# COMP348 — Document Processing and the Semantic Web

Week 01 Lecture 1: Introduction and Overview

Diego Mollá

COMP348 2018H1

**Abstract**

In this lecture we will do a brief overview of what the unit is about, and we will cover practical issues regarding the unit.

**Update February 20, 2018**

## Contents



## Reading

- Lecture Notes
- Unit guide

**Welcome to COMP348!**

. . . in which you will learn

- how to build software applications
- that use
  1. data mining
  2. knowledge about language
- to do useful things with documents
- with particular emphasis on Web solutions and documents.

# 1 Document Processing and the Semantic Web

### Document Processing

Information Overload

- A lot of information is available as free text.
- The most natural form to write information is through free text.
- A great deal of digital information is available as free text.
- People can read and understand free text easily.
- But it's very hard for machines!

### Document Processing and the Web

### The Web

- The Web was initially conceived as a means to hyperlink documents.
- Most of the information available on the Web is (still) as free text.
- This is what is often called unstructured data.

### Why Document Processing for the Web?

1. Web search: We want to find information.
2. Spam filtering: We want to ignore (some) information.
3. Sentiment analysis: We want to classify information.
4. Text mining: We want to discover information.

### The Semantic Web

Adding Semantics to the Web

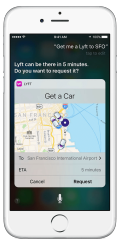
- Web 1.0: The good, old-fashioned Web.
- Web 2.0: The social web.
- Web 3.0: The semantic web.

The Semantic Web is about adding meta-data so that machines can process it.

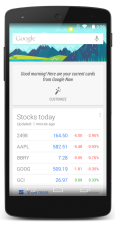

# 2 Example Applications

### Conversational Interfaces

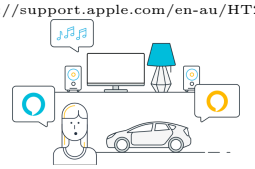
- Siri (Apple iOS), Google Now (Google, Android) are personal digital assistants that, among other things, answer your questions.
- Amazon's Echo and Google Home are products that use a speech interface to provide information and control smart devices.




https://support.apple.com/en-au/HT204389



http://www.androidcentral.com/google-now



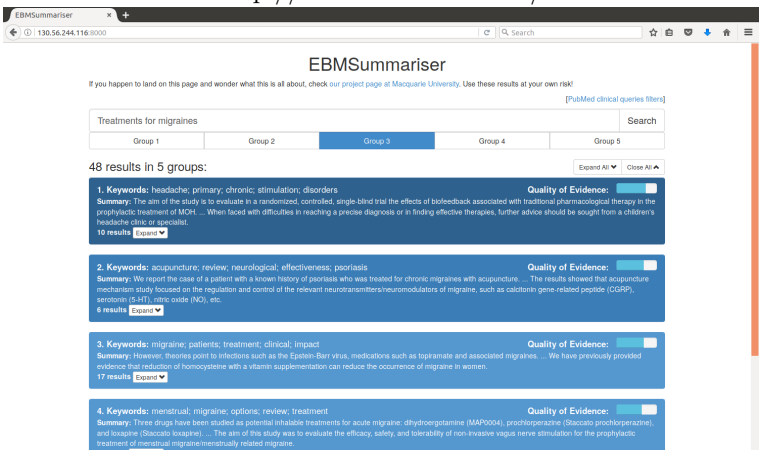
https://developer.amazon.com/alexa-voice-service

### Watson

- Watson: an information extraction system playing Jeopardy!
- Watson on Jeopardy! http://www.youtube.com/watch?v=WFR3lOm_xhE

**Diego's summariser for Evidence Based Medicine**

http://130.56.244.116:8000/




**Machine Learning and Data Mining**

Huge amounts of data are now on-line.

- much of it is unstructured text.

