Australian National University

COMP4650/6490 Document Analysis

Introduction

Prof. Graham Williams

Dr. Dawei Chen

Software Innovation Institute
ANU School of Computing

Outline

- Motivation
  - Why document analysis
  - What is this course about
- Course structure
  - Information retrieval (IR)
  - Natural language processing (NLP)
  - Machine learning (ML) for NLP
- Course logistics

Outline

- Motivation
  - Why document analysis
  - What is this course about
- Course structure
  - Information retrieval (IR)
  - Natural language processing (NLP)
  - Machine learning (ML) for NLP
- Course logistics

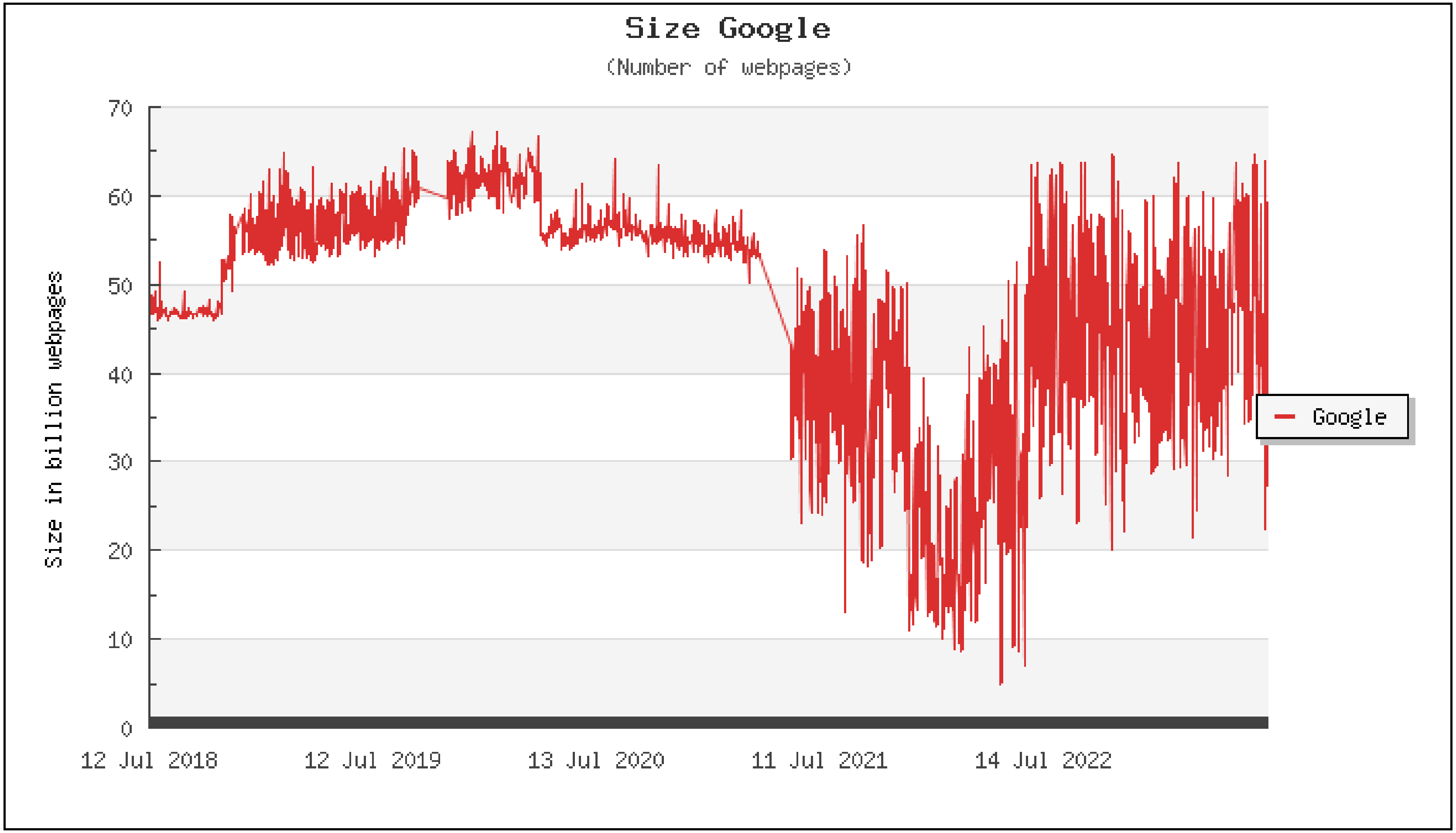
Motivation

Why document analysis?

