a search bar, and buttons for 'Google Search' and 'I'm feeling lucky'. A sidebar on the left of the Google! interface lists various categories like Arts, Business and Economy, Computers and Internet, Education, Entertainment, Government, Health, News, Recreation and Sports, Reference, Regional, and Science.

Yahoo!
NEW COOL RANDOM HEAD YAHOO! ADD LINES INFO URL
CLICK HERE TO VISIT THE STARS **YAHOO! LOS ANGELES** Weekly Picks
Yahoo! Deutschland

ALTAVISTA
Technology
View Multimedia From Our Vantage Point
AUTOTE! Car Buying & Car Insurance Pain Relief
USA CANADA Buy and insure new cars & trucks online LOW-COST
Click here for advertising information - reach millions every month!

Google! BETA
Search the web using Google!
Special Searches: Stanford Search, Linux Search
Help: About Google!, Company Info, Google! Logos
Get Google! updates monthly: your e-mail, Subscribe, Archive
Copyright ©1998 Google Inc.

Online shopping

★ macy's ORDER TRACKING STORES WEDDING REGISTRY SHIPPING TO 🇺🇸

SHOP BY DEPARTMENT ▾ Search or enter web ID 🔍 🛒

Macy's / Women / Sweaters

Cashmere Shop **1044 items in Sweaters**

Filter By

Offers

- Clearance/Closeout (197)
- Last Act (85)
- Sales & Discounts (437)

Sweater Style +

Size Range +

Size +

Brand +

Color -

- Black Blue Brown Gold
- Gray Green Ivory/Cr... Mult
- Orange Pink Purple Red
- Silver Taupe/Beige White Yellow

Sleeve Length +

Price +

Discount Range +

Sort by **Featured Items** ▾


 Charter Club Pure Cashmere Solid Crewneck Sweater in Regular & Petite Sizes, Created for Macy's
 USD 139.00 ★★★★★ (880) [More Like This](#)


 Charter Club Pure Cashmere Turtleneck Sweater in Regular & Petite Sizes, Created for Macy's
 USD 139.00 ★★★★★ (327) [More Like This](#)


 Charter Club Pure Cashmere V-neck Sweater in Regular & Petite Sizes, Created for Macy's
 USD 139.00 ★★★★★ (968) [More Like This](#)


 Charter Club Pure Cashmere Duster in Regular & Petite


 Charter Club Pure Cashmere Solid Basic Poncho, Created


 Charter Club Bell-Sleeve Cardigan, Created for Macy's

NEED IT FASTER?
 Choose **FREE PICK UP IN STORE** + get **EXTRA 20% OFF** your next store purchase. Exclusions apply. [FIND OUT MORE](#)

★ macy's ORDER TRACKING STORES WEDDING REGISTRY SHIPPING TO 🇺🇸

SHOP BY DEPARTMENT ▾ Search or enter web ID 🔍 🛒

Macy's / Men / Shirts

Polo Shirts T-Shirts **2691 items in Shirts**

Filter By

Offers

- Clearance/Closeout (99)
- Last Act (44)
- Sales & Discounts (117)

Shirt Type +

Size +

Brand +

Shirt Fit +

Color +

Fabric +

Pattern -

- Camo (11)
- Check (35)
- Colorblock (15)
- Dots (29)
- Embroidered (2)
- Floral (18)
- Geometric (6)
- Graphic (1063)
- Herringbone (1)
- Plaid (122)
- Print (487)

Customers Top Rated +

Price +

Sort by **Featured Items** ▾


 Weatherproof Vintage Men's Heathered Henley Special Savings
 USD 44.00 Sale USD 30.80 ★★★★★ (60) [More Like This](#)


 Weatherproof Vintage Men's Indigo Denim Shirt Special Savings
 USD 60.00 Sale USD 42.00 ★★★★★ (22) [More Like This](#)


 Weatherproof Vintage Men's Plaid Flannel Shirt Special Savings
 USD 60.00 Sale USD 42.00 ★★★★★ (44) [More Like This](#)


