Università
della
Svizzera
italiana

**Facoltà
di scienze
informatiche**

Information Retrieval
Academic Year 2024-25
Course Project
Prof. Monica Landoni

For this assignment you will work in pairs to implement a working prototype of an information retrieval system for a specific task and user needs. The goal of this assignment is to apply the concepts and tools you have learned in the class to a practical, real world, application.

Description

Each group was assigned a topic, together with an example website you need to crawl your data from. You need to build a system that gathers a large collection of samples associated with the topic and enable the search over this collection. To build the collections, you need to crawl multiple sources (websites), similar in content and topic to the website suggested. The system must provide an interface for searching, browsing, and presentation of the data to the user.

The system should also have two additional features. Additional features are categorized into two groups: complex and simple features.

Simple Features

Results presentation: results should be presented in a tabular format, so that many results could be seen at the same time. Each table cell should contain appropriate information for your project.

Filtering: in addition to being able to search by title, an user should be able to filter the results based on at least 3 relevant attributes for your project.

Results Snippets: present result snippets of each retrieved result (maximum 2-3 lines) in a kind of “Google style”, with query terms highlighted.

Complex Features

Automatic Recommendation: In addition to the relevant search results pertaining to the user query, the user should also be suggested “similar” products based on, say, category, description, price etc. The ordering among the recommended items is not important. However, you should mention how did you arrive at the recommendations. For this purpose, you can use any open-source recommenders available.

User Relevance Feedback: after presenting the search results to a user, the user may provide a positive or negative feedback on the results (i.e. mark relevant and irrelevant

documents). Based on this feedback the search results have to be updated and presented again.

Results clustering: the system should group results into topics. There should be a possibility for expanding/shrinking a topic to show all the results related to the topic. In addition, topics should be sorted in a descending order by the number of relevant results under them.

You can choose which features you want to implement yourself, provided there is at least one feature from each of the groups (two in total).

Submission procedure and evaluation

An important step is the evaluation of your tool. Three students of the class, of your choice, will act as test users of your system and help you in evaluating your system from the point of view of the user experience. **System evaluation (i.e. recall and precision) is not required**. Please, coordinate with the other groups in order to meet the deadline.

On 27 November end of the day each group will need to submit one page describing their progress so far.

Thereafter, you have to produce a report (max. 10 pages, including cover) of your work and of the evaluation. It will contain a concise explanation of how you tackled the design and implementation of the system. The code of the project and the report need to be submitted via iCorsi by end of day 16 December.

Projects:

N.	Project	Guidelines	Example Website
1	**IKEA hacks**	Choose at least 3 different websites with IKEA hacks \ articles about IKEA furniture \ reviews of IKEA products.	https://ikeahackers.net/category/hacks
2	**TV Series Episodes**	Choose at least 3 different TV series/Soap operas with a large number of episodes (for reference: https://en.wikipedia.org/wiki/List_of_television_programs_by_episode_count)	https://coronationstreet.fandom.com/wiki/Coronation_Street_episodes
3	**Volunteer camps**	Choose at least 3 websites listing volunteer camps or opportunities.	https://www.workcamps.sci.ngo/icamps/welcome.html or https://www.volunteerworld.com/en
4	**Hairdressers**	Choose at least 3 websites and then focus on a specific city, any other kind of professional (plumbers, lawyers etc) is also ok. If you want to mix and match (like: plumbers, electricians and painters, everything for the home, or hairdressers + aestheticians) it's also fine.	https://www.fresha.com/lp/en/bt/hair-salons/in/gb-london
5	**Clothing**	Choose at least 3 websites listing clothing (Asos, Zalando, Shein...)	https://www.asos.com/
6	**Short-term rentals**	Choose at least 3 websites, then focus on a specific city\area.	https://www.airbnb.co.uk/
7	**Textbooks**	Choose at least 3 websites listing textbooks (any subject)	https://open.umn.edu/opentextbooks
8	**Charities**	Choose at least 3 websites listing charities.	https://www.guidestar.org/NonprofitDirectory.aspx
9	**Indie Games**	Choose at least 3 websites listing indie videogames (commercial platforms like Steam are ok but only crawl indie games)	https://itch.io/games
10	**Stocks**	Choose at least 3 different stock exchanges.	https://www.borsaitaliana.it/borsa/azioni/listino-a-z.html?initial=A&lang=en
11	**Travel experiences**	Choose at least 3 websites listing travel experiences/tour, then focus on a specific city\area\country.	https://www.getyourguide.co.uk/london-l57/
12	**PhD positions**	Choose at least 3 websites listing PhD positions, any subject.	https://academicpositions.com/jobs/position/phd
13	**Content creators**	Choose at least 3 websites for Content Creators (like Patreon, Ko- fi...but please keep it safe for work!) and crawl those to create a search engine of mangas.	https://www.patreon.com/en-GB

14	**Concerts**	Choose at least 3 websites, then focus on a specific country, or 3 smaller countries (Switzerland is a small country!)	https://www.ticketmaster.ch/music/concerts/1221/events?language=en-us
15	**Pokemon**	Choose at least 3 websites listing pokemons.	https://www.pokemon.com/us/pokedex/
16	**Pets for adoption**	Choose at least 3 websites listing pets for adoption.	https://www.petfinder.com/
17	**Nannies/babysitters**	Choose at least 3 websites listing nannies\babysitters, then focus on a specific city\areas.	https://www.childcare.co.uk/find/Babysitters
18	**Motorcycles**	Choose at least 3 websites listing motorcycles.	https://www.motorcycle.com/
19	**Mangas**	Choose at least 3 websites listing Mangas. Manwha or Manhua (Korean or Chinese comics) are also ok!	https://www.viz.com/read
20	**Second-hand items**	Choose at least 3 websites offering second-hand items.	https://www.secondhand.org.uk/
21	**Football teams**	Choose at least 3 websites listing football teams, not necessarily from the same country.	https://footballdatabase.com/
22	**Medications**	Choose at least 3 websites listing medications.	https://www.drugs.com/drug_information.html
23	**Recipes**	Choose at least 3 websites listing recipes.	https://www.allrecipes.com/

Groups:

	Partner 1	Partner 2	Project number
1	Daniel Dorigo	Giovanni Elisei	9
2	Luca Rossinelli	Mychaylo Bozhko	2
3	Michelangelo Bettini	Ivan Angjelovski	7
4	Oubenali Jamila	Xincong Tong	16
5	Stefano Soardi	Etienne Orio	19
6	Vittoria Scocco	Mehmet Bayazit	22
7	Lino Candian	Wu Jiun-Yi	6
8	Adrian Biletskyi	Tyshchyk Nikita	14
9	Ioffe Andrii	Polad Bakhishzade	1
10	Mariia Batalkina	Linnikov Ivan	4
11	Matteo Ghilardini	Toscano Sasha	20
12	Costanza Rodrigues Gavazzi	Agnese Zamboni + Davide Frova	8
13	Giorgia Lillo	Boffi Sergio e Ambiveri Giulia.	17
14	Stipe Peran	Matteo Boioli	15
15	Noah Salvi	Alessandro Zaccaria	21
16	Edoardo Ababei	Joaquin Perdomo Roget	11
17	Eduard Bilous	Vladyslav Kotov	18
18	Mustafa Ozyurek	Peza Enio	10
19	Lorenzo Guideri	Leonardo Giacomo Asaro	3
20	Sergio De Crescenzo	Hector Garcia Rincon	13
21	Sofia Qu	Matthias Rota	12
22	Nicola Perucchi	Loren Shukry	5
23	Samuele Carmine Spanarelli		23