CAB431 Welcome
Text Analysis and Web Search

CAB431

Professor Yuefeng Li
School of Computer Science
Queensland University of Technology

The teaching team

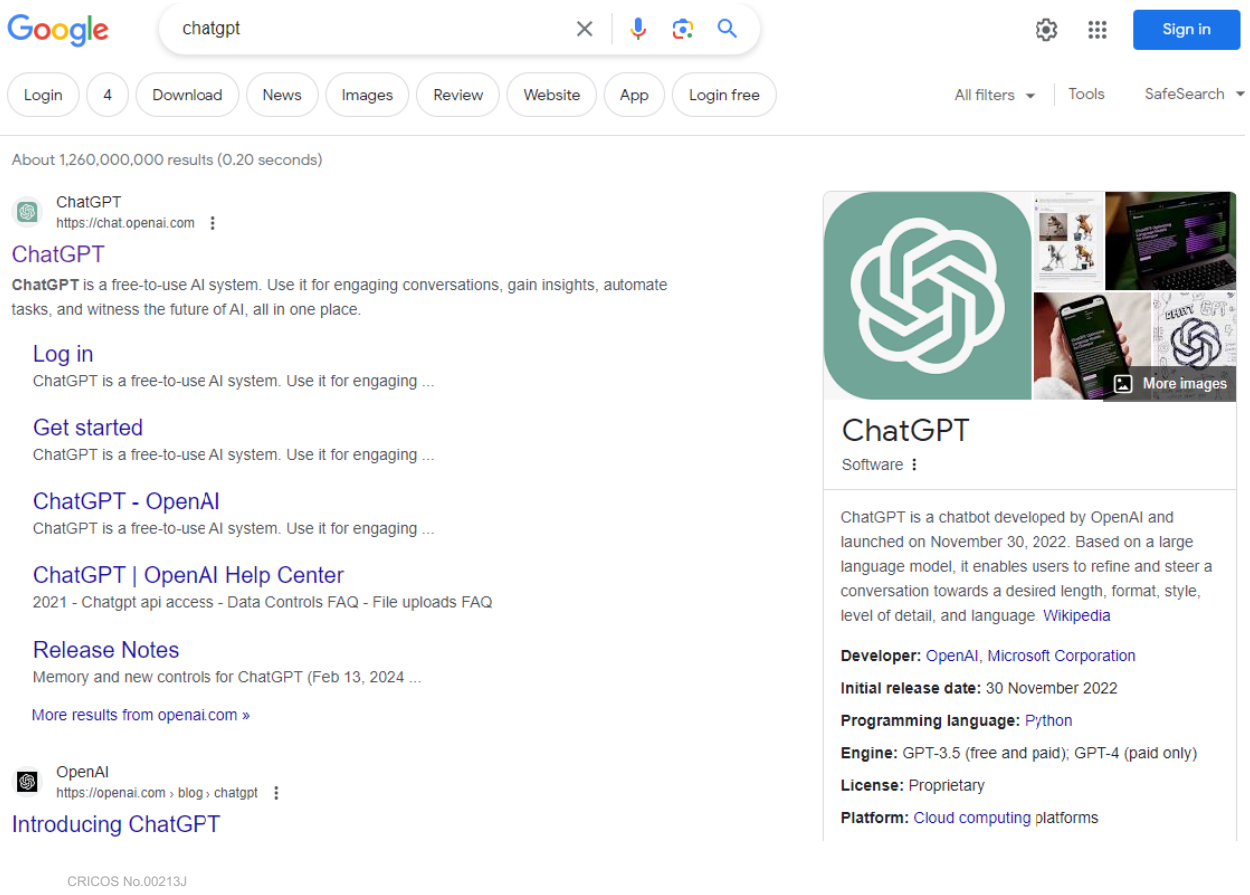
- Unit Coordinator
  - Professor Yuefeng Li
  - School of Computer Science
  - Email: y2.li@qut.edu.au
  - https://www.qut.edu.au/about/our-people/academic-profiles/y2.li
  - Related information for my research
    - QUT eprints
      - https://eprints.qut.edu.au/view/person/Li,_Yuefeng.html
    - Google scholar
      - https://scholar.google.com.au/citations?hl=en&user=IGIkD7oAAAAJ
    - DBLP (a computer science bibliography website)
      - https://dblp.org/pid/74/4581.html

