University of Macau
Faculty of Science and Technology
Department of Computer and Information Science
SFTW462 – Introduction to Natural Language Processing
Syllabus
1st Semester 2013/2014
Part A – Course Outline

Elective course in Computer Science

Course description:
(2-2) 3 credits. This course introduces fundamental concepts and skills associated with the design and implementation of different natural language processing systems covered from morphology, syntax and semantics. The main topics include regular expressions, (weighted) minimum edit distance, language modeling, Nävie Bayes (generative model), maximum entropy (discriminative model), text classification, sequence labeling, POS tagging, syntax parsing and computational lexical semantics. The course also includes an overview of practical natural language processing applications.

Course type:
Theoretical with substantial laboratory/practice content

Prerequisites:
- MATH111

Textbook(s) and other required material:
- Dan Jurafsky, and James H. Martin. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Pearson International Edition.

Reference:
- Steven Bird, Ewan Klein, and Edward Loper. (2009). *Natural language processing with Python*. O'reilly.

Major prerequisites by topic:
- Programming algorithms and formal structures.
- Basic knowledge in artificial intelligence.
- Basic familiarity with logic, linear algebra, probability theory.
- Mathematical principals in analyzing and problem modeling.

Course objectives:
- Learn the fundamental concepts, models, algorithms, and techniques. [a, e, k]
- Review basic knowledge of probability, formal language, computational linguistics, and programming skills. [a, e]
- Introduce engineering issues involved in the analysis and design natural language processing systems. [a, c, e]
- Practice of the techniques used in building natural language systems. [a, c, e, k]
- Appreciate the complexities of natural language. [a, c, e]

Topics covered:
- **Basic Concepts (2 hours)**: Introduce fundamental knowledge of natural language processing (NLP), and different analytical tasks at the morphology, part-of-speech (POS), syntactic structure and word sense. Discuss the problem of language ambiguities, and review the models and algorithms used in processing natural language.
- **Text Processing (4 hours)**: Introduce the fundamental techniques of text processing and string similarity measure, including regular expression, sentence segmentation, word tokenization, normalization and (weighted) minimum edit distance for string alignment. Those are the basic techniques that used in the first step for text preprocessing.

- **Probabilistic Models (8 hours)**: Introduce *N*-grams, Nävie Bayes, and Maximum Entropy Models, which are commonly used in language processing. Probabilistic models are crucial for capturing every kind of linguistic knowledge, and can be used to augment state machines and formal rule systems to solve many kinds of ambiguity problems.
- **Morphological Analysis (4 hours)**: Introduce the tasks of morphological analysis and part-of-speech tagging. Study the relevant algorithms and problem-solving techniques in morphological analysis.
- **Syntactic Parsing (6 hours)**: Study the fundamental concepts in syntax through the use of declarative formalisms: context-free grammars and dependency grammars. Learn parsing algorithms that employ grammars to automatically assign a syntactic structure to an input sentence.
- **Lexical Semantic (4 hours)**: Study the representation of meaning. Concern the issues of meaning that associated with lexicon, and introduce a computational problem of word sense disambiguation.
- **Applications (2 hours)**: Show how language-related algorithms and techniques can be applied to important real-world problems. This includes spelling checking and correction, text classification, named entity recognition, sentiment analysis, POS tagging and syntactic parsing.

Class/laboratory schedule:

Timetabled work in hours per week			No of teaching weeks	Total hours	Total credits	No/Duration of exam papers
Lecture	Tutorial	Practice				
2	2	Nil	14	56	3	1 / 3 hours

Student study effort required:

Class contact:	
Lecture	28 hours
Tutorial	28 hours
Other study effort	
Self-study	24 hours
Homework assignment	8 hours
Project / Case study	15 hours
Total student study effort	103 hours

Student assessment:
Final assessment will be determined on the basis of:

Homework	10%	Project	20%
Midterm	30%	Final exam	40%

Course assessment:
The assessment of course objectives will be determined on the basis of:
- Homework, project and exams
- Course evaluation

Course outline:

Weeks	Topic	Course work
1	**Introduction** Concepts of natural language processing (NLP), layers of language processing, morphology, part-of-speech, phrase structure and syntax tree, lexicon semantic, linguistic and computational issues	
2-3	**Text Processing** Regular expression, sentence segmentation, word tokenization and normalization, string matching, alignments, minimum edit distance, weighted minimum edit distance	Assignment#1
4-5	**Language Modeling** Probability foundations, noise channel, maximum likelihood estimation, model evaluation - perplexity, smoothing techniques, spelling checking and correction	Project Task #1

Weeks	Topic	Course work
6-7	**Classification Models** Generative and discriminative models, Näive Bayes, feature-based models, maximum entropy model, sequence labeling model	Assignment#2
8	**Text Classification** Classification algorithms, information extraction, named entity recognition and classification, sentiment analysis, feature selection, learning and evaluation	Project Task #2
9	**Part-Of-Speech (POS) Tagging** Word class, POS disambiguation, maximum entropy Markov model	Assignment#3 Midterm exam
10-12	**Syntax Parsing** Context-free grammar, dependency grammar, parsing strategy, statistical CYK parsing	Project Task #3
13	**Lexical Semantic** Representation of meaning, word sense relations, word sense disambiguation	Assignment#4
14	**Project Demonstration**	

Contribution of course to meet the professional component:
This course prepares students to work professionally in the area of human language processing.

