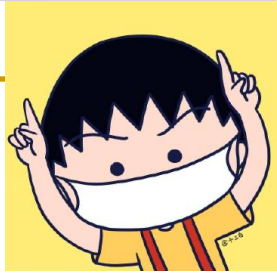
清華大學 Tsinghua University

请同学们在教室内佩戴口罩



Welcome to the class of
Web Information Retrieval !

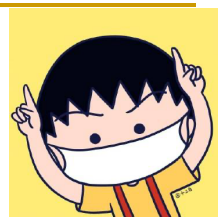
13:30-16:05, Tuesday 6B109 + Zhumu

* Please use your **Real Full Name** but not nickname on the online platform for our class attendances verification and record.

Min ZHANG （张敏）

z-m@tsinghua.edu.cn

During online course period

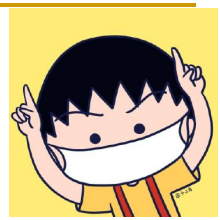


- **Use Zhumu as our online platform.**

  https://zhumu.me/

- Meeting room address will be **shared on the WeChat group** before the course.
- In the case that Zhumu is crashed, we will use WeMeet (腾讯会议 https://meeting.qq.com/) or "RainClassroom for THU" (荷塘雨课堂, via WeChat or https://pro.yuketang.cn) as the backups.

During online course period

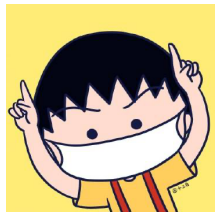


- Time: **13:30 - 16:05, Tues**. (with two breaks)
- QA: at WeChat Group (online), or Open Hour (offline)

- Homework assignment and submit:
  - As usual via **THU Web Learning** 网络学堂

    http://learn.tsinghua.edu.cn
- Slides will be shared at Web Learning platform after class

清華大學 Tsinghua University | 网络学堂 WEB LEARNING



Course basic information

Instructor

Instructor: Min ZHANG （张敏）

- Tenured Associate Professor
- IR group, DCST, Tsinghua Uni.; AI institute, THU.
- Office: FIT 1-506A Tel: 62798279
- Email: z-m@tsinghua.edu.cn
- http://www.thuir.cn/group/~mzhang

TA: WANG Zhen（王振）, WANG Chenyang (王晨阳）

- Email: wang-zhe18@mails.Tsinghua.edu.cn
  thuwangcy@gmail.com
- Office: FIT 1-506

A little bit about Min (1)

- **Experiences and Services**
  - Bachelor: 1999, PhD: 2003, in DCST, THU
  - Now Associate Professor in THUIR group, CST Dept.
  - Visited NUS, DFKI, Kyoto Uni., City Uni. of HK, MSRA
  - ACM SIGIR EC member, ACM TOIS Editor-in-Chief
- **Research Interests**
  - Information Retrieval & Recommendation, user behavior analysis, machine learning
- **Achievements and Awards**
  - Top conferences/journals publications, ~ 6000 citations, H-index 40 (Feb. 2021) SIGIR, IJCAI, WWW, WSDM, PNAS, TOIS, JIR, …
  - Multiple Top Performances in TREC and NTCIR since 2002.
  - **One of the top ranked World-wide IR researchers**

A little bit about Min (2)

CSRankings: Computer Science Rankings

CSRankings is a metrics-based ranking of top computer science institutions around the world. **Click on a triangle** (▶) to expand areas or institutions. **Click on a name** to go to a faculty member's home page. **Click on a pie** (the after a name or institution) to see their publication profile as a pie chart. **Click on a Google Scholar icon** () to see publications, and **click on the DBLP logo** () to go to a DBLP entry.

Applying to grad school? Read this first.

Rank institutions in the world by publications from 2011 to 2021

All Areas [off | on]

AI [off | on]

- ▶ Artificial intelligence
- ▶ Computer vision
- ▶ Machine learning & data mining
- ▶ Natural language processing
- ▼ The Web & information retrieval

ACM SIGIR

SIGIR ☑
WWW ☑

#	Institution	Count	Faculty
1	▼ Tsinghua University	39.6	35
	Faculty	*# Pubs*	*Adj. #*
	Min Zhang 0006 WEB+IR	41	7.0
	Shaoping Ma WEB+IR	41	7.1
	Yiqun Liu 0001 WEB+IR	39	6.5
	Yong Li 0008 HCI,WEB+IR	13	2.5
	Jie Tang 0001 ML,AI	8	1.9
	Minlie Huang NLP,AI	7	1.3

#	Institution	Count	Faculty
1	▶ Tsinghua University	39.6	35
2	▶ Cornell University	23.9	20
3	▶ Univ. of Illinois at Urbana-Champaign	20.7	21
4	▶ Chinese Academy of Sciences	16.3	15
5	▶ University of Michigan	16.2	17
6	▶ National University of Singapore	15.8	14
6	▶ Zhejiang University	15.8	24
8	▶ University of Amsterdam	15.1	5
9	▶ Stanford University	14.9	14
10	▶ Carnegie Mellon University	13.6	21

A little bit about Min (3)

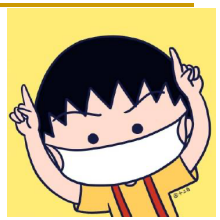
Course Experiences

- **Introduction to Machine Learning (2003 – Now)**
  - **The first** Machine Learning course in CST Dept., Tsinghua Univ.
- Advanced Topics in Information Retrieval (2008- Now)
  - Graduate Students
- Web and IR (for Advanced Computing Program, 2011 - Now)
  - Credit Course in THU-CMU Double Master Degree Program in CS

Awards

- Excellent Young Faculty Teaching Award, Tsinghua University, 2008
- Outstanding Teacher of Computer Science in Chinese Universities, China, 2018.
- Excellent Online Teaching Award, Tsinghua University, 2020

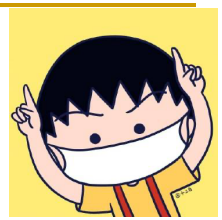
Course Arrangements



- The first 13 weeks
  - Lectures by the teacher on selected topics
  - With "Tea Time" discussions on the news/progresses on Web and IR industry/research (by one student)
    - Send me the email **by Mon. 09:00 (GMT+8)** to apply a tea time show on Tue. course.

