COMP348 — Document Processing and the Semantic Web

Week 01 Lecture 1: Introduction and Overview

Diego Mollá

Department of Computer Science
Macquarie University

COMP348 2019H1

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.

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!

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.

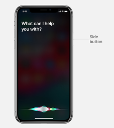
What we say to dogs
Okay, Ginger! I've had it! You stay out of the garbage! Understand, Ginger? Stay out of the garbage, or else!
What they hear
blah blah GINGER blah blah blah blah blah blah blah blah GINGER blah blah blah blah blah...

Programme

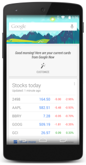
1. Document Processing and the Semantic Web
2. Example Applications
3. Unit Practicalities

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

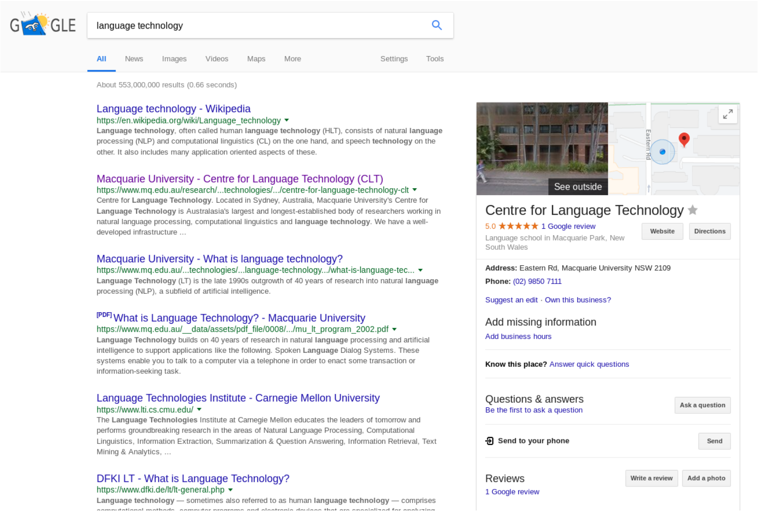
https://developer.amazon.com/alexa-voice-service

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 (13 Feb 2018)



About 553,000,000 results (0.66 seconds)

Language technology - Wikipedia
https://en.wikipedia.org/wiki/Language_technology
Language technology, often called human **language technology** (HLT), consists of natural **language** processing (NLP) and computational linguistics (CL) on the one hand, and speech **technology** on the other. It also includes many application oriented aspects of these.

Macquarie University - Centre for Language Technology (CLT)
https://www.mq.edu.au/research/...technologies/.../centre-for-language-technology-clt
Centre for **Language Technology**. Located in Sydney, Australia, Macquarie University's Centre for **Language Technology** is Australasia's largest and longest-established body of researchers working in natural language processing, computational linguistics and **language technology**. We have a well-developed infrastructure ...

Macquarie University - What is language technology?
https://www.mq.edu.au/...technologies/...language-technology.../what-is-language-tec...
Language Technology (LT) is the late 1990s outgrowth of 40 years of research into natural **language** processing (NLP), a subfield of artificial intelligence.

[PDF] What is Language Technology? - Macquarie University
https://www.mq.edu.au/__data/assets/pdf_file/0008/.../mu_lt_program_2002.pdf
Language Technology builds on 40 years of research in natural **language** processing and artificial intelligence to support applications like the following. Spoken **Language** Dialog Systems. These systems enable you to talk to a computer via a telephone in order to enact some transaction or information-seeking task.

Language Technologies Institute - Carnegie Mellon University
https://www.lti.cs.cmu.edu/
The **Language Technologies** Institute at Carnegie Mellon educates the leaders of tomorrow and performs groundbreaking research in the areas of Natural Language Processing, Computational Linguistics, Information Extraction, Summarization & Question Answering, Information Retrieval, Text Mining & Analytics, ...