**Data mining** extracting information from large data sources.

**Machine learning** techniques for generalising from data.

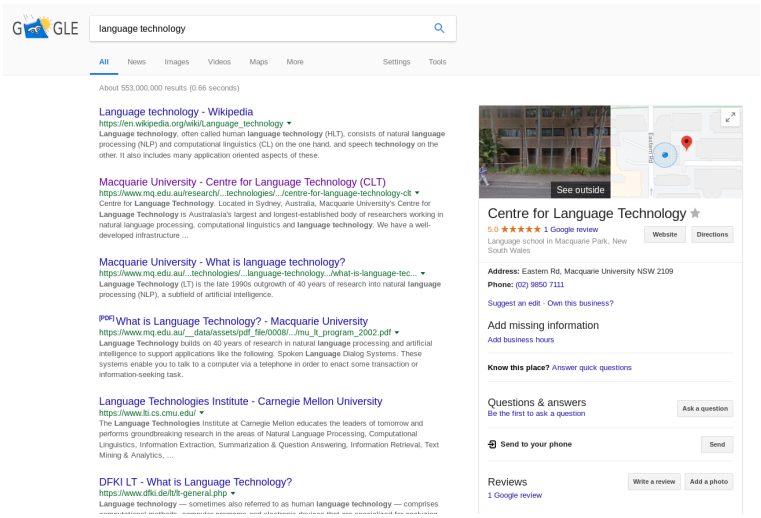
**Integrating Natural Language Processing and Data Mining**

Results to queries asked in current search engines may be enriched with information mined from:

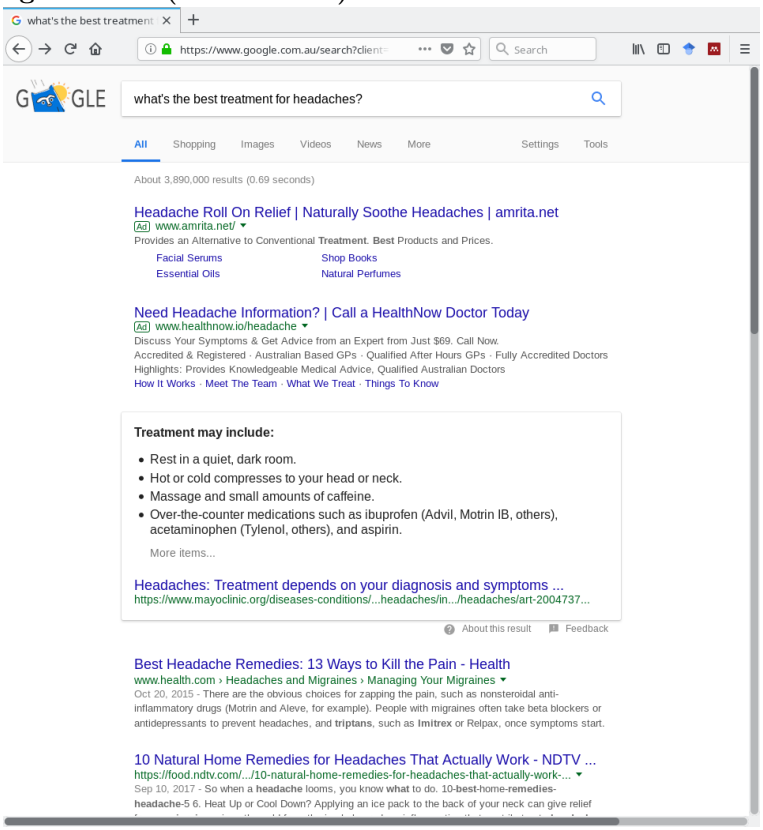
- Knowledge sources such as Google's Knowledge Graph.
- Text mining based on the characteristics of the query.


https://www.google.com/intl/bn/insidesearch/features/search/knowledge.html

**Google Search (13 Feb 2018)**




## Google Search (13 Feb 2018)




## The Semantic Web

*Berners Lee et al. (2001)*
The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.

- The Semantic Web annotates the contents of Web documents with meaning.

- The Semantic Web provides mechanisms to specify meaning and reason with meaning.
- Still largely unrealised, but it has developed various technologies that are becoming increasingly useful.

# 3 Unit Practicalities

### What This Unit is About

- COMP348 explores the issues involved in building significant text processing applications.
  - Emphasis on *non-interactive* natural-language text processing systems.
  - Emphasis also on text processing relative to the Web.
- Programming language: Python.
- This unit has COMP249 or COMP257 as a prerequisite.

### Staff and Consultation Times

**Diego Molla** Unit Convenor, Lecturer
E6A331, diego.molla-aliod@mq.edu.au

**Rolf Schwitter** Lecturer
E6A333, rolf.schwitter@mq.edu.au

**Bayzid Ashik Hossian** Workshops
bayzid-ashik.hossain@mq.edu.au

**Sonit Singh** Workshops
sonit.singh@mq.edu.au

### Web Resources

- The unit is available in iLearn http://ilearn.mq.edu.au.
- All the administrative material presented in this lecture is also available at this site.
  - Unit Outline.
  - Administrative Information.
  - Lecture Notes
  - Pointers to Reading.
  - Other Useful Stuff.
- You are expected to keep up-to-date by using iLearn for:
  - Relevant news and information.
  - Discussions.
  - Submission of assignments.