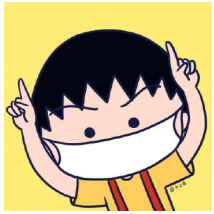
    Examples:
    - WolframAlpha
    - Adidas miCoach Smart Ball
    - Dark net search
    - Man vs. Machine

Course Arrangements



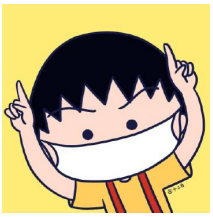
- The last 3 weeks -- A workshop
  - Seminar presented by the students
  - ~10 mins' talk + ~5 mins' QA (tentative, depending on #students)
  - Awards and celebration to "the Best Project".

Part I-1. Lectures on Web IR Fundamentals (subject to modifications)



Week	Date	Topic	Content
1	2.23	Introduction	Course intro, IR history, General procedure
2	3.2	Crawler and Indexing	Key tech: crawler, Index
3	3.9	Content analyses	Content-based Ranking Models – term weighting. Boolean, VSM, probability(classical), LM
4	3.16	Link Analysis Behavior Analysis	Link Analysis: Node degrees, HITS, PageRank, TrustRank; User Click models and satisfaction.
5	3.23	Evaluation	Evaluation: methodology; metrics (pre, recall, MAP, MRR, NDCG), Consistency. Rethinking of evaluation.

Homework assignments on Part I-1 (Candidates)



- 1. Crawler and Indexer
  - Writing a SE result crawler
- 2-3. Evaluation and Consistencies
  - Data understanding by manual annotation
  - Evaluation
  - Consistency
- 4. Link Analysis: PageRank
  - PageRank Calculation and optimization

Part I-2. Lectures on Advanced Topics (subject to modifications)

Week	Date	Topic	Content
6	3.30	Challenges	Anti-spam, User behavior, Scale, Multi-source fusion, UI
7	4.6	Social IR	Social network & Human computation, Crowd computing
8	4.13	Visual IR	Visual IR
9	4.20	Recommendation (1)	Tasks, evaluations, Techniques and algorithms(1)
10	4.27	Recommendation (2)	Techniques and algorithms (2)
11	5.4	National Holiday	
12	5.11	Users Modeling	User Modeling and satisfaction
13	5.18	Challenges in Recom.	Cold start, Explainability, Fairness…

Part II. Project and Course workshop

Week	Date	Topic	Content
14-16	5.25, 6.1, 6.8	Course Workshop	SE prototype design & implementation

- II-1. Everybody is asked to Design/build a prototype Search Engine
  - You can use general IR toolkits to build your search engine.
  - Both general and domain-specific SEs are acceptable.
    - A specific one with some interesting functions might be more attractive to your audiences

Part II. Project and Course workshop

- Two options:
  - 1. Design Only: Design the SE system with detail info.
    - Models, equations, mathematical foundations, analysis on the characteristics and difficulties……
    - Easier but with less scores on general
      - Scoring from 60-85
      - Longer QA time will be issued for *Design Only* talk in the workshop.
      - Challenging on different aspects of the design by the teacher, TA and the students.

Part II. Project and Course workshop

- Two options (cont.):
  - 2. Design & Implementation: Design and then implement your prototype search engine
    - Harder one with generally higher scores
    - Scoring from 75-100, or 0.
    - A demo system should be released online (accessible in THU net) at least one week before your talk on the workshop.
    - All the others are asked to test the demo in advance.
    - Final score including both pre-test and the live show.

Part II. Project and Course workshop

- II-2. Give a presentation (and show your Demo for Option 2) on the course workshop

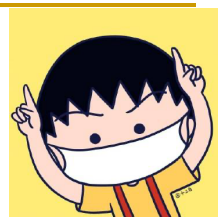
Option 1 (design only):

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

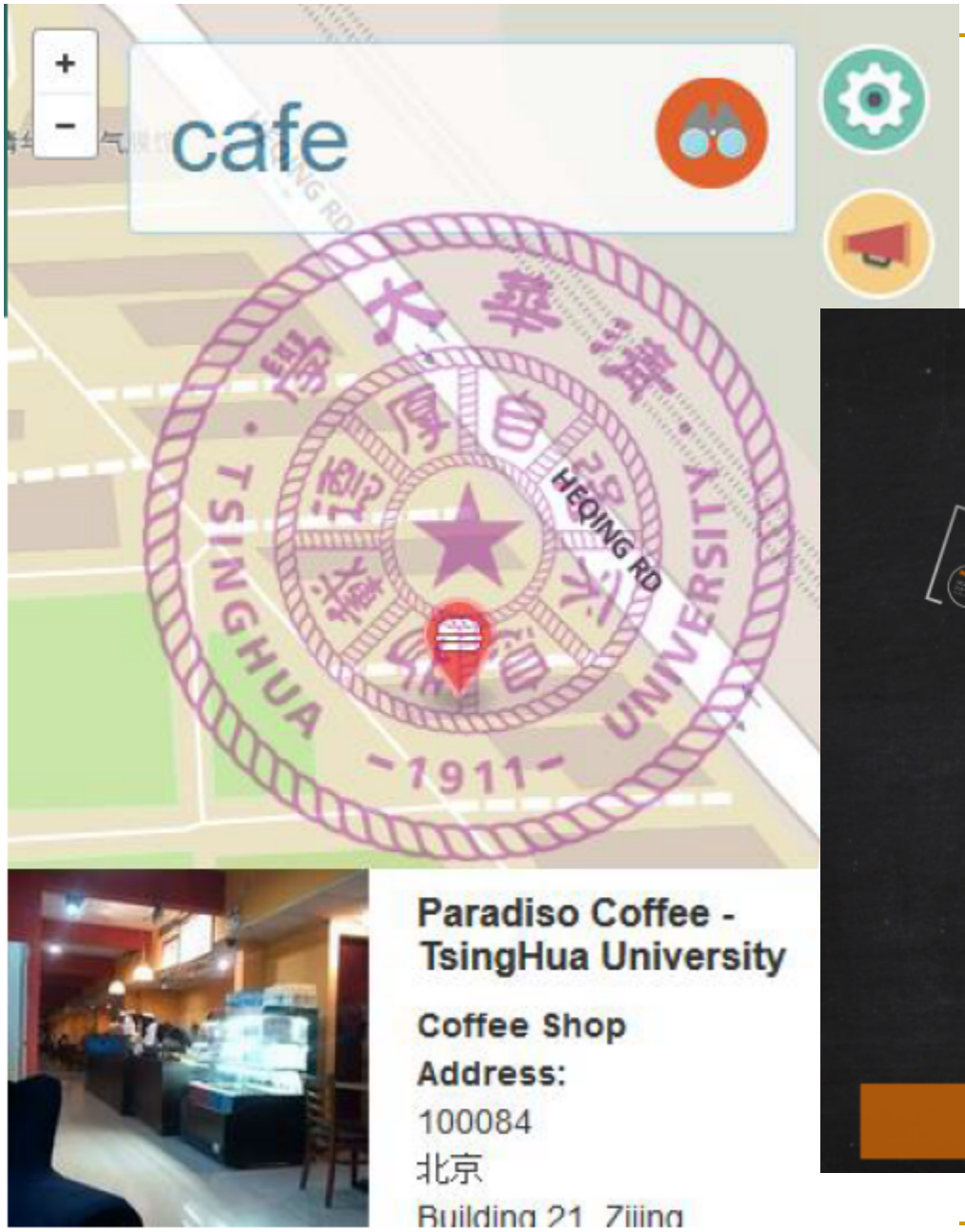
Option 2 (design and implementation):