DFKI LT - What is Language Technology?
https://www.dfki.de/lt/lt-general.php
Language technology — sometimes also referred to as human **language technology** — comprises computational methods, computer programs and electronic devices that are specialized for analyzing

Centre for Language Technology
5.0 ★★★★★ 1 Google review
Language school in Macquarie Park, New South Wales

Address: Eastern Rd, Macquarie University NSW 2109
Phone: (02) 9850 7111

Suggest an edit · Own this business?

Add missing information
Add business hours

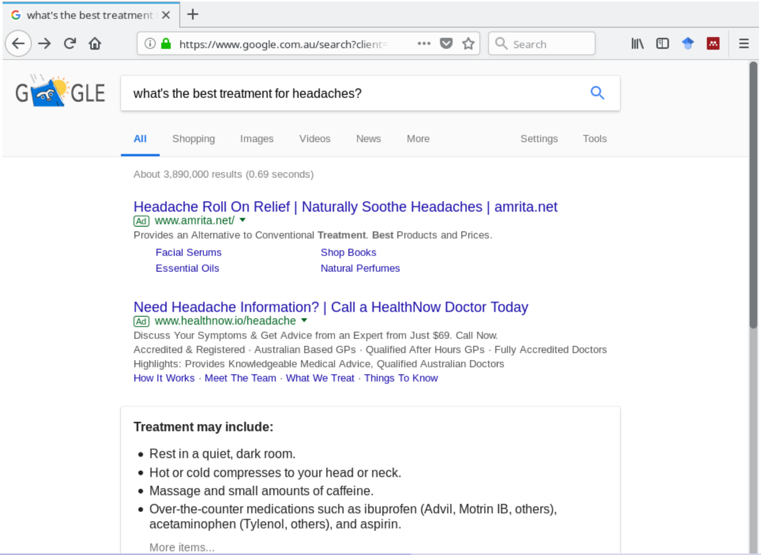
Know this place? Answer quick questions

Questions & answers
Be the first to ask a question

Send to your phone

Reviews
1 Google review

Google Search (13 Feb 2018)



what's the best treatment
https://www.google.com.au/search?client=
Search
GOOGLE
All
Shopping
Images
Videos
News
More
Settings
Tools

what's the best treatment for headaches?

About 3,890,000 results (0.69 seconds)

Headache Roll On Relief | Naturally Soothe Headaches | amrita.net
Ad www.amrita.net/
Provides an Alternative to Conventional **Treatment**. **Best** Products and Prices.
Facial Serums
Shop Books
Essential Oils
Natural Perfumes

Need Headache Information? | Call a HealthNow Doctor Today
Ad www.healthnow.io/headache
Discuss Your Symptoms & Get Advice from an Expert from Just $69. Call Now.
Accredited & Registered · Australian Based GPs · Qualified After Hours GPs · Fully Accredited Doctors
Highlights: Provides Knowledgeable Medical Advice, Qualified Australian Doctors
How It Works · Meet The Team · What We Treat · Things To Know

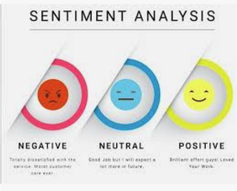
Treatment may include:

- Rest in a quiet, dark room.
- Hot or cold compresses to your head or neck.
- Massage and small amounts of caffeine.
- Over-the-counter medications such as ibuprofen (Advil, Motrin IB, others), acetaminophen (Tylenol, others), and aspirin.

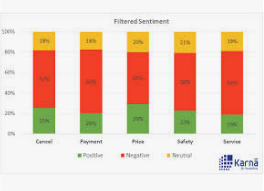
More items...

Sentiment Analysis

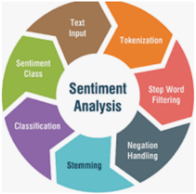
Very often used for analysis of opinions in social media.


SENTIMENT ANALYSIS
NEGATIVE
Totally dissatisfied with the service. Worst customer care ever.
NEUTRAL
Good Job but I will expect a lot more in future.
POSITIVE
Brilliant effort guys! Loved Your Work.

