

# COMP3220 — Document Processing and the Semantic Web

## Week 01 Lecture 1: Introduction and Overview

Diego Mollá

Department of Computer Science  
Macquarie University

COMP3220 2021H1

# Acknowledgement of Country

I would like to acknowledge the traditional custodians of the land where I am located, the Darug and Guringai peoples, and pay my respects to their Elders both past and present.

# Welcome to COMP3220!

... 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.

# Programme

- 1 Document Processing and the Semantic Web
- 2 Example Applications
- 3 Unit Practicalities

## Reading

- Lecture Notes
- Unit guide

# Programme

- 1 Document Processing and the Semantic Web
- 2 Example Applications
- 3 Unit Practicalities

# 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 to process!



# 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) in the form of free text.
- This is what is often called **unstructured data**.

## Why Document Processing for the Web?

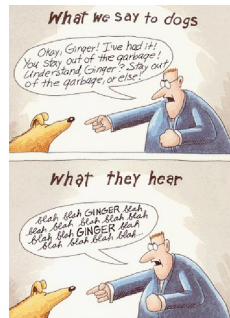
- 1 **Web search:** We want to find information.
- 2 **Spam filtering:** We want to ignore (some) information.
- 3 **Sentiment analysis:** We want to classify information.
- 4 **Text mining:** We want to discover information.

# 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.





# Programme

- 1 Document Processing and the Semantic Web
- 2 Example Applications
- 3 Unit Practicalities

# Conversational Interfaces

- Many platforms offer conversational interfaces where you can talk/write to in plain language.
- The aim is to produce a seamless user experience.
- Siri (Apple iOS), Google Assistant (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.

# Web Search

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.

# Google Search (16 Feb 2021)

Google

language technology

Explicit results filtered with SafeSearch. [Learn more](#) [Undo](#)


About 4,250,000,000 results (0.79 seconds)

**Language Technology** (LT) is the late 1990s outgrowth of 40 years of research into natural language processing (NLP), a subfield of artificial intelligence. ... This translates into applications such as spoken language dialog systems, intelligent Internet search engines, machine translation and automatic text summarisation.

[www.mq.edu.au](http://www.mq.edu.au) » research » centres » information » what...

**What is language technology? - Macquarie University**

About featured snippets Feedback




Rating Visit history

**Centre for Language Technology**  
5.0 ★★★★★ (1) · Language school  
Macquarie University NSW  
(02) 9850 7111

[WEBSITE](#) [DIRECTIONS](#)

**Language Technology Group**  
No reviews · Computer consultant

[WEBSITE](#)



**Language technology**

Language technology, often called human language technology, studies methods of how computer programs or electronic devices can analyze, produce, modify or respond to human texts and speech. Working with language technology often requires broad knowledge not only about linguistics but also about computer science. [Wikipedia](#)

Feedback

# Google Search (16 Feb 2021)

covid-19 treatment

[All](#) [News](#) [Images](#) [Videos](#) [Shopping](#) [More](#) [Settings](#) [Tools](#)

About 4,210,000,000 results (1.05 seconds)

**COVID-19 alert**

Coronavirus disease

[Overview](#)  
[Health info](#)  
[Testing](#)  
[Statistics](#)  
[News](#)  
[Coping](#)  
[Share](#)

Symptoms

Prevention

Treatments

**Self-care**

If you feel sick you should rest, drink plenty of fluid, and eat nutritious food. Stay in a separate room from other family members, and use a dedicated bathroom if possible. Clean and disinfect frequently touched surfaces.

Everyone should keep a healthy lifestyle at home. Maintain a healthy diet, sleep, stay active, and make social contact with loved ones through the phone or internet. Children need extra love and attention from adults during difficult times. Keep to regular routines and schedules as much as possible.

It is normal to feel sad, stressed, or confused during a crisis. Talking to people you trust, such as friends and family, can help. If you feel overwhelmed, talk to a health worker or counsellor.

[Learn more on who.int](#)

For informational purposes only. Consult your local medical authority for advice.

**Medical treatments**

If you have mild symptoms and are otherwise healthy, self-isolate and contact your medical provider or a COVID-19 information line for advice.

Seek medical care if you have a fever, cough, and difficulty breathing. Call in advance.

[Learn more on who.int](#)

For informational purposes only. Consult your local medical authority for advice.

**COVID-19 vaccine**  
See updates and latest info

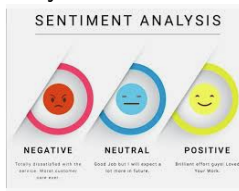
**Map of cases (last 14 days)**

**Cases overview**

New South Wales		
Total cases	Recovered	Deaths
5,138	3,262	54

# Sentiment Analysis

Very often used for analysis of opinions in social media.



