Text Retrieval & Search Engine (CP423A) Course Overview (Winter 2022)

Course Instructor: Stanley (Zhaohui) Liang, PhD

Course Information

- Offered by Dept. of Physics and Computer Science
- Credit: 0.50
- Prerequisite: CP317 Software Engineering
- Co-requisite: CP476 Internet Computing
- Days/Times: Monday & Wednesday, 5:30 PM-6:50 PM
- Time of Learning: 3 hours per week, 36 hours in total
- Term: Winter 2022 (Jan. 5 – Apr. 27)
- Course work: 11 weeks of course lecture, 1 mid-term exam, 1 final exam, 1 group (2 persons) presentation, 1 assignment (short-answer questions and coding a python App)

Course Syllabus

Week (Date)	Topic
2 (1/10 Mon, 1/12 Wed)	Text data understanding
3 (1/17 Mon, 1/19 Wed)	Text data access, text representation and preprocessing techniques
4 (1/24 Mon, 1/26 Wed)	Inverted indices, compression of indices and text
5 (1/31 Mon, 2/2 Wed)	IR Models: Boolean
6 (2/7 Mon, 2/9 Wed)	IR Models: Vector Space
7 (2/14 Mon, 2/16 Wed)	IR Models: Probabilistic Model
8 & 9	reading week and mid-term
10 (3/7 Mon, 3/9 Wed)	TR model evaluation
11 (3/14 Mon, 3/16 Wed)	Relevance feedback
12 (3/21 Mon, 3/23 Wed)	Latent semantic indexing
13 (3/28 Mon, 3/30 Wed)	Web search
14 (4/4 Mon, 4/6 Wed)	Information Filtering
15 – 17	Presentation, assignment due, final exam

Course Resource

- Textbook: ChengXiang Zhai & Sean Massung. Text Data Management and Analysis: Practical Introduction to Information Retrieval and Text Mining
- (https://storage.googleapis.com/pet-detect-239118/text_retrieval/textbook.pdf)
- Python notebooks for exercise
- (https://github.com/StanleyLiangYork/Text_retrieval_search_engine/tree/main/python_code)
- Lecture Notes
- (https://github.com/StanleyLiangYork/Text_retrieval_search_engine/tree/main/lecture_notes_2022W)
- Assignment – will be released after mid-term exam
- (https://github.com/StanleyLiangYork/Text_retrieval_search_engine/tree/main/assignment)

Self Introduction

- Course Instructor: Stanley (Zhaohui) Liang, PhD
- Research Assistant, York University
- Visiting Scientist in machine learning, National Library of Medicine (NLM), NIH, MD, USA
- Machine Learning Developer, Toronto, ON
- Contract Teaching Faculty member, Wilfrid Laurier University, Waterloo, ON
- Contact email:

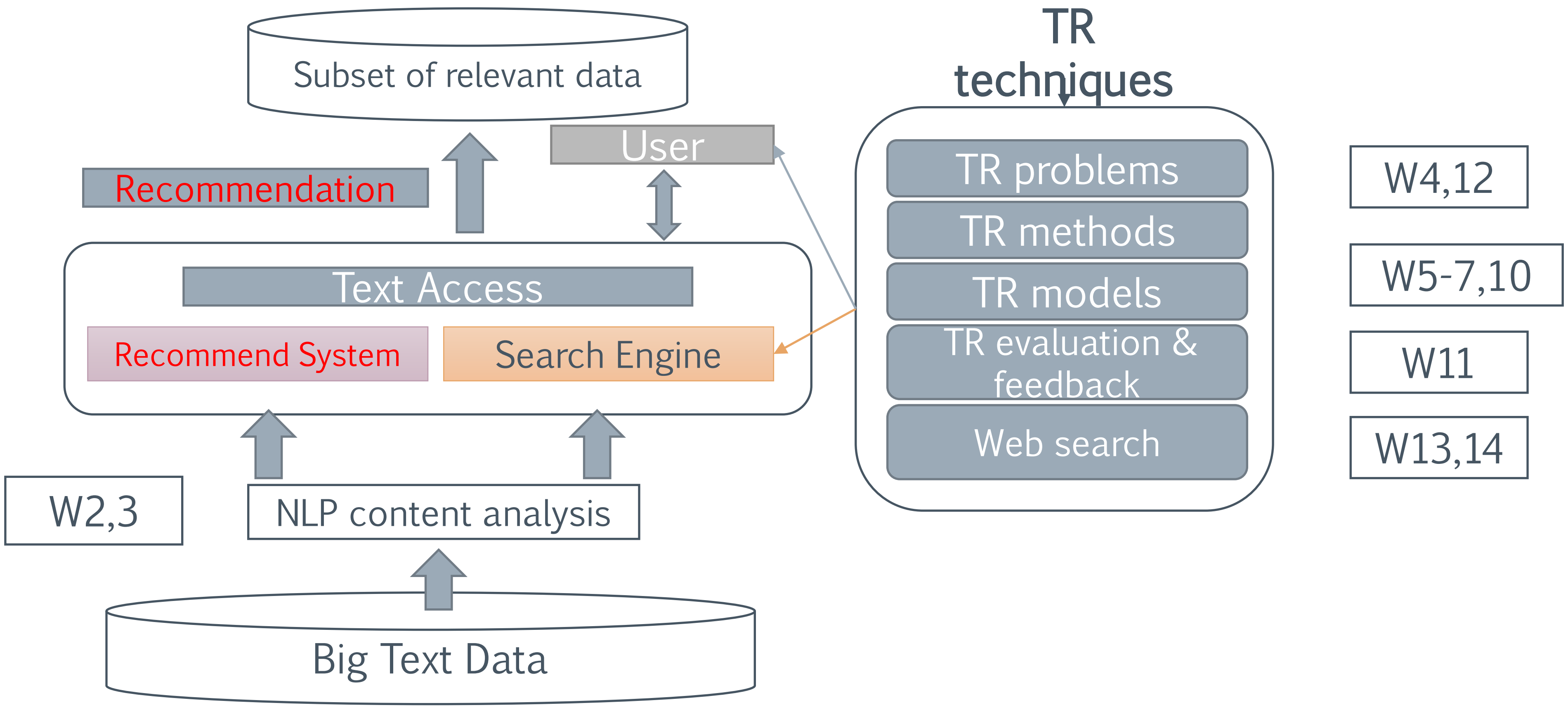
Text Retrieval Overview

-- the roadmap of Text Retrieval (TR)

What is Text Retrieval (TR)

- Having a collection of **text** documents
- Use a query to express the information requirements
- Search engine returns relevant documents picked by the TR system to users
- TR is a sub-task of information retrieval (IR), for IR can retrieve more than texts
- More popular named as "search technology" in the IT industry

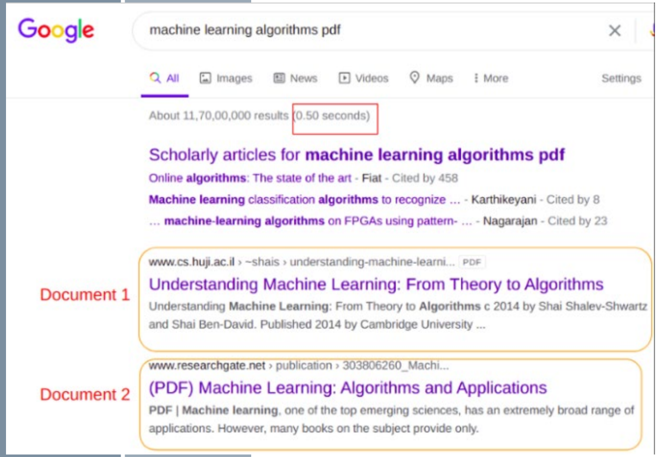
Roadmap of TR system



Some relevant ideals of TR

TR versus SQL database retrieval

- Free text vs structured data
- Ambiguous vs rigorous semantics
- Retrieve relevant docs vs matched records
- TR is an empirical problem, no best method, evaluation by users



About 11,70,00,000 results (0.50 seconds)

Scholarly articles for machine learning algorithms pdf

Online **algorithms**: The state of the art - Fiat - Cited by 458

Machine learning classification **algorithms** to recognize ... - Karthikeyani - Cited by 8

... **machine-learning algorithms** on FPGAs using pattern- ... - Nagarajan - Cited by 23