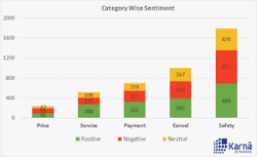
Sentiment Analysis and Classification
kdnuggets.com



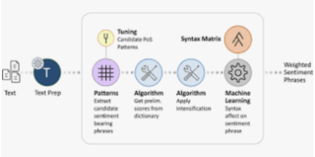
Sentiment Analysis: Concept, Analysis ...
towardsdatascience.com



A Sentiment Analysis Approach to ...
medium.com



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

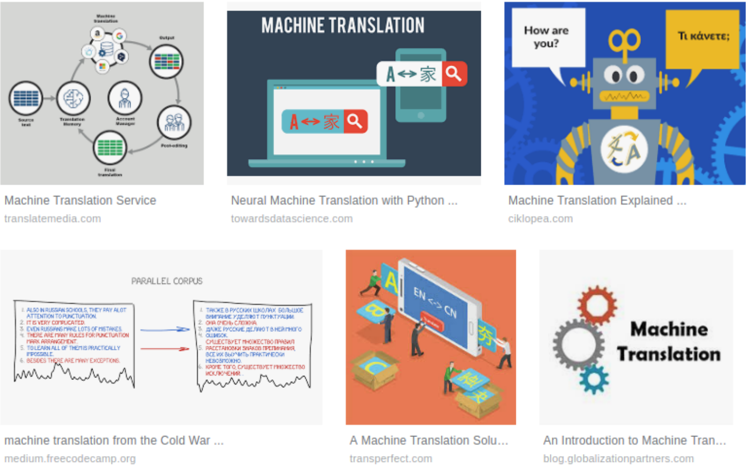

Text
Text Prep
Tuning
Candidate PoS Patterns
Patterns
Extract candidate sentiment bearing phrases
Algorithm
Get prelim. scores from dictionary
Algorithm
Apply intensification
Syntax Matrix
Machine Learning
Syntax affect on sentiment phrase
Weighted Sentiment Phrases

Sentiment Analysis | Lexalytics
lexalytics.com

Sentiment Analysis: How Dc
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. (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

- 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

Rolf Schwitter Unit Convenor, Lecturer
4 Research Park Drive, 359, rolf.schwitter@mq.edu.au

Diego Molla Lecturer
4 Research Park Drive, 358,
diego.molla-aliod@mq.edu.au

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

Abdus Salam Workshops
abdus.salam@hdr.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/comp348-2019
-
    - 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.

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.

Rooms and Times

Lectures

- Monday 9am-11am (3 Innovation Rd - G240 Faculty Tute Rm)
- Friday 10am-11am (3 Innovation Rd - G240 Faculty Tute Rm)

Workshops

One of these; check your timetable!

- Tuesday 11am-1pm (9 Wallys Wlk - 127 Faculty PC Lab)
- Tuesday 1pm-3pm (9 Wallys Wlk - 127 Faculty PC Lab)
- Tuesday 3pm-5pm (9 Wallys Wlk - 127 Faculty PC 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
  - "Deep Learning Book": François Chollet. Deep Learning with Python. (available in the library).
- 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.

Tentative Lecture Schedule — Diego

1. NLP Systems + Text Processing with Python (NLTK Ch 1)
2. Information Retrieval (Manning et al.)
3. Text Classification (NLTK Ch 6)
4. Deep Learning for Text Classification (Chollet, Ch. 2 & 3)
5. Text Sequences (Chollet, Ch. 6)
6. Generation of Text (Chollet, Ch. 8.1)

Lecture Schedule — Rolf

7. The Semantic Web; XML and XSLT (XSLT tutorial at W3School)
8. RDF, RDF Schema, SPARQL (RDF Primer, SPARQL 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

Tuesday

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

Reading

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