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

- III-3. Submit a paper (~6 pages) on your work.




POKÉMON
Video games search engine
Search mode : Title only
Lam Corporation - June 2013 - Information - Suggestion



Paradiso Coffee - TsingHua University

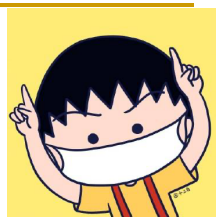
Coffee Shop

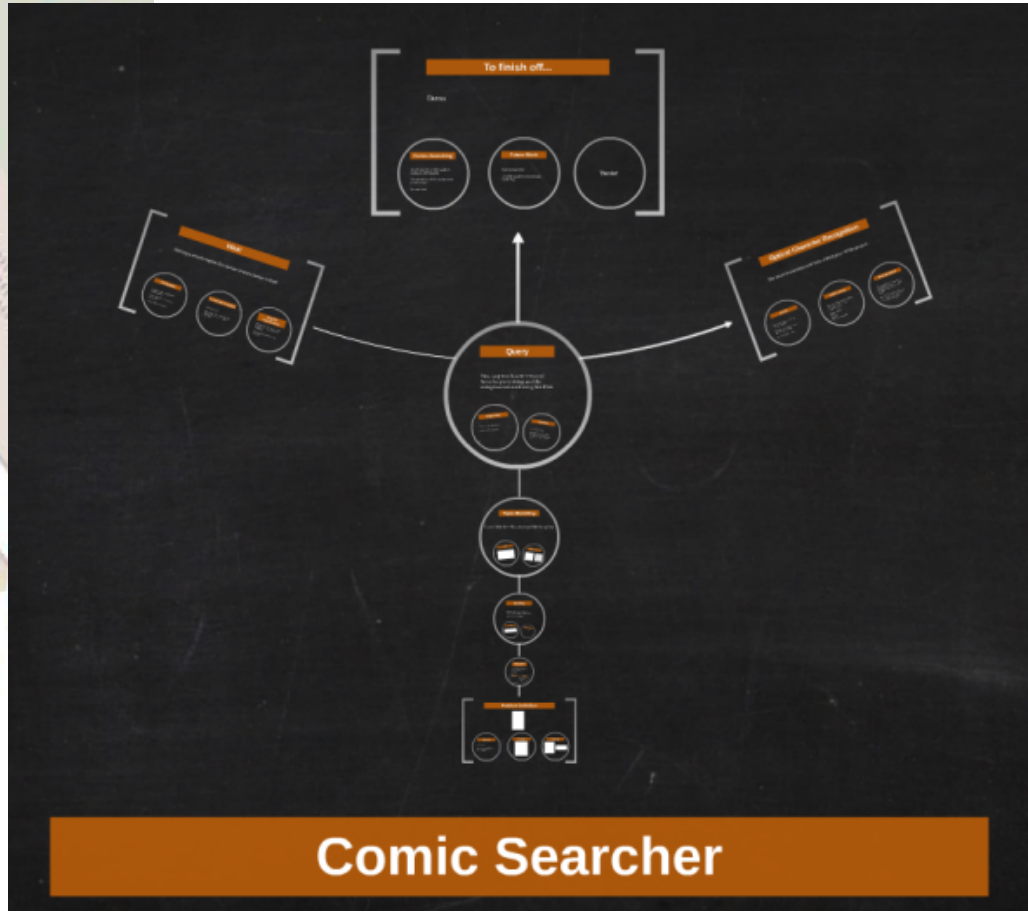
Address:

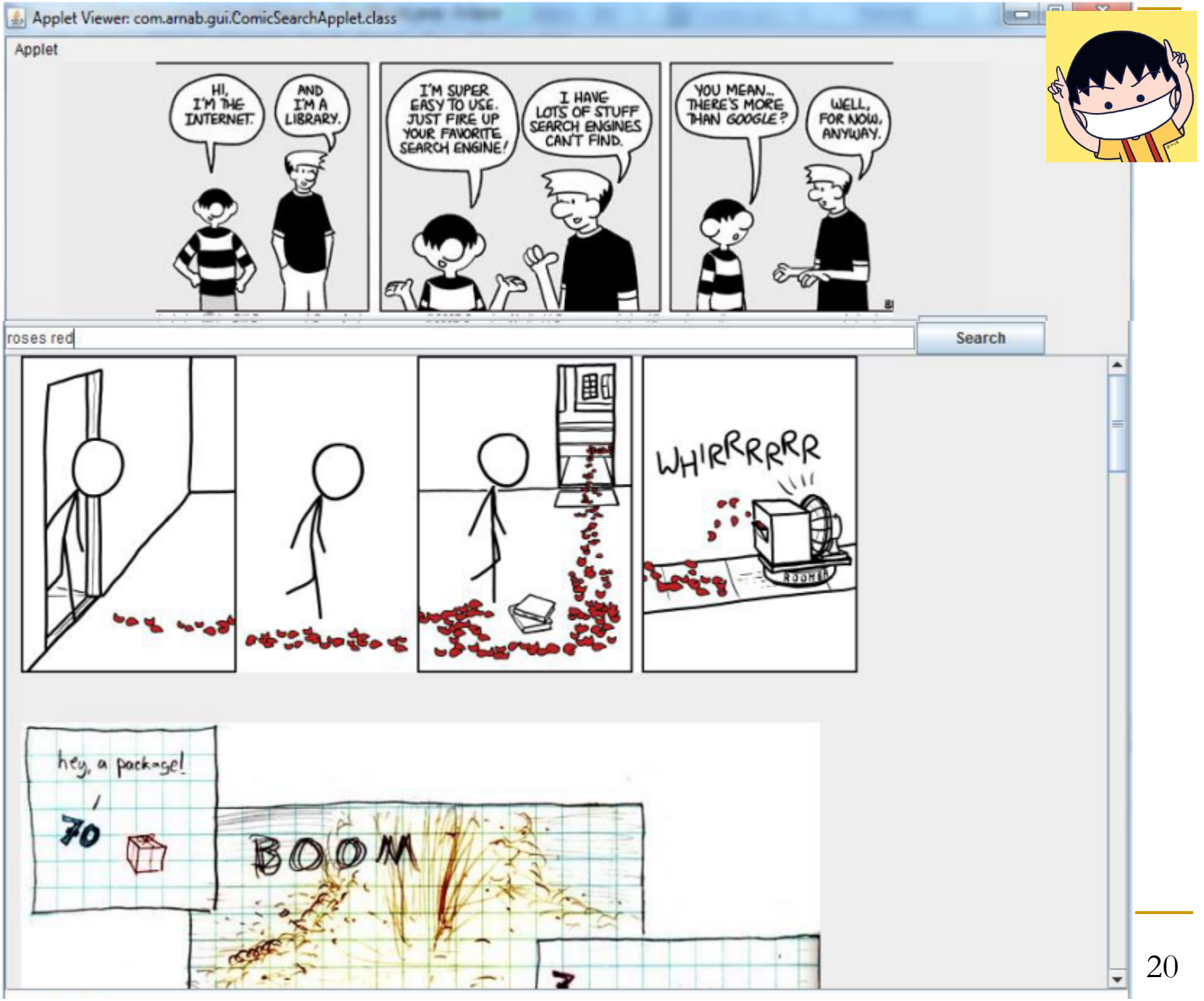
100084

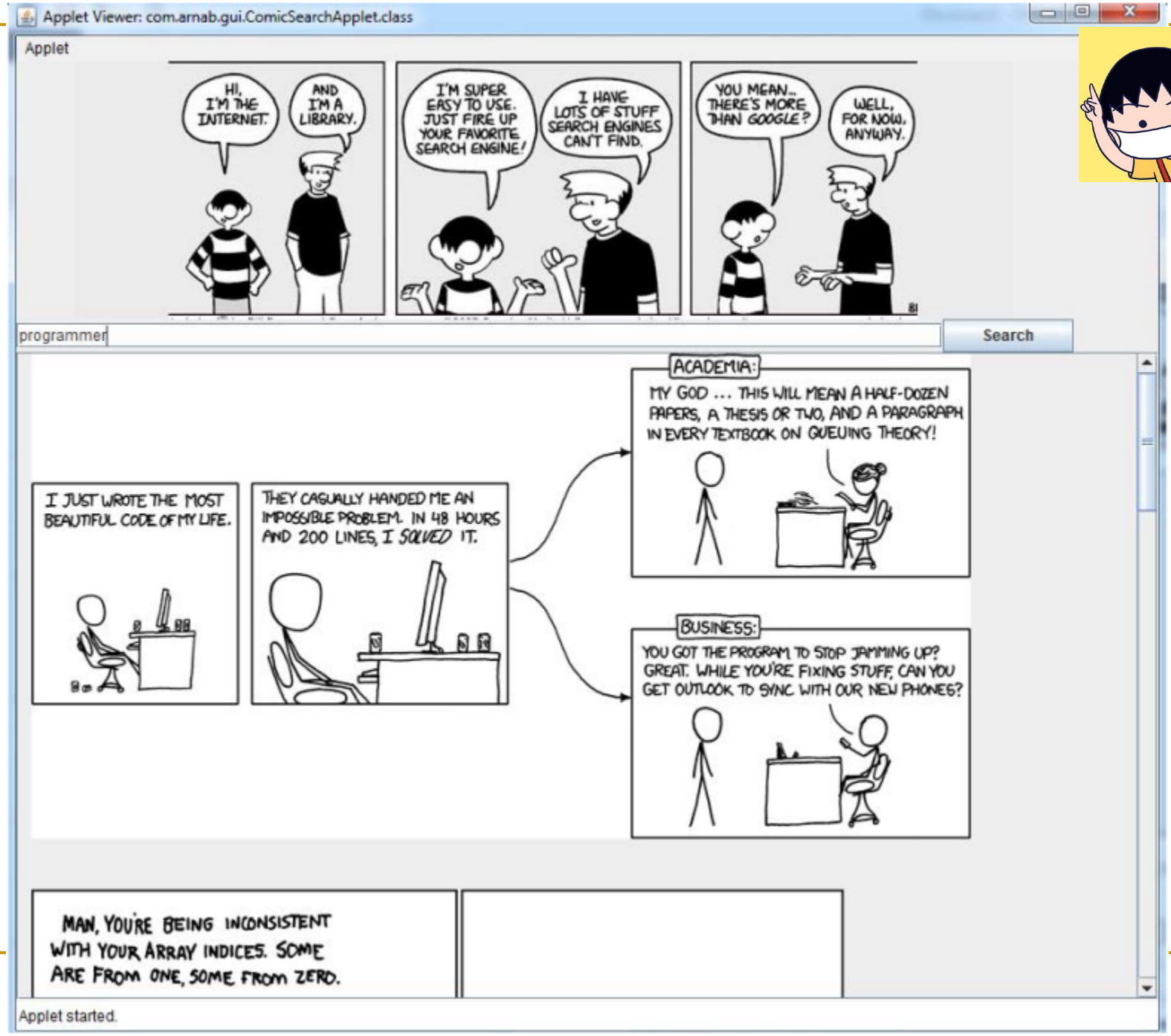
北京

Building 21, Zijing

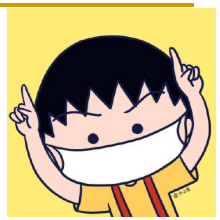





Applet Viewer: com.arnab.gui.ComicSearchApplet.class
Applet
HI, I'M THE INTERNET.
AND I'M A LIBRARY.
I'M SUPER EASY TO USE. JUST FIRE UP YOUR FAVORITE SEARCH ENGINE!
I HAVE LOTS OF STUFF SEARCH ENGINES CAN'T FIND.
YOU MEAN... THERE'S MORE THAN GOOGLE?
WELL, FOR NOW, ANYWAY.
roses red
Search
WHIRRRRRR
ROOMBA
hey, a package!
BOOM!