- There is a large amount of data out there, and much of it is text.
- Examples: Wikipedia, reddit, facebook, twitter, news websites, message boards, research papers.

Motivation

Why document analysis?

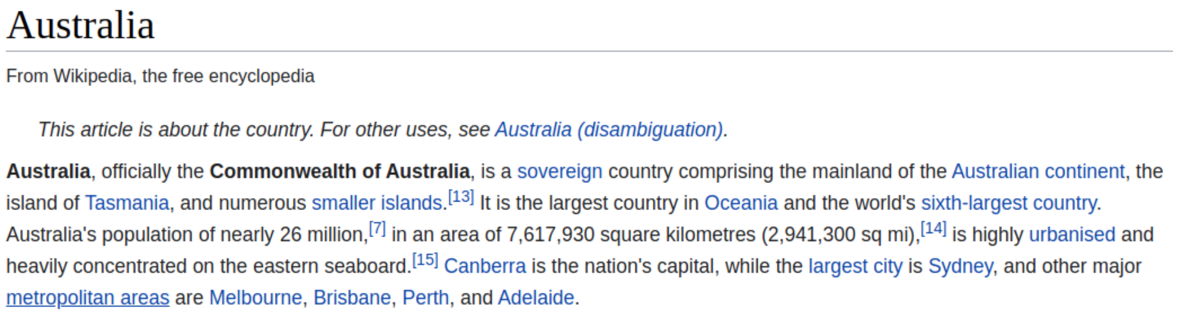


Estimated size of Google's index https://www.worldwidewebsize.com/

Motivation

Why document analysis?

- We want to make use of this data but …
  - Text data is unstructured
  - It certainly contains structure, but it is not easily manipulated in software.



Australia

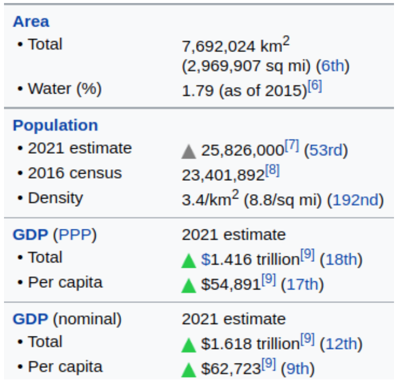
From Wikipedia, the free encyclopedia

This article is about the country. For other uses, see Australia (disambiguation).

Australia, officially the **Commonwealth of Australia**, is a sovereign country comprising the mainland of the Australian continent, the island of Tasmania, and numerous smaller islands.[13] It is the largest country in Oceania and the world's sixth-largest country. Australia's population of nearly 26 million,[7] in an area of 7,617,930 square kilometres (2,941,300 sq mi),[14] is highly urbanised and heavily concentrated on the eastern seaboard.[15] Canberra is the nation's capital, while the largest city is Sydney, and other major metropolitan areas are Melbourne, Brisbane, Perth, and Adelaide.

Unstructured data

- Text
- Images
- Videos
- Audio



Area	
• Total	7,692,024 km^2 (2,969,907 sq mi) (6th)
• Water (%)	1.79 (as of 2015)[6]
Population	
• 2021 estimate	▲ 25,826,000[7] (53rd)
• 2016 census	23,401,892[8]
• Density	3.4/km^2 (8.8/sq mi) (192nd)
GDP (PPP)	2021 estimate
• Total	▲ $1.416 trillion[9] (18th)
• Per capita	▲ $54,891[9] (17th)
GDP (nominal)	2021 estimate
• Total	▲ $1.618 trillion[9] (12th)
• Per capita	▲ $62,723[9] (9th)

Structured data

- Database tables
- XML
- Knowledge bases

Motivation

AI for Text Processing

- Creation of Language Models for
  - set top boxes – siri, alexis, ...
  - online support – telephone company enquiry lines
  - https://mlhub.ai
    - Speech synthesis and transcription
    - Language translations
    - Entities, language, sentiment.
- Traditionally considerable human effort required
  - Each new language model has to be built
- Large Language Models
  - Require massive data/compute

Motivation

Ethics and Document Analysis

- Becoming much easier to infringe on privacy
- Privacy is a Human Right
- Data can be misused:
  - Ridicule, Voyerism, Violence, Blackmail
  - Polical oppression, Domesitc abuse
  - Surveillance, Discredit, Misinformation
  - Kroeger et al “How Data Can Be Used Against People”

    https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3887097
- As practitioners, in every activity, ask yourself

  “**is this the right thing to be doing**”

Motivation

What is this course about?

