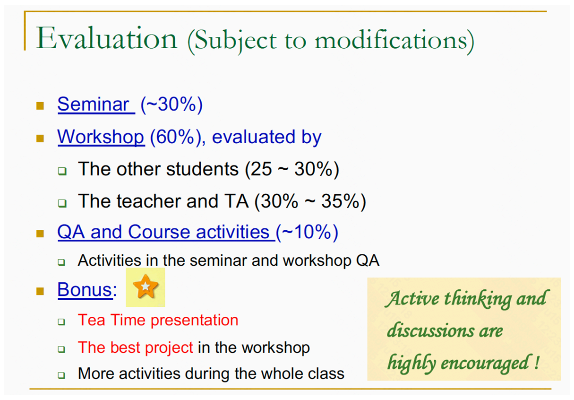
How to build a search engine

WebIR 25 Workshop Tutorial



Evaluation (Subject to modifications)

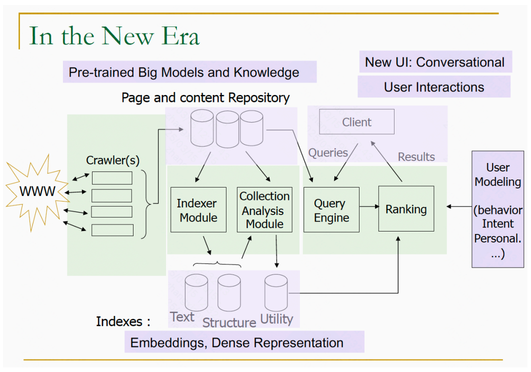
- Seminar (~30%)
- Workshop (60%), evaluated by
  - The other students (25 ~ 30%)
  - The teacher and TA (30% ~ 35%)
- QA and Course activities (~10%)
  - Activities in the seminar and workshop QA
- Bonus:
  - Tea Time presentation
  - The best project in the workshop
  - More activities during the whole class

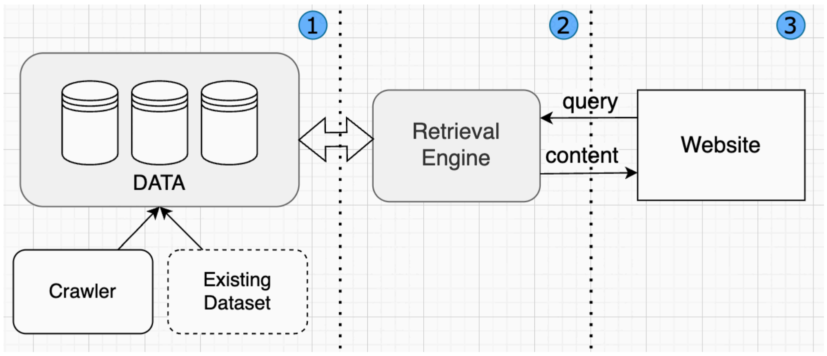
Active thinking and discussions are highly encouraged!

- What is include:
  - Basic solution & third-party package to build a SE
- What is not include:
  - UI & System design
  - Multimodality modules
  - Generation-combined retrieval
  - (encouraged but will not be detaily introduced)
- All in python (Demos)

Overall Pipeline

- In most workshop project :


In the New Era
Pre-trained Big Models and Knowledge
New UI: Conversational
User Interactions
Page and content Repository
Client
Queries
Results
Crawler(s)
WWW
Indexer Module
Collection Analysis Module
Query Engine
Ranking
User Modeling (behavior Intent Personal. ...)
Indexes :
Text
Structure
Utility
Embeddings, Dense Representation



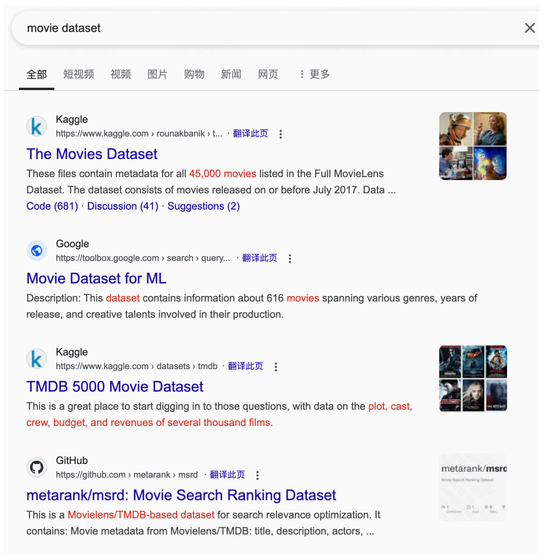
Data-Crawler

- **Request/urllib**: sending HTTP requests and handling response
- **BeautifulSoup**: scraping and parsing static web pages (fast)
- **Selenium**: simulate user interactions, load JavaScript-rendered content, and crawl dynamic websites (relatively slow)
- Others..