Applet Viewer: com.arnab.gui.ComicSearchApplet.class
Applet
HI, I'M THE INTERNET.
AND I'M A LIBRARY.
I'M SUPER EASY TO USE. JUST FIRE UP YOUR FAVORITE SEARCH ENGINE!
I HAVE LOTS OF STUFF SEARCH ENGINES CAN'T FIND.
YOU MEAN... THERE'S MORE THAN GOOGLE?
WELL, FOR NOW, ANYWAY.
programmer
Search
I JUST WROTE THE MOST BEAUTIFUL CODE OF MY LIFE.
THEY CASUALLY HANDED ME AN IMPOSSIBLE PROBLEM. IN 48 HOURS AND 200 LINES, I SOLVED IT.
ACADEMIA:
MY GOD ... THIS WILL MEAN A HALF-DOZEN PAPERS, A THESIS OR TWO, AND A PARAGRAPH IN EVERY TEXTBOOK ON QUEUING THEORY!
BUSINESS:
YOU GOT THE PROGRAM TO STOP JAMMING UP? GREAT. WHILE YOU'RE FIXING STUFF, CAN YOU GET OUTLOOK TO SYNC WITH OUR NEW PHONES?
MAN, YOU'RE BEING INCONSISTENT WITH YOUR ARRAY INDICES. SOME ARE FROM ONE, SOME FROM ZERO.
Applet started.

Supports fuzzy search (maximum edit distance of 3)



LaTeX Symbol Search

esplon Search

Mouse over equations to display their TeX commands

Symbol in TeX	TeX Command
E, ϵ and ε	`E`, `\epsilon` and `\vareps`

Can search content

LaTeX Symbol Search

the set of all real numbers Search

Mouse over equations to display their TeX commands

Symbol in TeX	TeX Command	Name / Read as / Category	Explanation
$\mathbb{R}$ $\mathbf{R}$	`\mathbb{R}` `\mathbf{R}`	real numbers R; the (set of) real numbers; the reals numbers	$\mathbb{R}$ means the set of real numbers.

theamazingemojisearcher.herokuapp.com/?q=happy

theamazingemojisearcher.herokuapp.com/emoji/3

The Amazing Emoji Searcher

Home / 😂 face with tears of joy

Face with tears of joy

Description

😂 A laughing emoji which at small sizes is often mistaken for being tears of sadness. In fact , this emoji is laughing so much that it is crying tears of joy. Tears are coming from both eyes, not due to sadness, but overwhelmed with happiness or laughter.

Also known as

- 😂 Laughing Emoji
- 😂 Laughing Crying Emoji
- 😂 Happy Tears Emoji
- 😂 Laughing Tears Emoji
- 😂 LOL Emoji

Related to

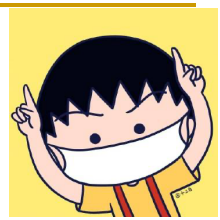
- 😄 Smiling Face With Open Mouth And Smiling Eyes
- 😹 Cat Face With Tears Of Joy
- 😢 Crying Face
- 😭 Loudly Crying Face
- 😆 Smiling Face With Open Mouth And Tightly-Closed Eyes

Different versions

iPhone	Android	Twitter	Gmail	Windows	Black and white
😂	😂	😂	😂	😂	😂

Made with love by Richard Luong. Available on Github.

素问　关于素问　联系我们　用微博登录



Tradition Chinese Medicine Doctors Search

素 問

专业、安全、免费的医疗信息服务

搜索病名, 例如：胃

Article Generator

Article Generator

Document | PDF Attachments | Videos

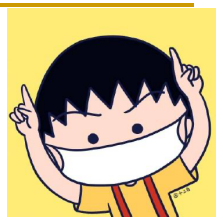
Keywords	Useful Links
learning	en.wikipedia.org
machine	azure.microsoft.com
data	online.stanford.edu
machine learning	www.springer.com
method	whatis.techtarget.com
algorithm	research.microsoft.com
model	ocw.mit.edu
program	azure.microsoft.com

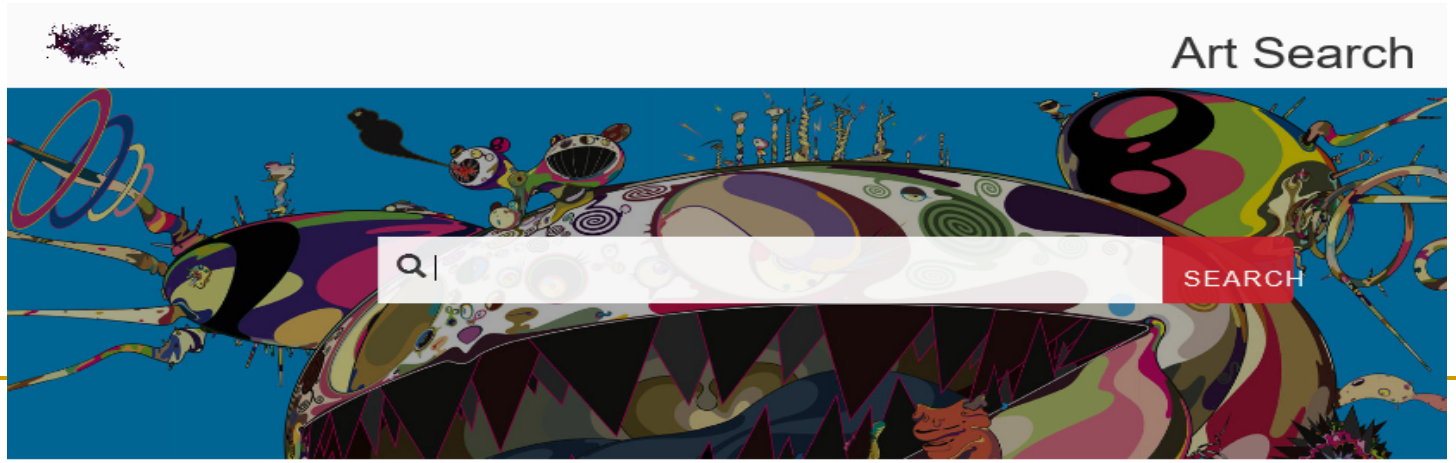
Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence. Machine learning explores the construction and study of algorithms that can learn from and make predictions on data. Machine learning is closely related to and often overlaps with computational statistics; a discipline that also specializes in prediction-making. Machine learning is employed in a range of computing tasks where designing and programming explicit, rule-based algorithms is infeasible.