- Automatic processing of text documents, e.g.
  - News articles
  - Academic papers
  - Webpage text
  - Social media posts
  - Sentences in books
- Documents are in natural human language
- In this course typically English

Outline

- Motivation
  - Why document analysis
  - What is this course about
- Course structure
  - Information retrieval (IR)
  - Natural language processing (NLP)
  - Machine learning (ML) for NLP
- Course logistics

Course Structure

- Information retrieval (IR)
  - How do you automatically identify relevant information from a huge collection of documents?
- Natural language processing (NLP)
  - How can computers understand human language?
- Machine learning (ML) for NLP
  - How can computers learn from text data?

Course Structure

Information Retrieval (IR)

- IR provides easy access to information that addresses an information need
  - Document search
    - Web search, enterprise search
    - Focus of this course
  - Media search
    - Image retrieval, speech retrieval, news search
  - Question answering
    - IBM Watson, Google Assistant, Siri, ChatGPT
  - Recommendation systems
    - Netflix/Youtube videos, Amazon products

Course Structure

IR Topics

- Introduction to IR
  - A model of a search system, Boolean retrieval
- Beyond Boolean retrieval
  - Ranked retrieval, TF-IDF and vector-space models
- Evaluating IR systems
  - Evaluation of unranked and ranked retrieval sets
- Web search basics
  - PageRank
  - Hyperlink-Induced Topic Search (HITS)

Course Structure

Natural Language Processing (NLP)

- NLP provides an interface between human language and computers
- Extract information from text
  - Is this email spam or not
  - Identify verbs in a sentence
  - Identify syntactic relationships between phrases
  - Identify relations in the text
- Generate new text
  - Machine translation
  - Document summarisation
  - Question answering

Course Structure

Natural Language Processing (NLP)

- NLP and IR are deeply connected
- NLP can make use of IR to find/rank/organise documents/sentences/words needed to solve an NLP task
- IR can make use of NLP to improve search and retrieval
  - All modern web search engines use extensive NLP
  - e.g. BERT embeddings, conversational IR such as Bing Chat and Google Bard

Course Structure

Natural Language Processing (NLP)

- Both NLP and IR must deal with human language and its characteristics
  - Ambiguity
    - Language is ambiguous
    - How to deal with language ambiguity?
  - Meaning
    - Where is the meaning of concepts/words/phrases coming from?
  - Multilinguality
    - There are many human languages!

Course Structure

Natural Language Processing (NLP)

- Ambiguity

I made her duck.

- I cooked waterfowl for her
- I cooked waterfowl belonging to her
- I created the duck she owns
- I caused her to quickly lower her head or body
- I waved my magic wand and turned her into a waterfowl

Course Structure

Natural Language Processing (NLP)

- Ambiguity

The police subdued the protestors because they were violent.

Who are we told was violent?

1. The police
2. The protestors

Course Structure

Natural Language Processing (NLP)

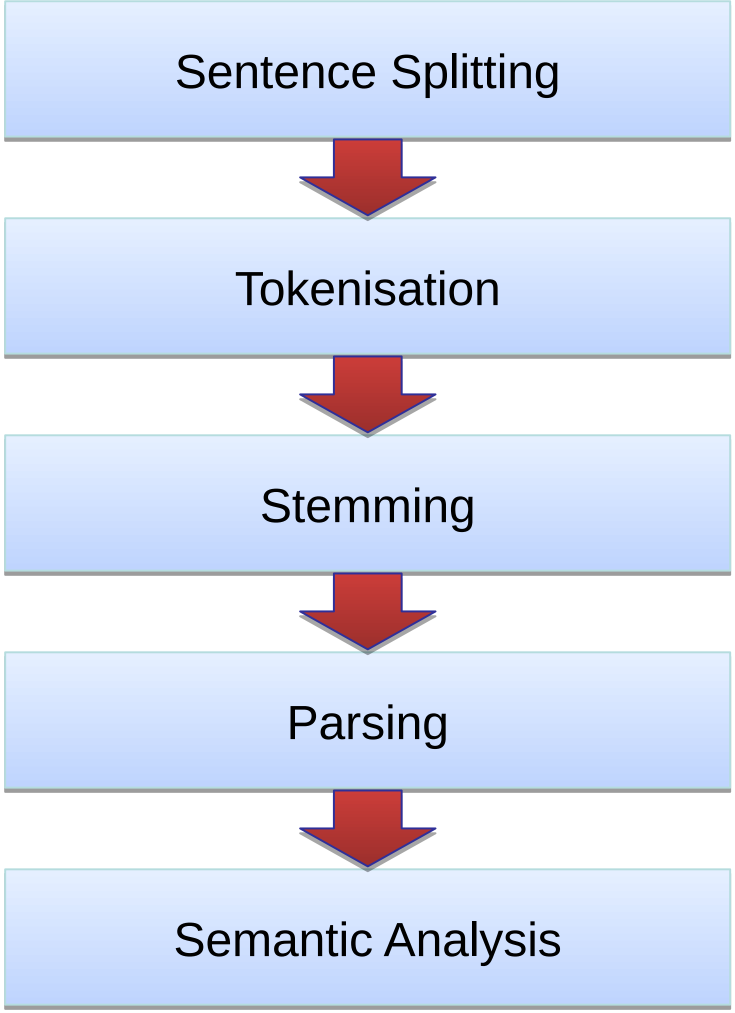
- Linguists defined structure for language at different levels so that it can be studied
- We can we use this knowledge to model language with computers