CAB431 Text Analysis and Web Search - Overview

- With the explosion of information resources on the Web, social media and corporate intranets, there is an imminent need for advanced technologies to help people deal with big text data.
- There are many practical applications of Web search and text analysis in the areas such as classification of news stories, academic papers or medical records; spam or junk email filtering, understand customers opinion or behaviors through their feedback in online-systems or social media, customer service promotion etc.
- Therefore, it is urgent for IT developers, Web analysts, information management consultants, or Web development & support officers to understand NLP (Natural Language Processing) techniques, popular text processing models (such as Web search engine, information retrieval models); advanced text mining techniques (such as supervised methods for information filtering or classification and unsupervised method for topic modelling); and future directions in Web Intelligence.

Professor Yuefeng Li
School of Computer Science

CAB431 Text Analysis and Web Search – Overview cont.



- A search engine is software designed to perform web searches. This is a great IT business product because most of us use search engines for our daily information needs, to increase our knowledge, to clarify things, and more.
- Google is a popular search engine. In terms of research, it extends traditional Information Retrieval (IR) research to include many advanced technologies (such as NLP, link analysis, social networks, semantic web, Web intelligence, cybersecurity, cloud computing, question answering systems, etc.).
- In practical applications, many new tools are involved, such as recommender systems, information filtering systems, sentiment analysis, document summarization, chatbot, ChatGPT, etc.
- The content of this unit is advanced and useful for the IT industry to deal with big data.

Why this unit important?

- Big data extends beyond structured data, including unstructured data: **text**, audio, video, click streams, log files and more. The dramatic increase in the availability of massive text data from various sources is creating a number of issues and challenges for **NLP (natural language processing), text analysis and Web search,** such as, the noisy and uncertain information, scalability and effectiveness.
- The future of work post-COVID
  - A lot of new roles will emerge across the care economy and in new industries such as AI (**Artificial Intelligence)** and content creation.
  - A rise in demand for workers able to fill green economy jobs, roles at the forefront of the **data and AI economy**, and new positions in engineering, cloud computing and product development.
- The target audience for this unit is primarily undergraduates in computer science or data science, but graduate students should also find this useful. We also consider the unit to be suitable for most students in information technology discipline.
- Finally, practicing data engineers (and data scientists) should benefit from the unit, whatever their background.
  - There is mathematics in the lecture notes and textbook, but nothing too esoteric. We will teach you how to understand and use them in terms of computer programs.
  - There are also Python code and programming exercises in the workshops and review questions, but nothing beyond the capabilities of someone who has taken some basic computer science and programming classes.

Previous study requirements

- Prerequisites (CAB201 or ITD121)
  - To enroll in this unit, you must have completed these prerequisite units (or have credit, advanced standing or exemption for them), or be able to demonstrate that you have equivalent background knowledge.
- Assumed knowledge
  - *We assume that you have knowledge in programming languages before you start this unit.*

Learning outcomes

- Understand, write, and explain fundamental Web search model, theories, techniques and algorithms.
- Design Web search solutions for user information needs.
- Demonstrate knowledge of advanced text analysis techniques for information filtering, text classification, topic modelling for text feature selection.
- Demonstrate knowledge of the principles and techniques of evaluating text analysis systems performance.
- Work independently or in a team to implement a major text analysis project.

Learning approaches

- This unit has 3-5 contact hours per week organized in Read and watch pre-recorded lectures (about 2 hours), attend lecture supports (zoom session 1 hour) and practical activities (workshops 1-2 hours).
- Students are expected to actively participate in all sessions, take part in the practical activities (both face-face and online workshops), complete review questions (about 2 hours) and assignments (2-3 hours).
- Lecture notes, workshop tasks and review questions (and solutions) will be made available weekly through CAB431 Canvas.

Feedback on learning and assessment

- You can obtain feedback on your progress throughout the unit through the following mechanisms:
  - Ask the teaching team for questions during the lecture support sessions and workshops, or via email and CAB431 slack;
  - You will receive a detailed marking criteria sheet for each assignment;
  - Your tutor or unit coordinator will provide constructive feedback on assessments upon completion.
- Assessment items:
  - **Assessment 1: Portfolio (20%)**
  - **Assessment 2: Project (applied) (35%)**
  - **Assessment 3: Examination (written) (45%)**

All details can be found on CAB431 Canvas – "About Assessment" module