COMP3220 — Document Processing and Semantic Technologies

Week 01 Lecture 1: Introduction and Overview

Diego Mollá

COMP3220 2022H1

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, 2022

Contents

1 Document Processing and Semantic Technologies **2**

2 Example Applications **3**

3 Unit Practicalities **5**

Reading

- Lecture Notes
- Unit guide

Acknowledgement of Country

I acknowledge the traditional custodians of the Macquarie University land, the Wallumattagal clan of the Dharug nation, whose cultures and customs have nurtured, and continue to nurture, this land, since the Dreamtime. We pay our respects to Elders past, present and future.

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.

1 Document Processing and Semantic Technologies

Document Processing

Information Overload

- A great deal of digital information is available as free text.
- People can read and understand free text easily.
- But it's very hard to process by 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) in the form of free text.

Examples of 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 and extract information.

Semantic Technologies

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.

What we say to dogs
Okay, Ginger! I've had it! You stay out of the garbage! Understand, Ginger? Stay out of the garbage, or else!
What they hear
blah blah GINGER blah blah blah blah blah blah blah GINGER blah blah blah blah blah blah...

(there is a newer definition of Web 3.0 related to blockchain technology)

2 Example Applications

Conversational Interfaces

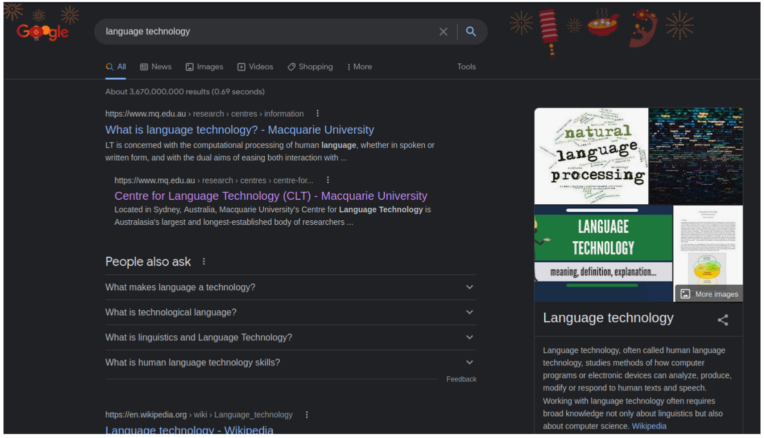
- Many platforms offer conversational interfaces where you can talk/write to in plain language.
- The aim is to produce a seamless user experience.
- Siri (Apple iOS), Google Assistant (Google, Android) are personal digital assistants that, among other things, answer your questions.
- Amazon's Echo and 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 (15 Feb 2022)



language technology

About 3,670,000,000 results (0.69 seconds)

https://www.mq.edu.au › research › centres › information

What is language technology? - Macquarie University

LT is concerned with the computational processing of human **language**, whether in spoken or written form, and with the dual aims of easing both interaction with ...

https://www.mq.edu.au › research › centres › centre-for...

Centre for Language Technology (CLT) - Macquarie University

Located in Sydney, Australia, Macquarie University's Centre for **Language Technology** is Australasia's largest and longest-established body of researchers ...

People also ask

What makes language a technology?

What is technological language?

What is linguistics and Language Technology?

What is human language technology skills?

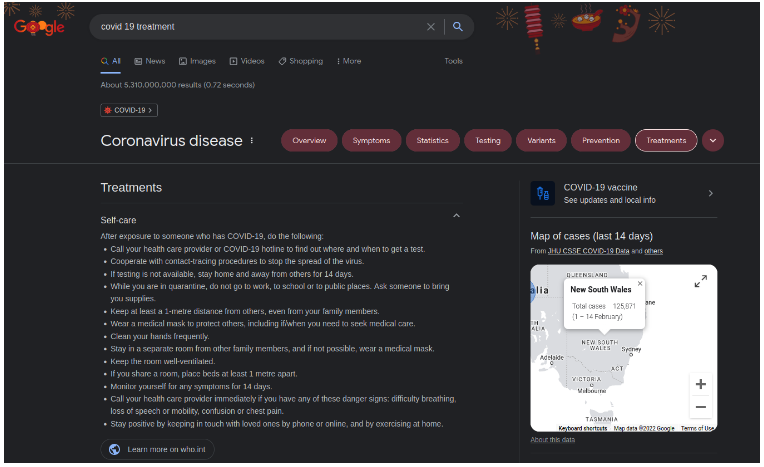
https://en.wikipedia.org › wiki › Language_technology

Language technology - Wikipedia

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

Google Search (15 Feb 2022)



About 5,310,000,000 results (0.72 seconds)

COVID-19 >

Coronavirus disease

Treatments

Self-care

After exposure to someone who has COVID-19, do the following:

- Call your health care provider or COVID-19 hotline to find out where and when to get a test.
- Cooperate with contact-tracing procedures to stop the spread of the virus.
- If testing is not available, stay home and away from others for 14 days.
- While you are in quarantine, do not go to work, to school or to public places. Ask someone to bring you supplies.
- Keep at least a 1-metre distance from others, even from your family members.
- Wear a medical mask to protect others, including if/when you need to seek medical care.
- Clean your hands frequently.
- Stay in a separate room from other family members, and if not possible, wear a medical mask.
- Keep the room well-ventilated.
- If you share a room, place beds at least 1 metre apart.
- Monitor yourself for any symptoms for 14 days.
- Call your health care provider immediately if you have any of these danger signs: difficulty breathing, loss of speech or mobility, confusion or chest pain.
- Stay positive by keeping in touch with loved ones by phone or online, and by exercising at home.

