COMP348

Document Processing and the Semantic Web

S1 Day 2019

Dept of Computing

Contents

General Information

Unit convenor and teaching staff
Unit convenor
Rolf Schwitter
rolf.schwitter@mq.edu.au
Contact via rolf.schwitter@mq.edu.au
4 Research Park Drive; office 359
Fridays 12-13

Lecturer
Diego Molla-Aliod
diego.molla-aliod@mq.edu.au
Contact via diego.molla-aliod@mq.edu.au
4 Research Park Drive; office 358
See http://web.science.mq.edu.au/~diego/

Credit points
3

Prerequisites
(39cp at 100 level or above) including (COMP249 or COMP257)

Corequisites

Co-badged status

Unit description
This unit explores the issues involved in building natural language processing (NLP) applications that operate on large bodies of real text such as are found on the world wide web. In this unit we discuss some core methods and tools for dealing with data on the web; in particular machine learning platforms widely used in industry. The unit also explores some recent developments of the web, such as emerging semantic web technologies and the corresponding standards promoted by the Word Wide Web Consortium (W3C). Application areas covered include web search, sentiment analysis, and information extraction.

Important Academic Dates

Information about important academic dates including deadlines for withdrawing from units are available at https://www.mq.edu.au/study/calendar-of-dates

Learning Outcomes

On successful completion of this unit, you will be able to:

Explain the main techniques that are used to develop and implement intelligent document processing applications.

Describe the functionality of the key components in document processing architectures.

Implement text processing applications using a programming language.

Apply web technology to document processing.

Assessment Tasks

Name	Weighting	Hurdle	Due
Assignment 1	5%	No	Week 3
Assignment 2	20%	No	Week 7
Assignment 3	15%	No	Week 12
Final exam	60%	No	Examination period

Assignment 1

Due: **Week 3**
Weighting: **5%**

In this assignment you will implement a simple document processing application that uses pre-packaged tools.

The assignment will be submitted using iLearn.

On successful completion you will be able to:

- Explain the main techniques that are used to develop and implement intelligent document processing applications.
- Implement text processing applications using a programming language.
- Apply web technology to document processing.

Assignment 2

Due: **Week 7**
Weighting: **20%**

This assignment will use more powerful techniques such as those used in commercial and research applications. You will experience the processing of real text data, which can be messy and unpredictable at times. At the end of the assignment you will submit a report describing the system, its implementation, and its evaluation.

The assignment will be submitted using iLearn.

On successful completion you will be able to:

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

Assignment 3

Due: **Week 12**
Weighting: **15%**

In this assignment you will experiment with the integration of Semantic Web technology into document processing. You will be asked to study a particular domain and report on the integration of Semantic Web technnologies suitable for the domain, including what sort of SPARQL queries would be applicable to solve specific user needs.

The assignment will be submitted using iLearn.

On successful completion you will be able to:

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

Final exam

Due: **Examination period**
Weighting: **60%**

The final exam will focus on the theoretical aspects of the unit. There will be few questions about implementation issues.

On successful completion you will be able to:

- Explain the main techniques that are used to develop and implement intelligent document processing applications.
- Describe the functionality of the key components in document processing architectures.

Delivery and Resources

Required and Recommended Texts

Most of the contents of the unit will be based on the following two books:

- Steven Bird, Ewan Klein, Edward Loper. *Natural Language Processing -- Analyzing Text with Python and the Natural Language Toolkit*. Online at http://www.nltk.org/book.
- F. Chollet (2017). Deep Learning with Python. Manning Publications. Available in the library.

Additional material will be made available during the semester, in conjunction with the lecture notes. See the unit schedule for a listing of the most relevant reading for each week.

Technology Used and Required

The following software is used in COMP348:

1. Anaconda for Python 3.6
2. NLTK (bundled with Anaconda)
3. Python SciKit-Learn (bundled with Anaconda)
4. gensim (bundled with Anaconda)
5. spaCy (can be installed using Anaconda)
6. Keras (can be installed using Anaconda)
7. Tensorflow (can be installed using Anaconda)
8. XML Copy Editor
9. BaseX (XML Database Engine)
10. Saxon (XSLT and XQuery Processor)
11. rdflib (can be installed using Anaconda)
12. Protege (Ontology Editor)

This software is installed in the labs; you should also ensure that you have working copies of all the above on your own machine. Note that many packages come in various versions; to avoid potential incompatibilities, you should install versions as close as possible to those used in the labs.

Unit Web Page

Note that the majority of the unit materials is publicly available while some material requires you to log in to iLearn to access it.

