Natural Language Processing

Module description

Target Audience: Students enrolled on MSc Data Science programme who are interested in learning about machine processing of natural language that is a key target for the application of Data Science techniques.

Content Differentiation: Natural language processing (NLP) involves machines processing and extracting information from natural human languages. NLP is a crucial target for the application of data science techniques. It consists of a range of specialized 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.

Module goals and objectives

Upon successful completion of this module, 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.

Module outline

The module consists of ten topics that focus on key areas of the fundamentals of computer science.

Topic 1.	**Key concepts:** • Alternative paradigms within NLP • NLP toolkits and libraries • Evaluation in NLP **Learning outcomes:** • Understand the scope and impact of NLP • Explore the development environment • Describe the evolution of NLP approaches

Topic 2.	**Key concepts:** • Word and sentence tokenization • Text normalization • Text corpora **Learning outcomes:** • Understand text processing fundamentals • Apply text processing techniques • Manipulate unstructured data
Topic 3.	**Key concepts:** • Word frequency distributions • ngram language models and perplexity • Topic models **Learning outcomes:** • Perform basic statistical analyses on language data • Understand how to statistically model natural language • Perform topic modelling on language data
Topic 4.	**Key concepts:** • Lexical semantics representations • Word embeddings • Similarity metrics **Learning outcomes:** • Understand how word meanings are represented

	• Analyse curated and distributed word representations • Apply semantic similarity techniques
Topic 5.	**Key concepts:** • Supervised classification • Feature extraction and selection • Sentiment lexicons **Learning outcomes:** • Understand the fundamentals of text categorization • Apply sentiment analysis techniques • Evaluate text categorization techniques
Topic 6.	**Key concepts:** • Context-free grammars • Dependency grammars • Probabilistic parsing **Learning outcomes:** • Understand the fundamentals of grammars and parsing • Apply practical syntax analysis techniques • Understand probabilistic approaches to parsing
Topic 7.	**Key concepts:** • Named entities • Relation extraction • Information extraction pipelines

	Learning outcomes: • Understand the definition and scope of information extraction • Apply entity recognition techniques • Create practical information extraction applications
Topic 8.	**Key concepts:** • Boolean search • Vector space models • Query expansion **Learning outcomes:** • Understand IR fundamentals • Analyse IR data structures • Apply IR techniques and principles
Topic 9.	**Key concepts:** • Speech acts & grounding • Dialog system architectures • Frames and slot filling **Learning outcomes:** • Understand the properties of human conversation • Anayse dialog system architectures • Create simple chatbots

Topic 10.	**Key concepts:** • NLP skills and competencies • Natural language engineering • NLP trends and developments **Learning outcomes:** • Understand how NLP concepts and principles are applied in industry • Gain insight into the challenges faced by NLP practitioners • Compare and contrast different contexts for NLP practice

Activities of this module

The course is comprised of the following elements:

- Lecture videos introduce the main concepts of the topics and illustrate them with examples
- Practice quizzes will be used to reinforce your learning and understanding
- Activities drive the work that you do for each topic, where you are asked to solve challenges of different types
- Graded assignments include a practical coursework assignment and a written exam.
- Discussions with your peers will help to guide your work and encourage you to explore different types of solutions to problems
- Readings will help to reinforce your learning of concepts.

How to pass this module

The module has two major assessments each worth 50% of your grade:

- Coursework: this will be assessed midway through the course. The coursework comprises a variety of exercises which in total will take up to 25 hours of study time to complete.

- The examination will be two hours long and consist of multiple-choice questions and longer written answers.

Activity	Required?	Deadline week	Estimated time per course	% of final grade
Written, staff graded coursework	Yes	12	Approximately 25 hours	50%
Written examination	Yes	20	2 hours	50%