

Contents

Preface

About this half unit	
Assessment	
The subject guide and other learning resources	
Suggested study time	
Acknowledgement	

1 Introduction: how to use this subject guide

1.1 Introduction	
1.2 Aims of the course	
1.3 Learning outcomes	
1.4 Reading list and other learning resources	
1.5 Software requirements	
1.6 How to use the guide/structure of the course	
1.6.1 Chapter 2: Introducing NLP: patterns and structures in language	
1.6.2 Chapter 3: Getting to grips with natural language data	
1.6.3 Chapter 4: Computational tools for text analysis	
1.6.4 Chapter 5: Statistically-based techniques for text analysis	
1.6.5 Chapter 6: Analysing sentences: syntax and parsing	
1.6.6 Appendices	
1.7 What the course does not cover	

2 Introducing NLP: patterns and structure in language

Essential reading	
Recommended reading	
Additional reading	
2.1 Learning outcomes	
2.2 Introduction	
2.3 Basic concepts	
2.3.1 Tokenised text and pattern matching	
Activity: Recognising names	
2.3.2 Parts of speech	
Activity: identify parts of speech	
2.3.3 Constituent structure	
Activity: Writing production rules	
2.4 A closer look at syntax	
2.4.1 Operation of a finite-state machine	
Activity: Finite-state machines	
2.4.2 Representing finite-state machines	
2.4.3 Declarative alternatives to finite-state machines	
Activity: Coding regular expressions	
Activity: tree diagrams for a regular language	
2.4.4 Limitations of finite-state methods – introducing context-free grammars	
Activity: Regular grammars	
Activity: Context-free grammar	
2.4.5 Looking ahead: some further uses of regular expressions	

2.4.6	Looking ahead: grammars and parsing	
2.5	Word structure	
	Activity: Past tense formation	
2.6	A brief history of natural language processing	
2.7	Summary	
2.8	Sample examination questions	
3	Getting to grips with natural language data	
	Essential reading	
	Recommended reading	
	Additional reading	
3.1	Learning outcomes	
3.2	Using the Natural Language Toolkit	
3.3	Corpora and other data resources	
3.4	Some uses of corpora	
3.4.1	Lexicography	
3.4.2	Grammar and syntax	
3.4.3	Stylistics: variation across authors, periods, genres and channels of communication	
3.4.4	Training and evaluation	
3.5	Corpora	
3.5.1	Brown corpus	
3.5.2	British National Corpus	
3.5.3	COBUILD Bank of English	
3.5.4	Penn Treebank	
3.5.5	Gutenberg archive	
3.5.6	Other corpora	
	Activity: Online corpus queries	
3.5.7	WordNet	
3.6	Some basic corpus analysis	
3.6.1	Frequency distributions	
	Activity: Using NLTK tools	
3.6.2	DIY corpus: some worked examples	
	Activity: building and analysing a DIY corpus	
3.7	Summary	
3.8	Sample examination question	
4	Computational tools for text analysis	
	Essential reading	
	Recommended reading	
	Additional reading	
4.1	Introduction and learning outcomes	
4.1.1	Learning outcomes	
4.2	Data structures	
	Activity: strings and sequences	
4.3	Tokenisation	
4.3.1	Some issues with tokenisation	
4.3.2	Tokenisation in the NLTK	
	Activity: Tokenising text	
4.4	Stemming	
	Activity: Comparing stemmers	
4.5	Tagging	
4.5.1	RE tagging	
	Activity: Tagging with REs	
4.5.2	Trained taggers and backoff	

4.5.3	Transformation-based tagging	
4.5.4	Evaluation and performance	
	Activity: Trained taggers	
4.6	Summary	
4.7	Sample examination question	

5 Statistically-based techniques for text analysis

	Essential reading	
	Recommended reading	
	Additional reading	
5.1	Learning outcomes	
5.2	Introduction	
5.3	Some fundamentals of machine learning	
5.3.1	Naive Bayes classifiers	
	Activity: Bayes' rule	
5.3.2	Hidden Markov models	
5.3.3	Information and entropy	
5.3.4	Decision trees and maximum entropy classifiers	
	Activity: further reading	
5.3.5	Evaluation	
5.4	Machine learning in action: document classification	
5.4.1	Summary: document classification	
	Activity: document classification	
5.5	Machine learning in action: information extraction	
5.5.1	Types of information extraction	
5.5.2	Regular expressions for personal names	
	Activity: coding regular expressions for proper names	
5.5.3	Information extraction as sequential classification: chunking and NE recognition	
	Activity: chunking and NE recognition	
5.6	Limitations of statistical methods	
5.7	Summary	
5.8	Sample examination question	

6 Analysing sentences: syntax and parsing

	Essential reading	
	Recommended reading	
	Additional reading	
6.1	Learning outcomes	
6.2	Grammars and parsing	
6.3	Complicating CFGs	
6.3.1	Verb categories	
	Activity: Verb categories	
6.3.2	Agreement	
	Activity: feature-based grammar	
6.3.3	Unbounded dependencies	
6.3.4	Ambiguity and probabilistic grammars	
	Activity: probabilistic grammar	
6.4	Parsing	
6.4.1	Recursive descent parsing	
6.4.2	Shift-reduce parsing	
6.4.3	Parsing with a well-formed substring table	
6.4.4	Finite-state machines and context-free parsing	
	Activity: Parsing	
6.5	Summary	

6.6	Sample examination question	
A	Bibliography	
B	Glossary	
C	Answers to selected activities	
	Chapter 2: Introducing NLP: patterns and structure in natural language . . .	
	Identify parts of speech, page 14	
	Operation of a finite-state machine, page 17	
	Coding regular expressions, page 19	
	Regular grammars, page 21	
	Past tense forms, page 25	
	Chapter 3: Getting to grips with natural language data	
	Online corpus queries, page 37	
	Using NLTK tools, page 39	
	Chapter 4: Computational tools for text analysis	
	Comparing stemmers, page 48	
	Tagging with REs, page 51	
	Chapter 5: Statistically-based techniques for text analysis	
	Activity: Bayes' Rule, page 59	
	Chapter 6: Analysing sentences: syntax and parsing	
	Activity: Verb categories, page 78	
	Activity: Feature-based grammar, page 80	
D	Trace of recursive descent parse	
E	Sample examination paper with answering guidelines	
	E.1 Sample examination questions	
	E.2 Answering guidelines for sample examination questions	