Machine learning is sometimes conflated with data mining, although that focuses more on exploratory data analysis. When employed in industrial contexts, machine learning methods may be referred to as predictive analytics or predictive modelling. In 1959, Arthur Samuel defined machine learning as a "Field of study that gives computers the ability to learn without being explicitly programmed". Already in the early days of AI as an academic discipline, some researchers were interested in having machines learn from data. Machine learning and data mining often employ the same methods and overlap significantly. The two areas overlap in many ways: data mining uses many machine learning methods, but often with a slightly different goal in mind. On the other hand, machine learning also employs data mining methods as "unsupervised learning"

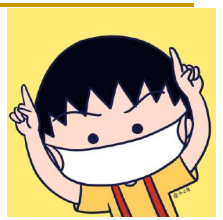
Back | PDF | Clear Highlights

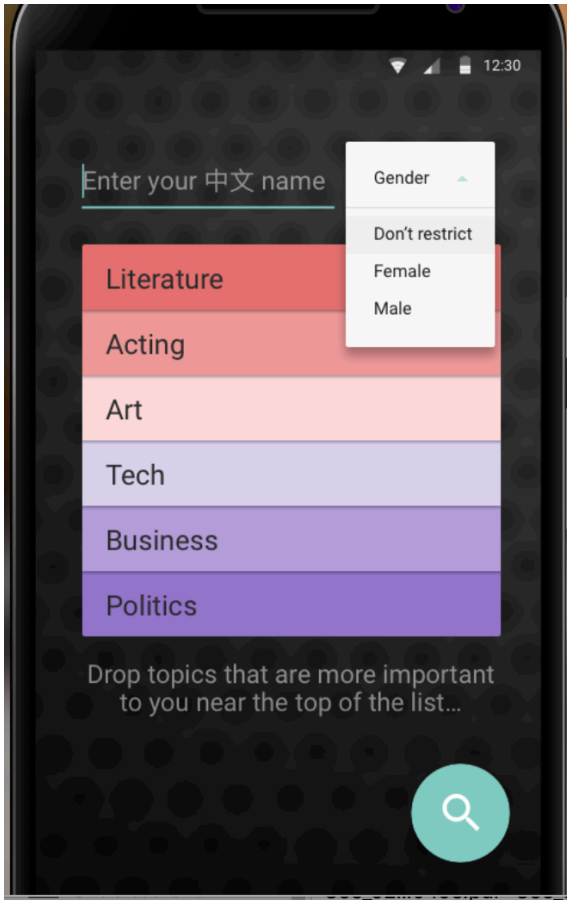
Web Art Search

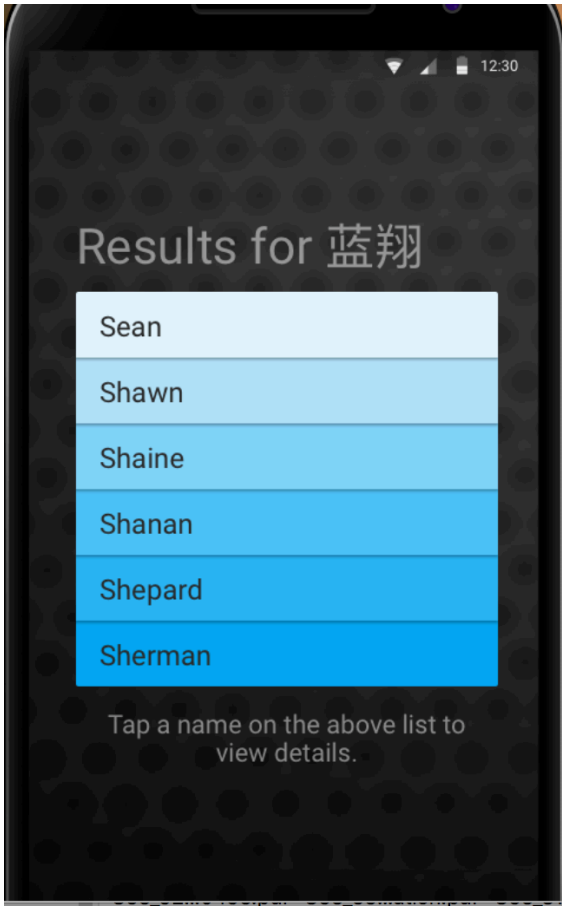


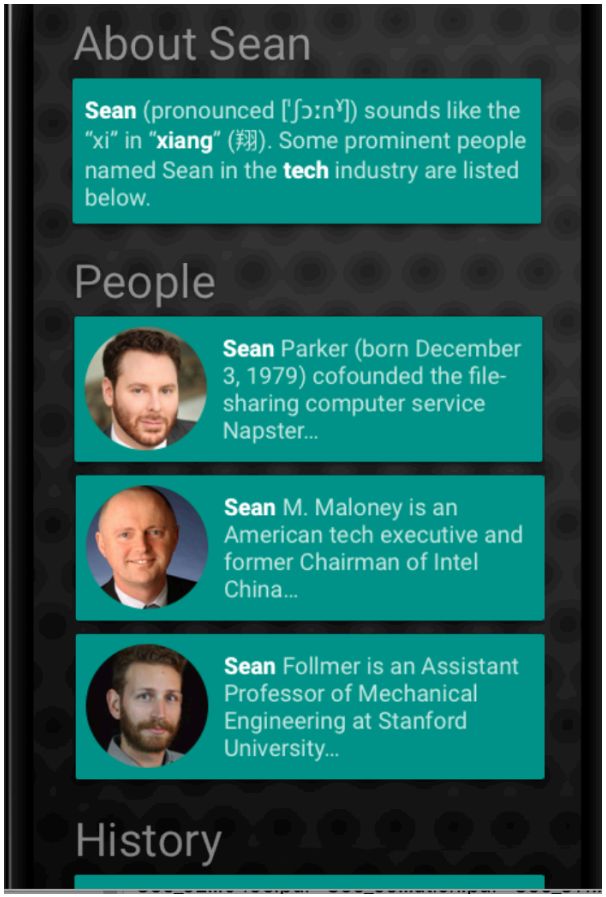


Bilingual Name Search Engine





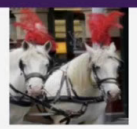




Results

Common Tags : Person

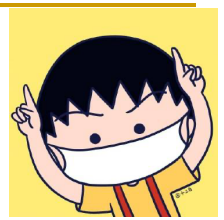


Results

Common Tags :


Reverse Image Search Engine



Reverse Image Search Engine

清華大學 Tsinghua University

HOME ABOUT INSTRUCTION SAMPLE

Papers M

rec sys

Advanced

Google It

Unread Only

Commit Index

Tags

Total	(101)
Rec	(4)
Deep	(1)
Ensemble	(1)

New tag +

Time

Any Time
Within 1 Year
Within 2 Years
Within 3 Years
Within 5 Years
Within 10 Years

Sort Order

By Default

RecSys Challenge 2016:Job Recommendations - 2 Cites

The 2016 ACM Recommender Systems Challenge focused on the problem of job recommendations. Given a large dataset from XING that consisted of anonymized user...