 Weatherproof Vintage Men's Pinstriped Flannel Shirt Special Savings
 USD 60.00


 NEW! Weatherproof Vintage Men's Brushed Jersey T-Shirt


 NEW! Tallia Men's Slim-Fit Ochre Floral Print Dress Shirt Special Savings

NEED IT FASTER?
 Choose **FREE PICK UP IN STORE** + get **EXTRA 20% OFF** your next store purchase. Exclusions apply. [FIND OUT MORE](#)

Medical domain

(B) Free text query →

(A) Image drop area →

Current query →

(C) Recognized medical terms

(D) Knowledge-based assisted expansion

(E) Case-based search result

The screenshot displays a search interface with the following components:

- Query Bar:** Contains tags for 'painless', 'hematuria', 'abdominal', 'tomography, spiral computed', 'renal mass', and 'pelvis'. A dropdown menu is open below 'tomography, spiral computed', showing 'Also matches' suggestions: 'spiral volumetric ct', 'tomography, helical computed', 'spiral ct', 'helical ct', 'tomography, spiral volumetric computed', 'helical computed tomography', and 'spiral computed tomography'.
- Image Drop Area:** Features a green button '+ Add images (JPG only)...' and a blue 'Search' button.
- Results Section:** Titled 'Results for: painless hematuria abdominal tomography spiral computed renal mass left renal pelvis ureter'. It includes two thumbnail images of CT scans and a case report titled 'Massive hematuria due to a congenital renal arteriovenous malformation mimicking a renal pelvis tumor: a case report' by Sountoulides, P; Zachos, I; Paschalidis, K; Asouhidou, I; Fotiadou, A; Bantis, A; Palasopoulou, M; Podimatas, T. The report text begins with 'Introduction Congenital renal arteriovenous malformations (AVMs) are very rare benign lesions. They are more common in women and rarely manifest in elderly people. In some cases they present with massive hematuria. Contemporary'.
- Case-based Search Results:** Three additional CT scan thumbnails are shown below the case report.

Course grading

- 40% theoretical part (1 test or 1 exam)
- 50% for a 3 parts project
 - Submission 1
 - Submission 2
 - 100% Final submission
- 10% case study presentation
- Groups of 3 students, maximum 8 groups
- Additional rules:
 - Minimum grade on the labs or theory: 9
 - You may use one sided A4 sheet handwritten by you with your notes
 - It must be handed at the end of the test.

Project grading

- Scoring:

- Implement. correctness 30%
- Results analysis 30%
- Critical discussion 40%

- Report:

- Maximum of 8 pages.
- No cover page.
- Must include graphs, tables, etc.

- Report organization:

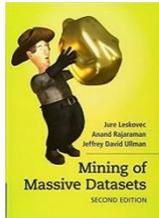
- Introduction
- Algorithms
- Implementation
- Evaluation
 - Dataset description
 - Baselines
 - Results analysis
- Critical discussion
- References

Case studies

1. Crowd Knowledge Enhanced **Multimodal Conversational** Assistant in Travel Domain
2. TweetFit: Fusing Multiple Social Media and Sensor Data for **Wellness Profile Learning**
3. Knowledge-aware **Multimodal Dialogue** Systems
4. Multi-modal Knowledge-aware Hierarchical Attention Network for Explainable **Medical Question Answering**
5. DeFacto - Temporal and Multilingual Deep **Fact Validation**
6. Modeling **Temporal Evidence** from External Web Collections
7. Ranking **News-Quality** Multimedia
8. **Crowdsourcing** facial expressions for affective-interaction

References

- Slides and articles provided during classes.
- Books:



Jure Leskovec, Anand Rajaraman, Jeff Ullman, “Mining of Massive Datasets”, Cambridge University Press, 2011.

<http://www.mmds.org/>



Aston Zhang, Zachary Lipton, Mu Li, and Alex Smola, “Dive into Deep Learning”

<http://d2l.ai/>