**Github**

- Some of the material of this unit is available in a public github repository.
- https://github.com/dmollaaliod/comp348-2018
- – Lecture notes
  – Workshop tasks
  – Code
- If you know how to use git, this will be the best way to make sure you have the latest versions.
  – git is one of the most popular version control systems.
  – Search the Web for tutorials and additional information on git.
- You can use the github browser interface to download individual files.

**Learning Outcomes**

1. Explain the main techniques that are used to develop and implement intelligent document processing applications.
2. Describe the functionality of the key components in document processing architectures.
3. Implement text processing applications using a programming language.
4. Apply web technology to document processing.

**Rooms and Times**

**Lectures**

- Monday 4pm-6pm (14 Sir Christopher Ondaatje Ave, 264 Tutorial room)
- Friday 10am-11am (25a Wallys Wlk, 209 Tutorial room)

**Workshops**
One of these; check your timetable!

- Friday 8am-10am (9 Wallys Wlk, 114 Faculty Unix Lab)
- Friday 2pm-4pm (9 Wallys Wlk, 114 Faculty Unix Lab)

***Please Note***
Workshops start from this week.

**Textbooks**

- Weeks 1 to 6 will use (mostly):
  - "NLTK Book": Steven Bird, Ewan Klein, Edward Loper. Natural Language Processing with Python — Analyzing Text with the Natural Language Toolkit. http://www.nltk.org/book
  - "IR Book": Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. http://www-nlp.stanford.edu/IR-book/.
- Weeks 7 to 12 are *not* based on any textbooks; we'll put a list of online texts.
- Every week there will be assigned readings; these readings are essential.
- The web site also has pointers to online resources.
  - Recommendations for additions are welcome.

**Workshops**

**Workshops**

- Tutorial/prac sessions begin from week 1.
- Tasks will typically cover practical assignment tasks and extensions/variations of exam questions.
- The practical exercises will focus on lab work on practical problems.
- This is also your opportunity to discuss and clarify content issues.

**Practical Assessed Assignments**

1. Simple Document Processing (5%, due Week 3)
   - Use of pre-packaged tools.
   - Can be used as a diagnostic test (before census date).
2. Document Processing (20%, due Week 7)
   - Use of techniques used in commercial and research applications.
   - Use of real (messy) text data.
3. Semantic Web (15%, due Week 12)
   - Integration of Semantic Web technologies.

**Submitting your Assignment**

- *Read the assignment specifications.*
- Submit in iLearn.
- Hard deadlines:
  - 20% of the **maximum** mark off per day of delay.

***Plagiarism***

- You may discuss but not write together.
- Read the Academic Honesty Policy. https://staff.mq.edu.au/work/strategy-planning-and-governance/university-policies-and-procedures/policies/academic-honesty

**Assessment**

**Assessment Components**

- Assignment 1: 5%
- Assignment 2: 20%
- Assignment 3: 15%
- Exam: 60%

***Final Assessment***

- Your final mark and grade are entirely determined by the sum of marks of the individual assessment tasks.
- To pass the unit, the sum of marks must be at least 50% of the total assessment marks.
- This unit does not have hurdle assessments.

**Lecture Schedule — Diego**

1. NLP Systems + Text Processing with Python (NLTK Ch 1)
2. Information Retrieval (Manning et al.)
3. Text Classification (I) (NLTK Ch 6)
4. Text Classification (II) (NLTK Ch 6, Manning et al Ch 13))
5. Sequence labelling (NLTK Ch 6)
6. Information Extraction and Summarisation (NLTK Ch 7; Hovy 2003)

**Lecture Schedule — Rolf**

7. The Semantic Web; XML (XSLT tutorial at W3School)
8. RDF, RDF Schema, SPARQL (RDF Primer, SPARQL 1.1 at W3C)
9. Linked Data (DBPedia)
10. Ontologies (Kroestzsch et al 2012, OWL Primer)
11. Rule Languages (RIF Primer)
12. Semantic Web Applications and Recent Trends
13. Revision

**Important Things To Do**

- Print out the lecture notes *before* going to the lecture.
- Read the workshop specification *before* going to the session.
  - time in the sessions is gold.
- Read the online Unit Outline; this is your "contract".
- Schedule an average of 9 hours per week for working on this unit:
  - As in every 3-credit-point unit.
  - This includes the mid-semester break.

**What's Next**

**Friday**

- Python for Text Processing
- Workshop: Python and Text Processing

**Reading**

- NLTK Chapter 1
- http://docs.python.org/tut/tut.html