Sentiment Analysis and Classification  
kdnuggets.com



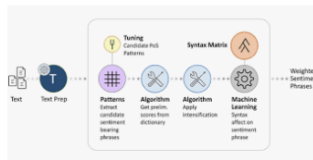
Sentiment Analysis: Concept, Analysis ...  
towardsdatascience.com



A Sentiment Analysis Approach to ...  
medium.com



Sentiment Analysis: Concept, Analysis ...  
towardsdatascience.com



Sentiment Analysis | Lexalytics  
lexalytics.com



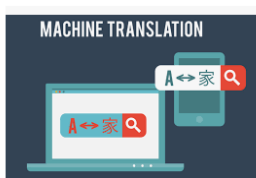
Sentiment Analysis: How Dr  
brandwatch.com

# Machine Translation

Deep learning has dramatically improved the quality of machine translation.



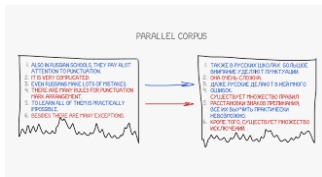
Machine Translation Service  
[translatemedia.com](https://translatemedia.com)



Neural Machine Translation with Python ...  
[towardsdatascience.com](https://towardsdatascience.com)



Machine Translation Explained ...  
[ciklopea.com](https://ciklopea.com)



machine translation from the Cold War ...  
[medium.freecodecamp.org](https://medium.freecodecamp.org)



A Machine Translation Solu...  
[transperfect.com](https://transperfect.com)



An Introduction to Machine Tran...  
[blog.globalizationpartners.com](https://blog.globalizationpartners.com)

# 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.



# Programme

- 1 Document Processing and the Semantic Web
- 2 Example Applications
- 3 Unit Practicalities**

# What This Unit is About

- COMP3220 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 the following prerequisites:
  - COMP2110/COMP249, or
  - COMP2200/COMP257.

# Staff

**Rolf Schwitter:** Unit convenor, lecturer (rolf.schwitter@mq.edu.au).

**Diego Molla:** Lecturer (diego.molla-aliud@mq.edu.au).

**Abdus Salam:** Tutor (abdus.salam@mq.edu.au).

# Delivery

- Lectures:
- Live zoom sessions on Monday 9-11am.
  - Recordings will be available in iLearn.

- Practicals:
- Register to your 2-hour block.
  - There are online sessions (via zoom) and in-campus sessions.
  - See [timetables.mq.edu.au/2021/](https://timetables.mq.edu.au/2021/)

## Please Note

Practicals start from this week.

# 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 and recordings.
  - 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/COMP3220/2021S1>
  - Lecture notes
  - Practicals
  - 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.

# 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.



# 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.

# 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.

# 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>
  - “Deep Learning Book”: François Chollet. Deep Learning with Python. (available in the library).
  - Dan Jurafsky, James H. Martin. Speech and Language Processing. 3rd ed. draft.  
<https://web.stanford.edu/~jurafsky/slp3/>
- 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.

# Assessment

## Assessment Components

- Assignment 1: 5%, due Week 3.
- Assignment 2: 20%, due Week 7.
- Assignment 3: 15%, due Week 12.
- Exam: 60%, online, during the examination period.

## 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.

# Practical 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>

# Tentative Lecture Schedule — Diego

- 1 Python for Text Processing (NLTK Ch 1)
  - 2 Information Retrieval (Manning et al.)
  - 3 Text Classification (NLTK Ch 6)
  - 4 Deep Learning for Text (Chollet, Ch. 2 & 3)
  - 5 Text Sequences (Chollet, Ch. 6)
  - 6 Advanced Deep Learning for Text (lecture notes)
- (recess) - use this time for working on the assignment

# Lecture Schedule — Rolf

- 7 Semantic Technologies (A Review of the Semantic Web Field)
- 8 RDF, RDF Schema and SPARQL (RDF Primer, SPARQL at W3C)
- 9 DBpedia and Wikidata (Wikipedia and DBpedia: a Comparative Study)
- 10 Ontologies (OWL Primer)
- 11 Rule Languages (Applications of Answer Set Programming)
- 12 Recent Trends in Semantic Technologies (lecture notes)
- 13 Revision



# Important Things To Do

- Print out the lecture notes **before** attending the lecture.
- Read the practical exercises **before** attending the session.
  - time in the sessions is gold.
- Read the online Unit Outline; this is your “contract”.
- Schedule an average of 10 hours per week for working on this unit:
  - As in every 10-credit-point unit.
  - This includes the mid-semester break.

# What's Next

## Week 1

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

## Reading

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