COMP348

Document Processing and the Semantic Web

S1 Evening 2015

Dept of Computing

Contents

General Information

Unit convenor and teaching staff
Unit Convenor
Diego Molla-Aliod
diego.molla-aliod@mq.edu.au
Contact via diego.molla-aliod@mq.edu.au
E6A331
See http://web.science.mq.edu.au/~diego/

Lecturer
Mark Johnson
mark.johnson@mq.edu.au
Contact via mark.johnson@mq.edu.au
E6A316
See http://web.science.mq.edu.au/~mjohnson/

Tutor
Charles Kuan

Credit points
3

Prerequisites
39cp and COMP249(P)

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. With the web full of unstructured and largely text-based data, the applications needed to handle this have their own particular characteristics. In this unit we discuss some core applications for dealing with data on the web, such as spam filtering and search engines. The unit also explores some developments of web, such as emerging semantic web technologies which support the exchange of XML metadata on the web, and Web 2.0 technologies (such as social networking, folksonomies, wikis and blogs). Application areas covered include information retrieval, web search, document summarisation, machine translation 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:

Describe the range of applications that require intelligent document processing.
Explain the main techniques that are used to develop and implement intelligent document processing applications.
Explain the main components of the Semantic Web and how they relate to Document Processing.
Implement text processing applications using a programming language such as Python.
Integrate Semantic Web technology into Document Processing

Assessment Tasks

Name	Weighting	Due
Assignment 1	10%	Week 5
Assignment 2	15%	Week 7
Assignment 3	15%	Week 12
Final exam	60%	Examination period

Assignment 1

Due: **Week 5**
Weighting: **10%**

In this assignment you will implement a simple document processing application that uses ad-hoc techniques. You will evaluate the quality of the application and describe the implementation and 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.
- Implement text processing applications using a programming language such as Python.

Assignment 2

Due: **Week 7**
Weighting: **15%**

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.
- Implement text processing applications using a programming language such as Python.

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 components of the Semantic Web and how they relate to Document Processing.
- Integrate Semantic Web technology into 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:

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

document processing applications.

- Explain the main components of the Semantic Web and how they relate to Document Processing.

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.
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008. Online at http://www-nlp.stanford.edu/IR-book/.

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. Python 3.4.2 : www.python.org
2. iPython notebook 2.3: ipython.org/notebook.html
3. NLTK 3.0: nltk.org
4. scikit-learn 0.15: scikit-learn.org
5. rdflib 4.1.2: pypi.python.org/pypi/rdflib/
6. pyld: github.com/digitalbazaar/pyld
7. SPARQLWrapper 1.6.4: pypi.python.org/pypi/SPARQLWrapper/1.6.4
8. MongoDB: www.mongodb.org
9. Protege Desktop: protege.stanford.edu
10. Saxon HE: saxon.sourceforge.net
11. BaseX: basex.org/products/download/all-downloads/
12. XML Copy Editor: xml-copy-editor.sourceforge.net

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

The webpage for this unit can be found at http://www.comp.mq.edu.au/units/comp348. Note that

the majority of the unit materials are 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	Classification with K-nearest neighbours	
4	Clustering with K-means	Manning et al Ch 16
5	Probabilistic models and naive Bayes classifiers	NLTK Ch 6 Manning et al Ch 13
6	Sequence Labelling and Hidden Markov Models	NLTK Ch 6
	RECESS	
7	The Semantic Web; XML	XSLT Tutorial at W3School
8	RDF	SPARQL 1.1 at W3C
9	Ontologies and Logic	Kroestzsch et al (2012)
10	Linked Data; Semantic Web Applications	Vinay et al (2013) Arnold (2008)
11	Advanced Text Classification	NLTK Ch 6 Manning et al Ch 15
12	Information Extraction and Summarisation	NLTK Ch 7 Hovy (2003)
13	Revision	

Policies and Procedures

Macquarie University policies and procedures are accessible from Policy Central. Students should be aware of the following policies in particular with regard to Learning and Teaching:

Academic Honesty Policy http://mq.edu.au/policy/docs/academic_honesty/policy.html

Assessment Policy http://mq.edu.au/policy/docs/assessment/policy.html

Grading Policy http://mq.edu.au/policy/docs/grading/policy.html

Grade Appeal Policy http://mq.edu.au/policy/docs/gradeappeal/policy.html

Grievance Management Policy http://mq.edu.au/policy/docs/grievance_management/policy.html

Disruption to Studies Policy http://www.mq.edu.au/policy/docs/disruption_studies/policy.html *The Disruption to Studies Policy is effective from March 3 2014 and replaces the Special Consideration Policy.*

In addition, a number of other policies can be found in the Learning and Teaching Category of 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/support/student_conduct/

Results

Results shown in *iLearn*, 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.

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

IT Help

For help with University computer systems and technology, visit http://informatics.mq.edu.au/help/.

When using the University's IT, you must adhere to the Acceptable Use 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 outcome

- Implement text processing applications using a programming language such as Python.

Assessment tasks

- Assignment 2
- Assignment 3

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

- Describe the range of applications that require intelligent document processing.
- Explain the main techniques that are used to develop and implement intelligent document processing applications.
- Explain the main components of the Semantic Web and how they relate to Document Processing.
- Implement text processing applications using a programming language such as Python.
- Integrate Semantic Web technology into Document Processing

Assessment tasks

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

Commitment to Continuous Learning

Our graduates will have enquiring minds and a literate curiosity which will lead them to pursue knowledge for its own sake. They will continue to pursue learning in their careers and as they participate in the world. They will be capable of reflecting on their experiences and relationships with others and the environment, learning from them, and growing - personally, professionally and socially.

This graduate capability is supported by:

Assessment task

- Assignment 3

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

- Describe the range of applications that require intelligent document processing.
- Explain the main techniques that are used to develop and implement intelligent document processing applications.
- Explain the main components of the Semantic Web and how they relate to Document Processing.
- Implement text processing applications using a programming language such as Python.
- Integrate Semantic Web technology into 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.
- Explain the main components of the Semantic Web and how they relate to Document Processing.
- Implement text processing applications using a programming language such as Python.
- Integrate Semantic Web technology into Document Processing

Assessment tasks

- 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

- Explain the main techniques that are used to develop and implement intelligent document processing applications.
- Implement text processing applications using a programming language such as Python.
- Integrate Semantic Web technology into Document Processing

Assessment tasks

- Assignment 2
- Assignment 3

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

- Describe the range of applications that require intelligent document processing.
- Explain the main techniques that are used to develop and implement intelligent document processing applications.
- Explain the main components of the Semantic Web and how they relate to Document Processing.

Assessment tasks

- Assignment 2
- Assignment 3

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.	Anility 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.

Your final grade depends on your performance in each part of the assessments. **Note that on occasion your raw mark (i.e. the total of your marks for each assessment item) may not be the same as the SNG which you receive**. In particular, if your exam marks are too low you may be awarded a lower grade than the one of the range of your raw marks.

You will obtain a grade of Pass if you meet the learning outcomes of this unit at a basic level. In particular:

- If you perform satisfactorily in the examination (at least 40% of the total exam marks); and
- if the total mark is at least 50%; and
- if you satisfy all the core sections of at least two of the assignments.

Changes Made to Previous Offerings

We try to adapt this unit to new developments in the area of natural language processing, and in response to feedback from students from past years.

Compared with last year, the contents of the Semantic Web has been reduced from five to three weeks. The contents on Machine Learning, while this year will take more relative space in the unit, will be roughly the same as last year, with the addition to one section on clustering.

In terms of software, the major change is in the use of Python 3 instead of Python 2. Python 3 has been available for several years but it is not backwards compatible with Python 2 and some of the libraries that we used in the past were not compatible with Python 3. Most of the key libraries are now compatible with Python 3, but note that code that was released in past offerings of the unit might not run on Python 3.