Document 1

www.cs.huji.ac.il › ~shais › understanding-machine-learni... PDF

Understanding Machine Learning: From Theory to Algorithms

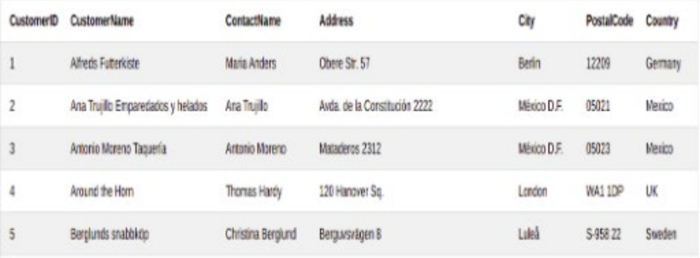
Understanding **Machine Learning**: From Theory to **Algorithms** c 2014 by Shai Shalev-Shwartz and Shai Ben-David. Published 2014 by Cambridge University ...

Document 2

www.researchgate.net › publication › 303806260_Machi...

(PDF) Machine Learning: Algorithms and Applications

PDF | **Machine learning**, one of the top emerging sciences, has an extremely broad range of applications. However, many books on the subject provide only.



CustomerID	CustomerName	ContactName	Address	City	PostalCode	Country
1	Alfreds Futterkiste	Maria Anders	Obere Str. 57	Berlin	12209	Germany
2	Ana Trujillo Emparedados y helados	Ana Trujillo	Avda. de la Constitución 2222	México D.F.	05021	Mexico
3	Antonio Moreno Taquería	Antonio Moreno	Mataderos 2312	México D.F.	05023	Mexico
4	Around the Horn	Thomas Hardy	120 Hanover Sq.	London	WA1 1DP	UK
5	Berglunds snabbköp	Christina Berglund	Berguvsvägen 8	Luleå	S-958 22	Sweden

```
SELECT CustomerName FROM Customers
WHERE city is "London"
```

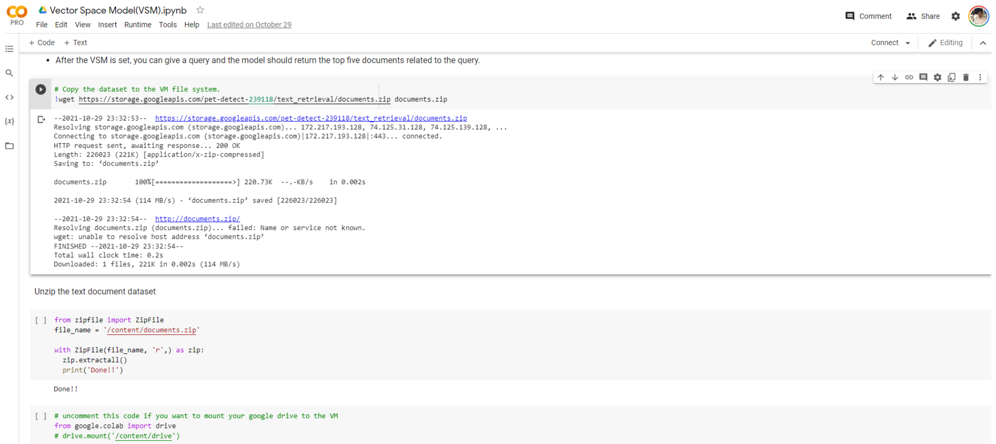
Some TR formulations

- Vocabulary: $V = \{w_1, w_2, w_3, \dots, w_N\}$, all words in the doc collection
- Query: q, $q_i \in V$
- Document: $d_i = d_{i1}, \dots, d_{im_j}, d_{ij} \in V$
- Collection: $C = \{d_{1,} \dots, d_M\},$
- Word Count: C(w, d) counting frequency of word *w* in *d*
- Set of relevant documents: $R(q) \subseteq C$
- TR task: compute $R'(q) \rightarrow R(q)$
- Two strategies:
  - Document selection: $R'(q) = \{d \in C | f(d, q) = 1\}, where\ f(d, q) \in \{0,1\}$
  - Document ranking: $R'(q) = \{d \in C | f(d, q) > \theta\}, where\ f(d, q) \in \Re\ is\ a\ relevance\ measure\ function$
  - Ranking is often preferred