- ⚠️Minimize the risk of being detected and **banned** by a website:
  - *Check robots.txt*
  - ***Control request frequency***
  - *Use proper headers*
  - *…*

Data-Existing Dataset

- **Getting a perfectly matched dataset is difficult, but not impossible**


movie dataset

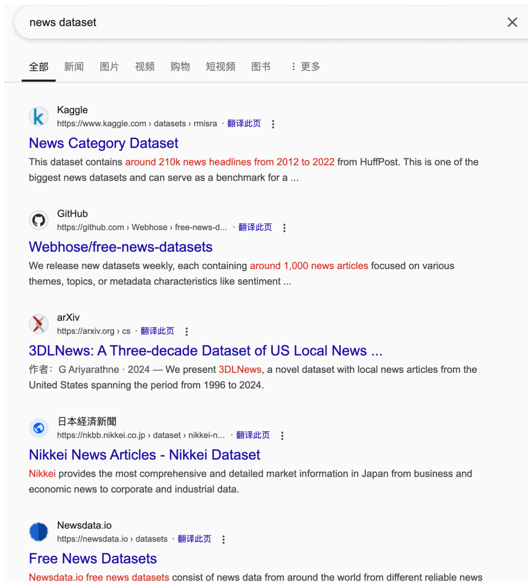
全部 短视频 视频 图片 购物 新闻 网页 ⋮ 更多

Kaggle
https://www.kaggle.com › rounakbanik › t... · 翻译此页
The Movies Dataset
These files contain metadata for all 45,000 movies listed in the Full MovieLens Dataset. The dataset consists of movies released on or before July 2017. Data ...
Code (681) · Discussion (41) · Suggestions (2)

Google
https://toolbox.google.com › search › query... · 翻译此页
Movie Dataset for ML
Description: This dataset contains information about 616 movies spanning various genres, years of release, and creative talents involved in their production.

Kaggle
https://www.kaggle.com › datasets › tmdb · 翻译此页
TMDB 5000 Movie Dataset
This is a great place to start digging in to those questions, with data on the plot, cast, crew, budget, and revenues of several thousand films.

GitHub
https://github.com › metarank › msrd · 翻译此页
metarank/msrd: Movie Search Ranking Dataset
This is a Movielens/TMDB-based dataset for search relevance optimization. It contains: Movie metadata from Movielens/TMDB: title, description, actors, ...


news dataset

全部 新闻 图片 视频 购物 短视频 图书 ⋮ 更多

Kaggle
https://www.kaggle.com › datasets › rmisra · 翻译此页
News Category Dataset
This dataset contains around 210k news headlines from 2012 to 2022 from HuffPost. This is one of the biggest news datasets and can serve as a benchmark for a ...

GitHub
https://github.com › Webhose › free-news-d... · 翻译此页
Webhose/free-news-datasets
We release new datasets weekly, each containing around 1,000 news articles focused on various themes, topics, or metadata characteristics like sentiment ...

arXiv
https://arxiv.org › cs · 翻译此页
3DLNews: A Three-decade Dataset of US Local News ...
作者： G Ariyarathne · 2024 — We present 3DLNews, a novel dataset with local news articles from the United States spanning the period from 1996 to 2024.

日本経済新聞
https://nkbb.nikkei.co.jp › dataset › nikkei-n... · 翻译此页
Nikkei News Articles - Nikkei Dataset
Nikkei provides the most comprehensive and detailed market information in Japan from business and economic news to corporate and industrial data.

Newsdata.io
https://newsdata.io › datasets · 翻译此页
Free News Datasets
Newsdata.io free news datasets consist of news data from around the world from different reliable news

Retrieval Engine

- **Index, Search, Rank**

whoosh-reloaded elastic

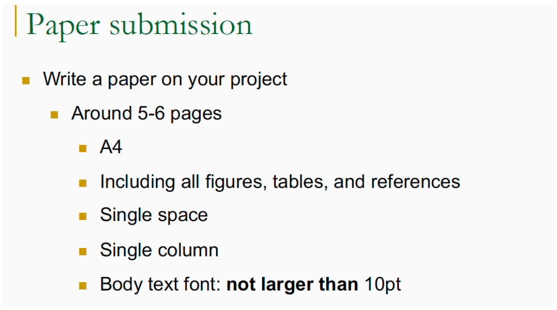
- **Whoosh**: lightweight, fast, full-text indexing and searching library
- **Elastic Search**: open source distributed, RESTful search and analytics engine)
- …
- **Build it on your own:**
  - Actually feasible, especially when dealing with data at a non-commercial scale.
  - Manage the data on your own
  - Using BM25/TF-IDF or novel algorithms specifically designed for specific tasks.