- Phonetics and phonology
  - Knowledge about linguistic sounds, i.e. distributional patterns of sounds, pronunciations
- Morphology
  - Knowledge of the meaningful components of words, i.e. structure of words, e.g. bathroom
- Syntax
  - Knowledge of the structural relationships between words, i.e. how words combine to form phrases and sentences
- Semantics
  - Knowledge of meaning, i.e. how language conveys meaning, e.g. Queen (Royal, band?)
- Discourse
  - Knowledge about linguistic units larger than a single utterance, e.g. in conversation, relationship of current sentence to previous
- Pragmatics
  - Knowledge of the relationship of meaning to the goals, i.e. how language is used to do things. Context and purpose of an utterance.

Course Structure

Natural Language Processing (NLP)

A typical NLP & IR Pipeline



Sentence Splitting

Tokenisation — Morphology

Stemming — Morphology

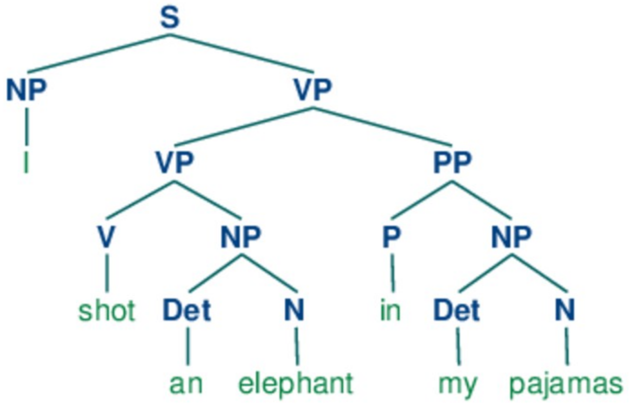
Parsing — Syntax

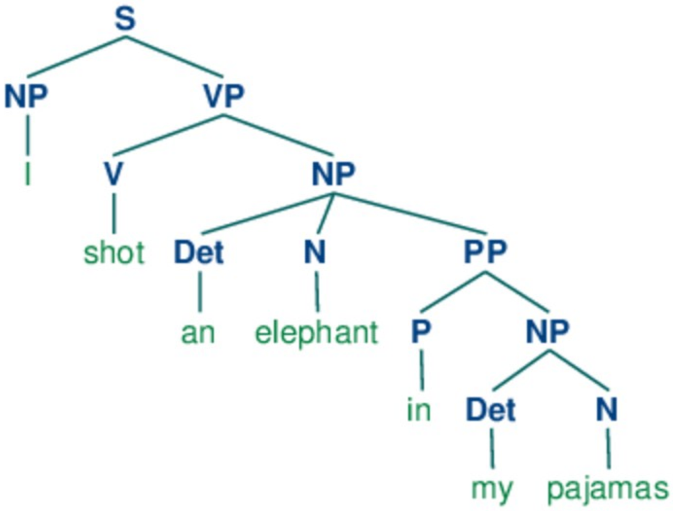
Semantic Analysis — Semantics & Discourse

Course Structure

NLP Topics

- Language modelling & smoothing
- Syntactic parsing
  - Constituency & dependency parsing
- Semantics
  - Lexical, logical, predicate-argument semantics
- Evaluation in NLP
  - Evaluation metrics
  - Methods comparison
- Multilingual & low resource NLP





Course Structure

Machine Learning (ML) for NLP Topics

- Machine learning basics
  - Linear & logistic regression
  - Practical considerations in ML
- Representation
  - Bag-of-word model: One-hot, PPMI
  - Word vectors: LSA, word2vec
- Clustering
- Deep neural networks
  - Feedforward NN, Backpropagation, SGD
  - RNN, LSTM, GRU
- Attention & Transformers
- Pre-trained language models

Outline

- Motivation
  - Why document analysis
  - What is this course about
- Course structure
  - Information retrieval (IR)
  - Natural language processing (NLP)
  - Machine learning (ML) for NLP
- Course logistics