Abel, Fabian and Kohlsdorf, Daniel and Larson, Martha

(2016) ACM Conference on Recommender Systems

Rec Deep Manage tags Mark read

An ensemble method for job recommender systems - 1 Cites Read

Abstract In this paper, we present an ensemble method for job recommendation to ACM RecSys Challenge 2016. Given a user, the goal of a job recommendatio...

Zhang, Chenrui and Cheng, Xueqi

(2016) Recommender Systems Challenge

Rec Ensemble Manage tags Mark read

MoviExplain:a recommender system with explanations - 46 Cites

Providing justification to a recommendation gives credibility to a recommender system. Some recommender systems (Amazon.com etc.) try to explain their recommendations, in an effort to regain customer ...

Symeonidis, Panagiotis and Nanopoulos, Alexandros and Manolopoulos, Yannis

(2009) Conference on Recommender Systems

Rec Manage tags Mark read

A preliminary study on a recommender system for the job recommendation challenge - 2 Cites Read

A preliminary study on a recommender system for the job recommendation challengecollaborative filteringjob recommendation challengetop-n recommendationAbstract In this paper we present our method used...

Polato, Mirko and Aiolli, Fabio

Papers Mine Master anagement

研究生必备

你每天都会用到的搜索引擎

多类别管理

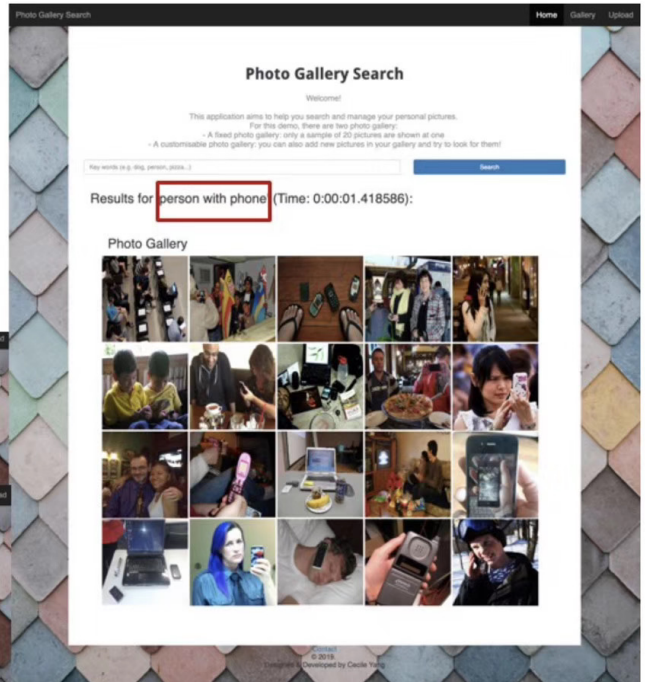
一篇文章多个tag，筛选管理更轻松

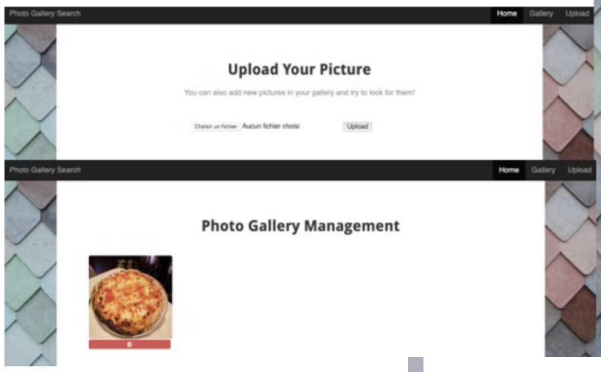
界面清新

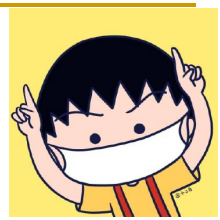
用户友好的界面，简洁大方

PHOTO GALLERY SEARCH

Cecile Yang

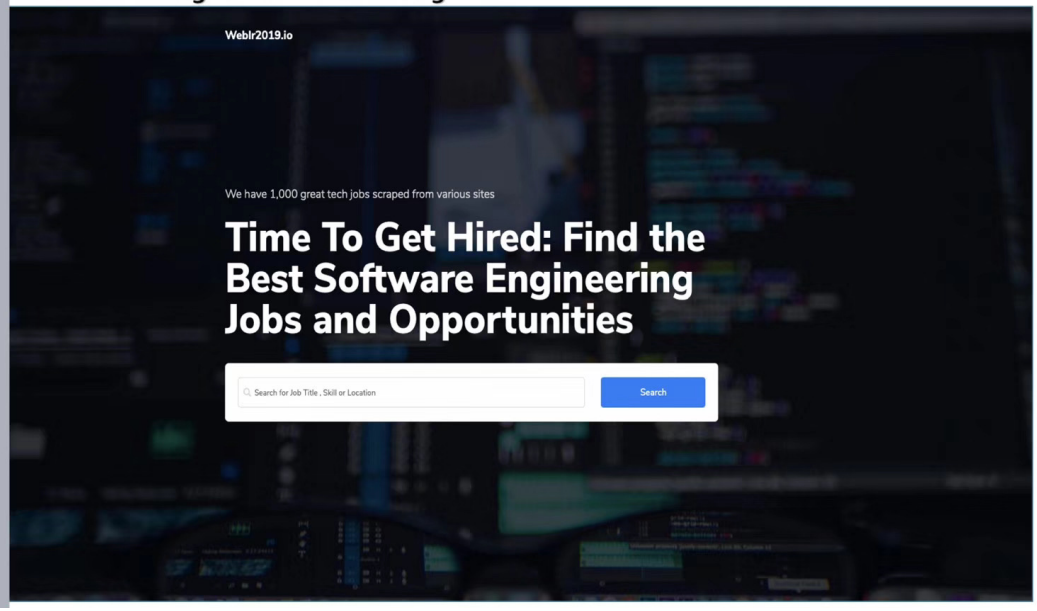


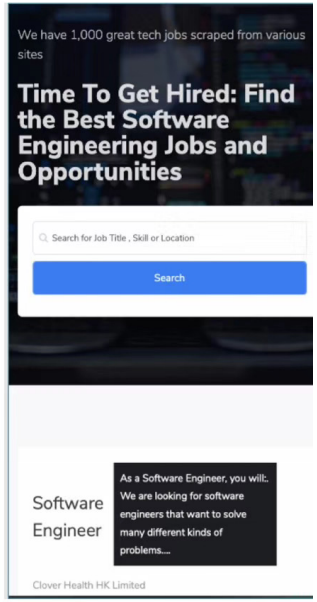




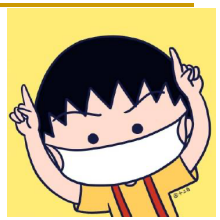
Niche Job Aggregator

Jobs Search Engine for Software Engineers





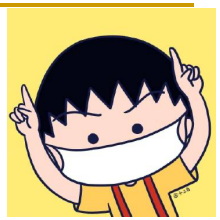
Evaluation (Subject to modifications)



- Homework (40%~45%)
  - Around 4 assignments.
- Workshop (40%~45%), evaluated by
  - The other students (half)
  - The teacher and TA (half)
- Paper (~15%)
- Bonus:
  - The best project on the workshop
  - Tea time presentations on weekly classes
  - Active question asking on Workshop

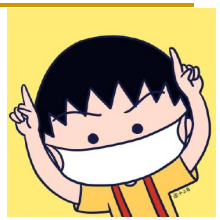
Active thinking and discussions are highly encouraged !

What you will learn

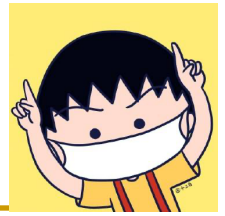


- Overview and In-sight of Search (and Recommendation)Technologies
  - Beat 90% general SE users (but still not an expert, to be honest).
- Obtaining knowledge for (web) information processing and big data analysis (not only SE and IR), Such as
  - Evaluation strategies
    - Consistencies, correlation, and metrics (precision, recall, AUC, NDCG,…)
  - Classical and state-of-arts models
  - Better understanding on challenges
- Getting to know about new techniques in Web information services
  - Search, Recommender Systems, User modeling, Social analysis, Visual processing …
- Be used to always keep an eye on the new trends of Internet and AI

References



