CM3060 Natural Language Processing

Course Description

Natural language processing (NLP) involves machines processing and extracting information from natural human languages. NLP is a crucial target for the application of computer science techniques. It consists of a range of specialised techniques that researchers are developing in the significant and growing field of Natural Language Processing. By taking this module, you will gain a solid grasp and practical experience with those techniques. The module complements other modules in the programme which involve the processing and interpretation of data by machines.

This module will provide you with a grounding in both rule-based and statistical approaches to NLP, and it combines theoretical study with hands-on work employing widely used software packages. The module focuses on text processing, and by taking this module you will learn about how you can work with text based natural language in your computer programs. You will learn about grammars and how they can be used to analyse text. You will learn how statistical analysis can be used to extract information from and classify text. You will work in an appropriate programming environment for NLP, using libraries to implement NLP workflows.

Course Goals and Objectives

Upon successful completion of this course, you will be able to:

- Explain differences between rule-based and statistical approaches to NLP, and evaluate their relative merits
- Select appropriate statistical language analysis techniques for a particular problem
- Utilize software tools such as corpus readers, stemmers, taggers and parsers and carry out analysis of existing texts by writing software using existing NLP libraries
- Define formal grammars for fragments of a natural language
- Evaluate applications of statistical techniques to natural language analysis such as classification, information extraction and probabilistic parsing

Textbook and Readings

Specific essential readings for each week from the following list are included in the Readings page for each week:

- Bird, Steven, Ewan Klein, and Edward Loper. Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc., 2009.
- Jurafsky, Dan, and James H. Martin. "Speech and Language Processing (3rd draft ed.)." (2019).
- Perkins, Jacob. Python 3 text processing with NLTK 3 cookbook. Packt Publishing Ltd, 2014.
- Python Natural Language Processing Cookbook: Over 50 recipes to understand, analyze, and generate text for implementing language processing tasks, Zhenya Antić, Packt Publishing Ltd, 2021
- Provost, Foster, and Tom Fawcett. Data Science for Business: What you need to know about data mining and data-analytic thinking. O'Reilly Media, Inc., 2013.
- Schütze, Hinrich, Christopher D. Manning, and Prabhakar Raghavan. Introduction to information retrieval. Vol. 39. Cambridge: Cambridge University Press, 2008.
- Hovy, Dirk. Text Analysis in Python for Social Scientists: Discovery and Exploration. Cambridge University Press, 2020.
- VanderPlas, Jake. Python data science handbook: Essential tools for working with data. O'Reilly Media, Inc., 2016.

Course Outline

The course consists of ten topics divided into 20 weeks that focus on key concepts.

Topic 1: Introduction to NLP	**Key concepts:**
	This week we examine the various paradigms and disciplines contributing to NLP and take a look at how the field has evolved over the decades.
	Learning outcomes:
	- Understand the scope and impact of NLP
	- Explore the development

	environment • Describe the evolution of NLP approaches
Topic 2: Basic text processing	**Key concepts:** This week we introduce some basic techniques for text processing, such as tokenization and various forms of normalisation. **Learning outcomes:** • Manipulate unstructured data • Apply text processing techniques • Understand text processing fundamentals
Topic 3: Language modelling	**Key concepts:** This week we introduce the topic of language modelling and learn how to represent text using simple frequency distributions. **Learning outcomes:** • Perform topic modelling on language data • Understand how to statistically model natural language • Perform basic statistical analyses on language data
Topic 4: Lexical semantics	**Key concepts:** This week we learn how word meanings can be represented using resources such as WordNet and explore how to use semantic similarity techniques to build useful NLP applications. **Learning outcomes:** • Apply semantic similarity techniques • Analyse curated and distributed word representations • Understand how word meanings are represented
Topic 5: Text categorisation and sentiment analysis	**Key concepts:** This week we explore the theory underpinning generative approaches to text categorisation and experiment with applying this to real world data.

	Learning outcomes: • Evaluate text categorisation techniques • Apply sentiment analysis techniques • Understand the fundamentals of text categorisation
Topic 6: Syntax and parsing	**Key concepts:** This week we examine some of the linguistic principles that shape human language and learn how to apply different syntactic approaches and formalisms. **Learning outcomes:** • Understand probabilistic approaches to parsing • Apply practical syntax analysis techniques • Understand the fundamentals of grammars and parsing
Topic 7: Information extraction	**Key concepts:** This week we learn how to apply various NLP techniques to extract useful information from unstructured data. **Learning outcomes:** • Understand the definition and scope of information extraction • Apply entity recognition techniques • Create practical information extraction applications
Topic 8: Information retrieval	**Key concepts:** This week we explore how NLP ideas are used in constructing search engines and study some of the foundational models for representing text. **Learning outcomes:** • Understand Information Retrieval fundamentals • Manipulate data structures for information retrieval • Apply information retrieval models and principles
Topic 9: Chatbots and dialogue systems	**Key concepts:**

	This week we study the fundamental principles of human conversation and learn how to apply them in building chatbots and dialogue systems. **Learning outcomes:** • Understand the properties of human conversation • Anayse dialogue system architectures • Create simple chatbots
Topic 10: NLP in practice	**Key concepts:** This week we interview various NLP practitioners and examine how they have used their NLP knowledge to build practical NLP systems. **Learning outcomes:** • Compare and contrast different contexts for NLP practice • Gain insight into the challenges faced by NLP practitioners • Understand how NLP concepts and principles are applied in industry

Learning Activities of This Course

The course is comprised of the following elements:

Lecture videos. In each week the concepts you need to know will be presented through a collection of short video lectures. You may stream these videos for playback within the browser by clicking on their titles or download the videos. You may also download the slides that go along with the videos.

Readings. Each topic may include several suggested readings. These are a core part of your learning and, with the videos, will cover all the concepts you need for this course.

Practice Quizzes. Each week will include one or more practice quizzes, intended for you to assess your understanding of the topics. You will be allowed unlimited attempts at each practice quiz. Each attempt may present a different selection of questions to you. There is no time limit on how long you take to complete each attempt at the quiz. These quizzes do not contribute toward your final score in the module.

Discussion Prompt. Each week includes one or more discussion prompts. You will see the discussion prompt alongside other items in the lesson. Each prompt provides a space for you to respond. After responding, you can see and comment on your peers' responses.

How to Pass This Module

The module will contain a range of summative and formative assessments. Summative assessments are assessments which contribute directly towards your final grade. Formative assessments do not count directly towards your final grade. Instead, they provide you with opportunities for low stakes practice and will often provide some sort of feedback about your progress. For example, a practice quiz might provide you with feedback about why a particular answer was wrong.

The module has two assessments: a midterm assessment and a final exam. The midterm assessment comprises a project. The exam comprises a quiz section and a written section. The midterm and exam each carry 50% of the grade.

This is a detailed breakdown of all the marks:

Activity	Required? (Summative)	Deadline week	Estimated time per course	% of final grade
Written, staff graded coursework	Yes	Approximately week 13	25 hours	50%
Written examination	Yes	Approximately week 22	3-4 hours	50%