Python as the coding tool

MOST EXERCISE WILL BE ONLINE

- Google Colab is a free online Jupyter Notebook Platform for Python programming
- REPL (Read, Evaluation, Print, Loop) style coding
- Free use (need google account)
- Good pre-installed packages
- Can directly use bash command
- Easy access of GPU
- https://colab.research.google.com



CO PRO
Vector Space Model(VSM).ipynb
File Edit View Insert Runtime Tools Help Last edited on October 29
Comment Share
+ Code + Text
Connect
Editing

- After the VSM is set, you can give a query and the model should return the top five documents related to the query.

```
# Copy the dataset to the VM file system.
!wget https://storage.googleapis.com/pet-detect-239118/text_retrieval/documents.zip documents.zip
```

```
--2021-10-29 23:32:53--  https://storage.googleapis.com/pet-detect-239118/text_retrieval/documents.zip
Resolving storage.googleapis.com (storage.googleapis.com)... 172.217.193.128, 74.125.31.128, 74.125.139.128, ...
Connecting to storage.googleapis.com (storage.googleapis.com)|172.217.193.128|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 226023 (221K) [application/x-zip-compressed]
Saving to: ‘documents.zip’

documents.zip       100%[===================>] 220.73K  --.-KB/s    in 0.002s

2021-10-29 23:32:54 (114 MB/s) - ‘documents.zip’ saved [226023/226023]

--2021-10-29 23:32:54--  http://documents.zip/
Resolving documents.zip (documents.zip)... failed: Name or service not known.
wget: unable to resolve host address ‘documents.zip’
FINISHED --2021-10-29 23:32:54--
Total wall clock time: 0.2s
Downloaded: 1 files, 221K in 0.002s (114 MB/s)
```

Unzip the text document dataset

```
[ ] from zipfile import ZipFile
    file_name = '/content/documents.zip'

    with ZipFile(file_name, 'r',) as zip:
      zip.extractall()
      print('Done!!')
```

```
Done!!
```

```
[ ] # uncomment this code if you want to mount your google drive to the VM
    from google.colab import drive
    # drive.mount('/content/drive')
```

ASSIGNMENT NEED OFFLINE

- Standard Python interpreter
- https://www.python.org/
- Use python 3.7 or 3.8
- Anaconda distribution (individual edition)
- https://www.anaconda.com/products/individual
- Pycharm as IDE (cooperate with anaconda)
- https://www.jetbrains.com/pycharm/
- https://docs.anaconda.com/anaconda/user-guide/tasks/pycharm/

Coding Lab and course exercises

- The implementation examples of the relevant TR algorithms and models are available in the course GitHub
- https://github.com/StanleyLiangYork/Text_retrieval_search_engine
- In the "python_code" folder
- Download the ZIP repository, then you can upload the python notebook (.ipynb) to Colab for exercise
- Use Colab notebook as code test /debug environment
- If you are not familiar with python, run the python tutorial notebook (Python_coding_tutorial.ipynb) at first

Course Evaluation

- Paper presentation – 10%
- Midterm exam – 30%
- Written assignment – 20%
- Final exam – 40%
- Attendance – 5% extra

Paper presentation

- 10% for your overall grade
- 2 students in a group
- Select 1 paper from the list with 35 papers
- Prepare 8 PowerPoint slides to present the main idea in 6 minutes
- Evaluation metrics
  - Completeness 3%
  - Clearness 3%
  - Time management 2%   ±1 min
  - Involvement 2%

Assignment

- 20% for your overall grade
- Will be released on Feb 21 (reading week)
- 5 questions – 50%
  - 50% of the assignment marks, 10% each
  - The answer should be 200 – 400 words, including simple formula and computation
- 1 programming task (python) – 50%
  - Download the python template from Github
  - Implement the missing code
  - Will be tested with real data

Midterm and Final Exam

- Midterm: 30% for the overall grade
- Final: 40% for the overall grade
- Both will be in 60 mins
- 6 True/False questions (5 points each)
- 6 multiple choice questions (5 points each)
- 4 short answer questions (10 points each)
- Cues for success
  - Good time management
  - Try to answer all questions

Attendance

- Attendance to lectures and presentations is mandatory
- Will randomly take attendance
- 2% will be deducted for one missing attendance, up to 5%
- Punctuality is a good habit
- Try to run all exercise notebook after each lecture to get your hand wet to the relevant issues
- The evaluation will focus on the completeness and involvement