Course Logistics

- Review the following documents on Wattle
  - Course Outline
  - Course Schedule
  - Learning Expectations
- Course team
- Reference books
- Delivery mode
- Communication, assessments, and academic integrity
- Class representatives
- What to do next

Australian National University

Course Logistics

Course Team

- Conveners & Lecturers
  - Prof. Graham Williams
  - Dr. Dawei Chen
    - Administrative questions: dawei.chen@anu.edu.au
- Tutors
  - Han Zhang
  - Jinghan Zhang
  - Karthik Vemireddy
  - Mingrui Gao
  - Naisheng Liang
  - Qingzheng Xu
  - Ruiqi Li
  - Shashank Gummuluru
  - Wei Zhou
  - Yiran Wang
  - Ziyu Chen

Course Logistics

Reference Books (not required)

- Introduction to Information Retrieval
  - Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze
  - Cambridge University Press
  - PDF online https://nlp.stanford.edu/IR-book/
- Speech and Language Processing (3rd ed. draft)
  - Dan Jurafsky and James H. Martin
  - Latest draft https://web.stanford.edu/~jurafsky/slp3/

Course Logistics

Delivery Mode

- Lectures
  - 2 lectures per week presented in-person
  - Recorded and uploaded to Echo360
- Labs
  - 6 labs (weeks 2,3,5,6,8,9), in-person
  - Register for a lab session in MyTimetable
  - Documentation distributed via Wattle
  - Lab sessions to ask questions and get personalised help with the material

Course Logistics

Communication

- Wattle
  - Lecture slides will be posted on Wattle
  - Quizzes and assignments will be provided and submitted through Wattle
- Questions
  - Ed discussion forum: All course/content/exam related questions to be posted to the course forum
  - All admin/personal requests to be emailed to dawei.chen@anu.edu.au
- Information
  - Check the Ed discussion forum regularly!

Course Logistics

Assessments

- See **Course Outline** on Wattle for details
- 3 Assignments (30%)
- 3 Quizzes (10%) and 1 self-assessment (Quiz 0, ungraded)
- Final Exam (60%, hurdle assessment)
- Coding in Python required for Assignments
- Late submission
  - No late submission of the online quizzes and assignments will be permitted, without a pre-arranged extension.
  - If an assessment task is not submitted by the due date, a mark of 0 will be awarded.

Course Logistics

Academic Integrity

- Group work
  - **No group work** is permitted in any part of the assessment in this course
  - Plagiarism will not be tolerated
  - Moss or Turnitin may be used to check assignments
- Plagiarism:
  Passing someone else's work off as your own.
- See university resources for more details
  https://www.anu.edu.au/students/academic-skills/academic-integrity

Class Representatives

- Class Student Representation is an important component of the teaching and learning quality assurance and quality improvement processes within the ANU College of Engineering, Computing and Cybernetics (CECC).
- The role of Class Representatives is to provide ongoing constructive feedback on behalf of the student cohort to Course Conveners and to Associate Directors (Education) for continuous improvements to the course.
- Roles and responsibilities:
  - Act as the official liaison between your peers and convener.
  - Be creative, available and proactive in gathering feedback from your classmates.
  - Attend regular meetings, and provide reports on course feedback to your course convener
  - Close the feedback loop by reporting back to the class the outcomes of your meetings.

Course Logistics

Class Representatives

Why become a class representative?

- **Ensure students have a voice** to their course convener, lecturer, tutors, and College.
- **Develop skills sought by employers**, including interpersonal, dispute resolution, leadership and communication skills.
- **Become empowered**. Play an active role in determining the direction of your education.
- **Become more aware of issues influencing your University** and current issues in higher education.
- **Course design and delivery**. Help shape the delivery of your current courses as well as future improvements for following years.
- **Note**: Class representatives will need to be comfortable with their contact details being made available via Wattle to all students in the class.
- For more information regarding roles and responsibilities, contact ANUSA CECC representatives: sa.cecc@anu.edu.au

Want to be a class representative?

Nominate today!

Please nominate yourself to your course convener by end of Week 2.

What to do next?

1. Read the **Course Outline** and the **Course Schedule**.
2. Sign up for a lab session via MyTimetable.
3. Take **Quiz 0** to assess if this course is right for you.
4. Make sure you are familiar with Python programming and go through the *Python Basics Tutorial*.
5. Start reading any parts of the course textbooks that interest you.
6. Sign up and ask any questions you have about the course on the Ed discussion forum.
7. If you are interested in becoming a course representative email dawei.chen@anu.edu.au