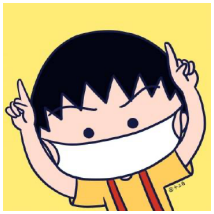
- We're not having an official textbook
  - There isn't one with good coverage of all & only the topics we'll discuss
  - A changing field, advanced topics
- A list of references:
  - Books
    - W. Bruce Croft, Donald Metzler, Trevor Strohman, **Search Engine: information retrieval in practice**
    - Christopher D. Manning, Prabhakar Raghavan ,Hinrich Schütze, **Introduction to information retrieval**
    - I. Witten, A. Moffat, and T. Bell, **Managing Gigabytes**
  - Proceedings of Conferences
    - SIGIR, WWW, IJCAI, WSDM, CIKM, TREC, NTCIR ...
  - Very important: Web resources, Search engines



What's IR?

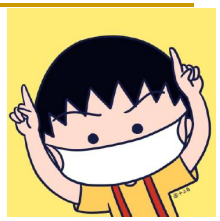
Figure Copyright by TREC

What is Information Retrieval (IR)?



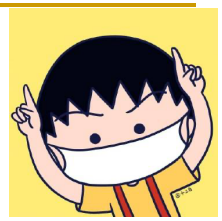
- Narrow-sense:
  - IR = Search Engine Technologies (i.e. IR =
    - Google, Yahoo, Bing, Ask, Baidu, Sogou, …
    - Library info search, enterprise search, in-site search, destop search…
    - PicSearch, Greplin, Blekko, SkyScanner, KooXoo, Qunar, …

What's IR? (cont.)



- Broad-sense: IR ~ **Information Management**
  - General problem: how to manage information?
  - How to **find** useful information? (retrieval & recommendation)
    - Beyond search engine:
    - e.g. in news feed, movie, travel, e-commerce, financial… scenarios
    - e.g. in social media platform, e.g. Twitter, Facebook, YouTube, WeChat, Weibo, Zhihu, ……
  - How to **organize** information? (classification & filtering)
    - e.g., automatically assign email to different folders
  - How to **discover** information (or even knowledge) from the data? (mining)
    - e.g., discover correlation of events

What's IR? (cont.)

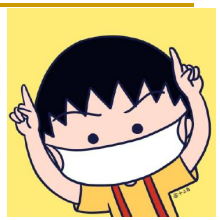


- Goal:
  - Find documents *relevant* to an information need from a large document set

- And now:
  - Beyond relevance
  - Multi-modal documents
  - Users' (implicit) information need
  - Heterogeneous environment

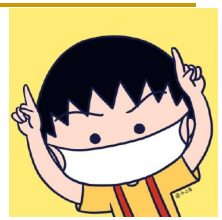
Figure Copyright by TREC

IR is Hard!



- Under/over-specified query
  - Ambiguous: “buying CDs” (certificate deposit? or compact disc?)
  - Incomplete: what kind of CDs?
  - What if “CD” is never mentioned in document?
- Vague semantics of documents
  - Ambiguity: word-sense, structural
    - e.g. “bank”
  - Incomplete: Inferences required
    - E.g. “windows” “apple”
- A difficult task even for human beings!
  - Only 80% agreement in human judgments

IR is "Easy"!



- IR CAN be easy in a particular case
  - Ambiguity in query/document is RELATIVE to the database
  - So, if the query is SPECIFIC enough, just one keyword may get all the relevant documents
- PERCEIVED IR performance is usually better than the actual performance
  - Users can NOT judge the completeness of an answer
  - E.g. Web Search vs. Machine Translation

(To be continued)