The unit will make extensive use of discussion boards hosted within iLearn. Please post questions there, they will be monitored by the staff on the unit.

Unit Schedule

Week	Topic	Reading
1	NLP Systems + Text Processing in Python	NLTK Ch 1
2	Information retrieval	Manning et al. (2008)

3	Text Classification	NLTK Ch 6
4	Deep Learning for Text	Chollet, Ch. 6
5	Text Generation with LSTM	Chollet, Ch. 8.1
6	Advanced Usage of Deep Learning for Text	Chollet, Ch. 6
7	XML and XSLT	XSLT Tutorial at W3School
	Recess	
8	RDF, RDF Schema and SPARQL	RDF Primer SPARQL
9	Linked Data	DBpedia
10	Ontologies	Kroetzsch et al (2012) OWL Primer
11	Rule Languages	RIF Primer
12	Semantic Web Applications and Recent Trends	
13	Revision	

Policies and Procedures

Macquarie University policies and procedures are accessible from Policy Central (https://staff.mq.edu.au/work/strategy-planning-and-governance/university-policies-and-procedures/policy-central). Students should be aware of the following policies in particular with regard to Learning and Teaching:

- Academic Appeals Policy
- Academic Integrity Policy
- Academic Progression Policy
- Assessment Policy
- Fitness to Practice Procedure
- Grade Appeal Policy
- Complaint Management Procedure for Students and Members of the Public
- Special Consideration Policy *(**Note:** The Special Consideration Policy is effective from 4 December 2017 and replaces the Disruption to Studies Policy.)*

Undergraduate students seeking more policy resources can visit the Student Policy Gateway (https://students.mq.edu.au/support/study/student-policy-gateway). It is your one-stop-shop for the

key policies you need to know about throughout your undergraduate student journey.

If you would like to see all the policies relevant to Learning and Teaching visit Policy Central (https://staff.mq.edu.au/work/strategy-planning-and-governance/university-policies-and-procedures/policy-central).

Student Code of Conduct

Macquarie University students have a responsibility to be familiar with the Student Code of Conduct: https://students.mq.edu.au/study/getting-started/student-conduct

Results

Results published on platform other than eStudent, (eg. iLearn, Coursera etc.) or released directly by your Unit Convenor, are not confirmed as they are subject to final approval by the University. Once approved, final results will be sent to your student email address and will be made available in eStudent. For more information visit ask.mq.edu.au or if you are a Global MBA student contact globalmba.support@mq.edu.au

Final Examination

If you receive special consideration for the final exam, a supplementary exam will be scheduled in the interval between the regular exam period and the start of the next session. By making a special consideration application for the final exam you are declaring yourself available for a resit during the supplementary examination period and will not be eligible for a second special consideration approval based on pre-existing commitments. Please ensure you are familiar with the policy prior to submitting an application. You can check the supplementary exam information page on FSE101 in iLearn (bit.ly/FSESupp) for dates, and approved applicants will receive an individual notification one week prior to the exam with the exact date and time of their supplementary examination.

Student Support

Macquarie University provides a range of support services for students. For details, visit http://students.mq.edu.au/support/

Learning Skills

Learning Skills (mq.edu.au/learningskills) provides academic writing resources and study strategies to improve your marks and take control of your study.

- Workshops
- StudyWise
- Academic Integrity Module for Students
- Ask a Learning Adviser

Student Services and Support

Students with a disability are encouraged to contact the Disability Service who can provide appropriate help with any issues that arise during their studies.

Student Enquiries

For all student enquiries, visit Student Connect at ask.mq.edu.au

If you are a Global MBA student contact globalmba.support@mq.edu.au

IT Help

For help with University computer systems and technology, visit http://www.mq.edu.au/about_us/offices_and_units/information_technology/help/.

When using the University's IT, you must adhere to the Acceptable Use of IT Resources Policy. The policy applies to all who connect to the MQ network including students.

Graduate Capabilities

Creative and Innovative

Our graduates will also be capable of creative thinking and of creating knowledge. They will be imaginative and open to experience and capable of innovation at work and in the community. We want them to be engaged in applying their critical, creative thinking.

This graduate capability is supported by:

Learning outcomes

- Implement text processing applications using a programming language.
- Apply web technology to document processing.

Assessment tasks

- Assignment 1
- Assignment 2
- Assignment 3
- Final exam

Capable of Professional and Personal Judgement and Initiative

We want our graduates to have emotional intelligence and sound interpersonal skills and to demonstrate discernment and common sense in their professional and personal judgement. They will exercise initiative as needed. They will be capable of risk assessment, and be able to handle ambiguity and complexity, enabling them to be adaptable in diverse and changing environments.