Website

- **Front-End**
  - Html/css
    - https://jekyllthemes.io/resources (template)
  - Vue
  - React
  - Element UI
  - …
- **Back-End**
  - Django
  - Flask
  - …
  - Flask and Django can be easily deployed on laptops and accessed within the campus network
- **Other tools for website development..**
  - streamlit

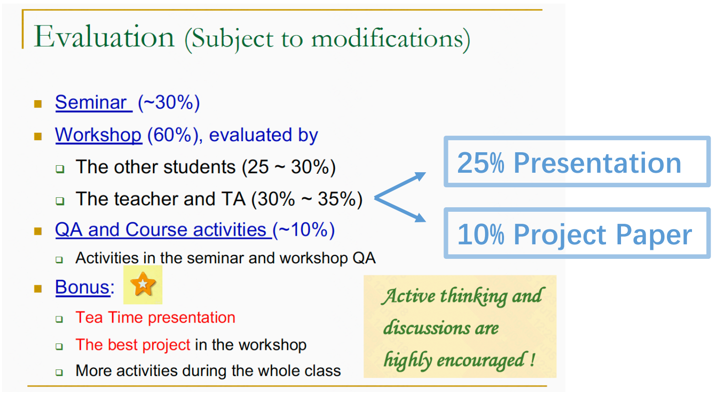
Submission & Scoring

- Generally submitted at 17-18 weeks
  - *Workshop Slides*
  - **Project Paper**
- Form groups of **1-2 students** freely.
- Be graded **separately** and submit a paper focusing on their own work.



Paper submission

- Write a paper on your project
  - Around 5-6 pages
    - A4
    - Including all figures, tables, and references
    - Single space
    - Single column
    - Body text font: **not larger than** 10pt



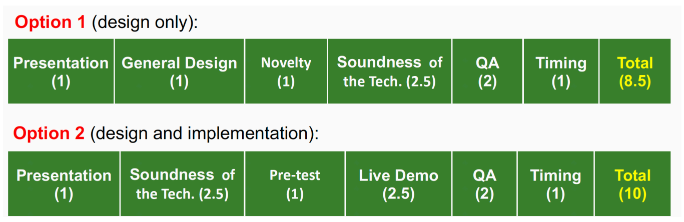
Evaluation (Subject to modifications)

- Seminar (~30%)
- Workshop (60%), evaluated by
  - The other students (25 ~ 30%)
  - The teacher and TA (30% ~ 35%)
- QA and Course activities (~10%)
  - Activities in the seminar and workshop QA
- Bonus:
  - Tea Time presentation
  - The best project in the workshop
  - More activities during the whole class

25% Presentation

10% Project Paper

Active thinking and discussions are highly encouraged !



Option 1 (design only):

Presentation (1)	General Design (1)	Novelty (1)	Soundness of the Tech. (2.5)	QA (2)	Timing (1)	Total (8.5)

Option 2 (design and implementation):

Presentation (1)	Soundness of the Tech. (2.5)	Pre-test (1)	Live Demo (2.5)	QA (2)	Timing (1)	Total (10)

What Makes a Good Course Project

- Identify the differences between your project and existing SEs on the market
  - New scenarios
  - New UI interfeces
  - New technologies
  - …
  - Even new SE paradigms

- The best project is determined by voting
  - Presentation is also important
  - Impress the listeners

Incorporating Large Language Models for Free

- Some models provide free API tokens for new users.
- One example: **ChatGLM**
  - New users receive a certain number of free tokens.
  - The **GLM4-flash** model is available for free API calls, which can meet some basic needs.
  - https://chatglm.cn/

Demo

- Proposal
  - Developing an SE for retrieving the latest ArXiv papers
  - Personalizing recommendations based on user profiles
  - Allowing users to set "Read Later" lists with corresponding deadlines