Relationship to CS program objectives and outcomes:
This course primarily contributes to the Computer Science program outcomes that develop student abilities to:
(a) an ability to apply knowledge of mathematics, science, and engineering.
(c) an ability to design a system, component, or process to meet desired needs within realistic constraints such as economic, environmental, social, political, ethical, health and safety, manufacturability, and sustainability.
(e) an ability to identify, formulate, and solve engineering problems.
(k) an ability to use the techniques, skills, and modern engineering tools necessary for engineering practice.

Relationship to CS program criteria:

Criterion	DS	PF	AL	AR	OS	NC	PL	HC	GV	IS	IM	SP	SE	CN
Scale: 1 (highest) to 4 (lowest)	4		2							1	3		2	

Discrete Structures (DS), Programming Fundamentals (PF), Algorithms and Complexity (AL), Architecture and Organization (AR), Operating Systems (OS), Net-Centric Computing (NC), Programming Languages (PL), Human-Computer Interaction (HC), Graphics and Visual Computing (GV), Intelligent Systems (IS), Information Management (IM), Social and Professional Issues (SP), Software Engineering (SE), Computational Science (CN).

Course content distribution:

Percentage content for			
Mathematics	Science and engineering subjects	Complementary electives	**Total**
10%	80%	10%	100%

Persons who prepared this description:
Dr. Fai Wong, Dr. Sam Chao

Part B – General Course Information and Policies

1st Semester 2013/2014

Instructor:	Dr. Fai Wong	Office:	R108
Office hour:	Mon ~ Fri 15:00 – 18:00, or by appointment	Phone:	8397 8051
Email:	derekfw@umac.mo		

Time/Venue: Mon 11:00 – 13:00, WLG113 (lecture)
Wed 14:00 – 16:00, RLG302 (tutorial)

Grading distribution:

Percentage Grade	Final Grade	Percentage Grade	Final Grade
100 - 93	A	92 - 88	A–
87 - 83	B+	82 - 78	B
77 - 73	B–	72 - 68	C+
67 - 63	C	62 - 58	C–
57 - 53	D+	52 - 50	D
below 50	F		

Comment:
The objectives of the lectures are to explain and to supplement the text material. Students are responsible for the assigned material whether or not it is covered in the lecture. Students who wish to succeed in this course should read the textbook prior to the lecture and should work all homework and project assignments. You are encouraged to look at other sources (other texts, etc.) to complement the lectures and text.

Homework policy:
The completion and correction of homework is a powerful learning experience; therefore:
- There will be approximately 4 homework assignments.
- Homework is due one week after assignment unless otherwise noted, no late homework is accepted.
- The course grade will be based on the average of the HW grades.

Course project:
The project is probably the most exciting part of this course and provides students with meaningful experiences to design and implement an NLP system in the course:
- The application domain will be discussed further in class.
- The project will be presented at the end of the semester.

Exams:
One midterm exam will be held during the semester. Both the midterm and final exams are closed book, 2-hour examinations. There will be occasional in-class assignment.

Note:
- Check UMMoodle (https://ummoodle.umac.mo/) for announcement, homework and lectures. Report any mistake on your grades within one week after posting.
- No make-up exam is given except for CLEAR medical proof.
- Cheating is absolutely prohibited by the university.

Appendix:

Rubric for Program Outcomes

Rubric for (a)	**5 (Excellent)**	**3 (Average)**	**1 (Poor)**
Understand the theoretic background	Students understand theoretic background and the limitations of the respective applications.	Students have some confusion on some background or do not understand theoretic background completely.	Students do not understand the background or do not study at all.
Rubric for (c)	**5 (Excellent)**	**3 (Average)**	**1 (Poor)**
Design capability and design constraints	Student understands very clearly what needs to be designed and the realistic design constraints such as economic, environmental, social, political, ethical, health and safety, manufacturability, and sustainability.	Student understands what needs to be designed and the design constraints, but may not fully understand the limitations of the design constraints.	Student does not understand what needs to be designed and the design constraints.
Rubric for (e)	**5 (Excellent)**	**3 (Average)**	**1 (Poor)**
Identify applications in engineering systems	Students understand problem and can identify fundamental formulation.	Students understand problem but cannot apply formulation, or cannot understand problem.	Students cannot identify correct terms for engineering applications.
Rubric for (k)	**5 (Excellent)**	**3 (Average)**	**1 (Poor)**
Use modern principles, skills, and tools in engineering practice	Student applies the principles, skills and tools to correctly model and analyze engineering problems, and understands the limitations.	Student applies the principles, skills and tools to analyze and implement engineering problems.	Student does not apply principles and tools correctly and/or does not correctly interpret the results.