COMP90042 LECTURE 1 A

SUBJECT OVERVIEW

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

COURSE OVERVIEW

Text processing

- Machine learning from words and documents
- Structure prediction, words as sequences and trees

Search

- Efficient information retrieval
- Exploiting the structure of the web

End tasks

- Translation, information extraction, question answering

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

PREREQUISITES

- COMP90049 / COMP30018 “Knowledge Technologies”
- Some Python programming experience
- No knowledge of linguistics or advanced mathematics is assumed
- Caveats – Not “vanilla” computer science
  - Involves some basic linguistics, e.g., syntax and morphology
  - Requires some maths, e.g., algebra, derivatives, linear algebra, dynamic programming

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

EXPECTATIONS AND OUTCOMES

- Expectations
  - develop Python skills
  - keep up with readings
  - classroom participation
- Outcomes
  - Practical familiarity with range of text analysis technologies
  - Understanding of theoretical models underlying these tools
  - Competence in reading research literature

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

ASSESSMENT: ASSIGNMENTS AND EXAM

- Homework (20% total = 4 × 5% each)
  - Small activities building on workshop
  - Released every 2-3 weeks, due the following week
- Project (30% total)
  - Individual work
  - Released before Easter break & due near end of semester
- Exam (50%)
  - two hour, closed book
  - covers content from lectures, workshop **and prescribed reading**
- **Hurdle** >50% exam, and >50% on homework + project

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

TEACHING STAFF

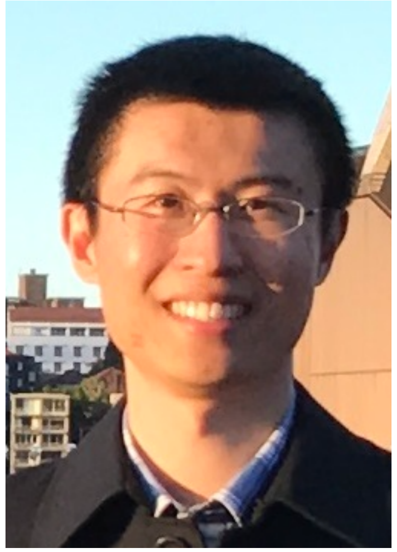
- Lecturers
  - Julian Brooke Trevor Cohn

- Teaching Assistants

Jeremy
Nicholson

Karl
Grieser

Yuan
Li



COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

COURSE OVERVIEW

Introduction to text processing

- Text classification, word meaning and document representations

Structure learning

- Sequence tagging, n-gram language modelling, parsing & translation

Information Retrieval

- Vector space model, efficient indexing, query expansion and using the web as a graph

Larger tasks in Text Analysis

- Information extraction, question answering

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

RECOMMENDED TEXTS

- Use a mixture of texts
  - *Daniel Jurafsky and James H. Martin*, Speech and Language Processing, 2nd & 3rd eds., Prentice Hall. 2009 (out of print) & 2016 draft (free online).
  - *Christopher D. Manning, Prabhakar Raghavan and Hinrich Schtze*, Introduction to Information Retrieval, Cambridge University Press. 2008. $105 (free online)
- Recommended for learning python:
  - *Steven Bird, Ewan Klein and Edward Loper*, Natural Language Processing with Python, O'Reilly, 2009. (free online)
- Reading links or PDFs will be posted to LMS

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

CONTACT HOURS

- Lectures
  - Tue 11-12pm FBE-G06 (Prest Theatre)
  - Wed 2:15-3:15pm Chemistry-189 (Masson Theatre)
- Workshops: enrol in one of
  - Mon 11am, 7:15pm Alice Hoy 108
  - Tue 10am Alice Hoy 222
  - Fri 2:15pm, 5:15pm Alice Hoy 236/211
- Office hour, casual drop in session
  - Bring any questions you have to Julian / Trevor
  - Tues 2.15-3.15pm Doug McDonell 7.02

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

PYTHON

- Making extensive use of python
  - workshops feature programming challenges
  - provided as interactive ‘notebooks’ for workshops
  - homework and project in python
- Using several great python libraries
  - NLTK (text processing)
  - Numpy, Scipy, Matplotlib (maths, plotting)
  - Scikit-Learn (machine learning tools)

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

PYTHON

- Python ‘*Canopy EPD*’ installed on workshop machines
  - Can use this at home (free download, but register with your unimelb email)
  - Based on Python 2.7
- New to Python?
  - Expected to pick this up during the subject, on your own time
- Introductory Python session **this week**
  - Fri 2:15pm-3:15pm   Alice Hoy 236
    Run by Jeremy, covering Python programming fundamentals.

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

WHY PROCESS TEXT?

- Masses of information ‘trapped’ in unstructured text
  - How can we find this information?
  - Let computers automatically reason over this data?
  - First need to understand the structure, find important elements and relations, etc…
  - Over 1000s of languages….
- Challenges
  - Search, displaying results
  - Information extraction
  - Translation
  - Question answering
  - …

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE

A MOTIVATING APPLICATION

- IBM ‘Watson’ system for Question Answering
  - QA over large text collections
    - Incorporating speech recognition, speech synthesis and more
  - https://www.youtube.com/watch?v=FC3IryWr4c8
  - https://www.youtube.com/watch?v=lI-M7O_bRNg (from 3:30-4:30)
- Research behind Watson is *not* revolutionary
  - But this is a transformative result in the history of AI
  - Combines cutting-edge text processing components with large text collections and high performance computing

COPYRIGHT 2017, THE UNIVERSITY OF MELBOURNE