This graduate capability is supported by:

Learning outcomes

- Explain the main techniques that are used to develop and implement intelligent document processing applications.

- Describe the functionality of the key components in document processing architectures.
- Implement text processing applications using a programming language.
- Apply web technology to document processing.

Assessment tasks

- Assignment 1
- Assignment 2
- Assignment 3
- Final exam

Discipline Specific Knowledge and Skills

Our graduates will take with them the intellectual development, depth and breadth of knowledge, scholarly understanding, and specific subject content in their chosen fields to make them competent and confident in their subject or profession. They will be able to demonstrate, where relevant, professional technical competence and meet professional standards. They will be able to articulate the structure of knowledge of their discipline, be able to adapt discipline-specific knowledge to novel situations, and be able to contribute from their discipline to inter-disciplinary solutions to problems.

This graduate capability is supported by:

Learning outcomes

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

Assessment tasks

- Assignment 1
- Assignment 2
- Assignment 3
- Final exam

Critical, Analytical and Integrative Thinking

We want our graduates to be capable of reasoning, questioning and analysing, and to integrate and synthesise learning and knowledge from a range of sources and environments; to be able to critique constraints, assumptions and limitations; to be able to think independently and systemically in relation to scholarly activity, in the workplace, and in the world. We want them to have a level of scientific and information technology literacy.

This graduate capability is supported by:

Learning outcomes

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

Assessment tasks

- Assignment 1
- Assignment 2
- Assignment 3
- Final exam

Problem Solving and Research Capability

Our graduates should be capable of researching; of analysing, and interpreting and assessing data and information in various forms; of drawing connections across fields of knowledge; and they should be able to relate their knowledge to complex situations at work or in the world, in order to diagnose and solve problems. We want them to have the confidence to take the initiative in doing so, within an awareness of their own limitations.

This graduate capability is supported by:

Learning outcomes

- Implement text processing applications using a programming language.
- Apply web technology to document processing.

Assessment tasks

- Assignment 1
- Assignment 2
- Assignment 3
- Final exam

Effective Communication

We want to develop in our students the ability to communicate and convey their views in forms effective with different audiences. We want our graduates to take with them the capability to read, listen, question, gather and evaluate information resources in a variety of formats, assess, write clearly, speak effectively, and to use visual communication and communication technologies as appropriate.

This graduate capability is supported by:

Learning outcomes

- Explain the main techniques that are used to develop and implement intelligent document processing applications.
- Describe the functionality of the key components in document processing architectures.

Assessment tasks

- Assignment 1
- Assignment 2
- Assignment 3
- Final exam

Assessment Standards

The following table shows an indication of achievements required for each learning outcome. The standards of a level also include the standards of a lower level. For example, the standards of Proficient includes the standards of Functional and Developing.

Learning Outcome	Developing	Functional	Proficient
1. Describe the range of applications that require intelligent text processing.	Limited ability to describe the main applications.	Ability to describe the main characteristics of the main applications.	Ability to describe and compare a wide range of applications, providing insight about their key issues and current state of the art.
2. Explain the main techniques that are used to develop and implement intelligent document processing applications.	Ability to describe only some of the main techniques.	Ability to describe the main techniques.	Ability to apply the techniques to specific problem instances.
3. Explain the main components of the Semantic Web and how they relate to Document Processing.	Limited ability to explain the main components of the Semantic Web.	Ability to describe the main components of the Semantic Web.	Ability to explain the main components of the Semantic Web, with insightful references about the interplay between Semantic Web and document processing.
4. Implement text processing applications using a programming language such as Python.	Ability to implement trivial applications that are not much more complex than the examples given in standard textbooks and tutorials.	Ability to implement, document and evaluate simple end-to-end intelligent text-processing applications.	Ability to implement and evaluate complex intelligent text-processing applications. Ability to document and evaluate the implementation in a manner that shows insight.
5. Integrate Semantic Web technology into Document Processing.	Limited ability to implement core elements of Semantic Web applications.	Ability to implement and document simple Semantic Web applications.	Ability to implement and document Semantic Web applications that require the use of Document Processing technology, in a manner that shows insight.

All the unit assessed tasks will be marked on a numerical scale that reflects how well the student meets the relevant learning outcomes. This mapping of learning outcomes to marks will be

specified in the assignment descriptions.

In order to pass the unit, you need to obtain at least 50% of the sum of marks of all individual assessments.

Changes Made to Previous Offerings

This year, we will introduce students to the field of deep learning in Week 4, 5 and 6 using Python and the Keras library as well as TensorFlow.