Learn more on who.int

COVID-19 vaccine
See updates and local info

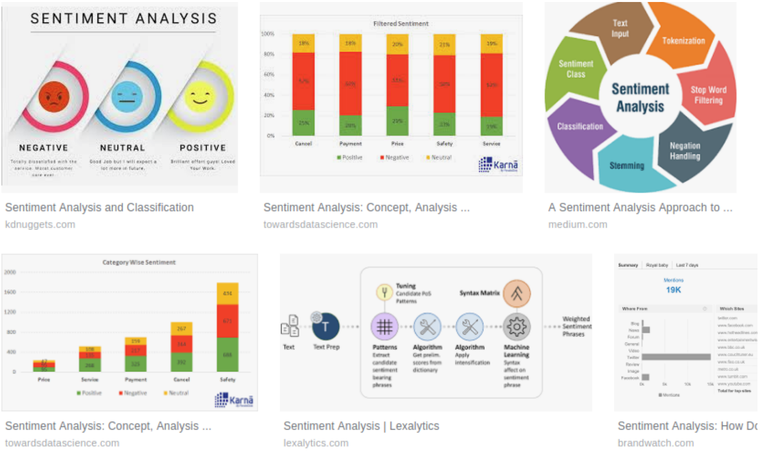
Map of cases (last 14 days)
From JHU CSSE COVID-19 Data and others

New South Wales
Total cases 125,871
(1 – 14 February)

About this data

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

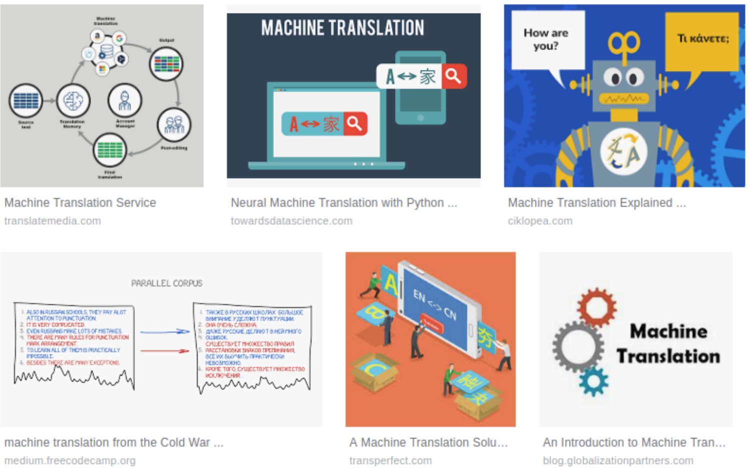
Sentiment Analysis: Concept, Analysis ...
towardsdatascience.com

Sentiment Analysis | Lexalytics
lexalytics.com

Sentiment Analysis: How Do
brandwatch.com

Machine Translation

Deep learning has dramatically improved the quality of machine translation.



Machine Translation Service
translatemedia.com

Neural Machine Translation with Python ...
towardsdatascience.com

Machine Translation Explained ...
ciklopea.com

machine translation from the Cold War ...
medium.freecodecamp.org

A Machine Translation Solu...
transperfect.com

An Introduction to Machine Tran...
blog.globalizationpartners.com

The Semantic Web

Berners Lee et al, The Semantic Web. Scientific American, 2001
"The Semantic Web is not a separate Web but an extension of the current one, 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

- 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, tutor (rolf.schwitter@mq.edu.au).

Diego Molla: Lecturer, tutor (diego.molla-aliod@mq.edu.au).

Jason Ng: Tutor (kingtao.ng@mq.edu.au).

Delivery

Lectures:
- On campus sessions on Monday 9-11am: 14SCO T2.
- Recordings will be available in iLearn / Echo360.

Practicals:
- Register to your 2-hour block.
- These are in-campus sessions.
- See timetables.mq.edu.au/2022/.

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/COMP3220/2022S1`
  - 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.

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 iLearn page also has pointers to online resources.
  - Recommendations for additions are welcome.

Assessment

Assessment Components

- Assignment 1: 10%, due Week 3.
- Assignment 2: 20%, due 2nd week of recess.
- Assignment 3: 20%, due Week 12.
- Exam: 50%, 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 (10%, due Week 3)
   - Use of pre-packaged tools.
   - Can be used as a diagnostic test (results will be out before census date).
2. Document Processing (20%, due 2nd week of recess)
   - Use of techniques used in commercial and research applications.
   - Use of real (messy) text data.
3. Semantic Web (20%, due Week 12)
   - Integration of Semantic Web technology.

Submitting your Assignment

- *Read the assignment specifications.*
- Submit in iLearn.
- Hard deadlines:
  - Late submissions will not be accepted without an approved special consideration request.
  - Assessments submitted after the due date will receive a mark of 0.

Plagiarism

- You may discuss but not write together.
- Read the Academic Integrity Policy. `https://policies.mq.edu.au/document/view.php?id=3`

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)

Lecture Schedule — Rolf

7. Semantic Technologies (A Review of the Semantic Web Field)

   (recess - use this time for working on the assignment)
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

- Download 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.
  - The total work load of all 10-credit points units is 150 hours.

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`