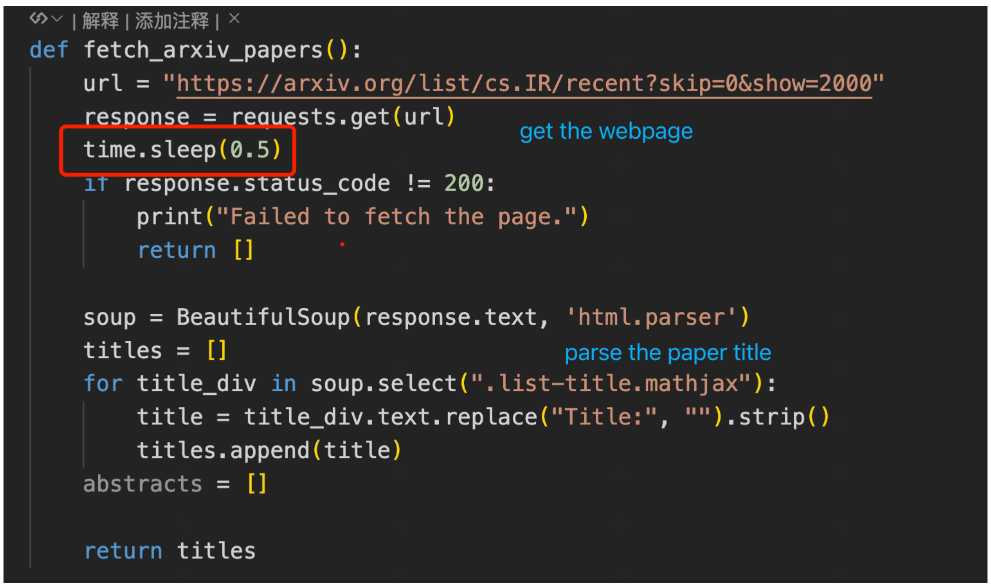
- A toy example

- Crawler: bs4
- Search Algorithm: precise match w/o ranking
- Website: Flask+HTML

Crawler

# robots.txt for http://arxiv.org/ and mirror sites http://*.arxiv.org/ # Indiscriminate automated downloads from this site are not permitted # See also: http://arxiv.org/help/robots
User-agent: * Crawl-delay: 15
Allow: /archive
Allow: /year
Allow: /list
Allow: /abs
Allow: /pdf Allow: /html
Allow: /catchup

bs4



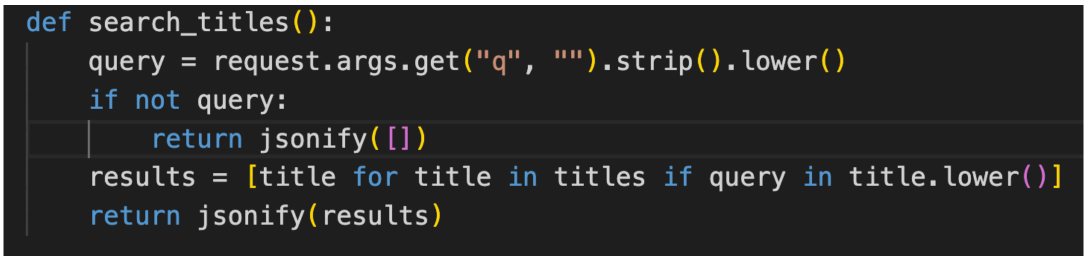
```python
def fetch_arxiv_papers():
    url = "https://arxiv.org/list/cs.IR/recent?skip=0&show=2000"
    response = requests.get(url)
    time.sleep(0.5)
    if response.status_code != 200:
        print("Failed to fetch the page.")
        return []

    soup = BeautifulSoup(response.text, 'html.parser')
    titles = []
    for title_div in soup.select(".list-title.mathjax"):
        title = title_div.text.replace("Title:", "").strip()
        titles.append(title)
    abstracts = []

    return titles
```

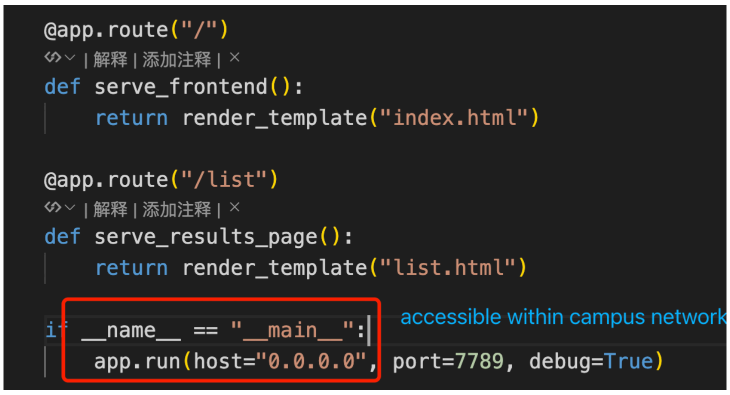
Retrieve

Extremely simple
Do not follow this



```python
def search_titles():
    query = request.args.get("q", "").strip().lower()
    if not query:
        return jsonify([])
    results = [title for title in titles if query in title.lower()]
    return jsonify(results)
```

Website



```python
@app.route("/")
<> | 解释 | 添加注释 | ×
def serve_frontend():
    return render_template("index.html")

@app.route("/list")
<> | 解释 | 添加注释 | ×
def serve_results_page():
    return render_template("list.html")

if __name__ == "__main__":
    app.run(host="0.0.0.0", port=7789, debug=True)
```

accessible within campus network

Try it out
http://101.5.198.232:7789

Thanks!

Q&A