Natural Language Processing with Python, by Steven bird, Ewan Klein and Edward Loper, first edition, 2009. O'Reilly Media, Inc, ISBN 978-0-596-51649-9.

Reviewed by: Ripan Hermawan, Indonesia University of Education, Indonesia.

Natural Language Processing (NLP) is a 'theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications' (Liddy, 2001, p.1). NLP is an increasingly popular field of study nowadays. The applications of NLP are widespread. There are many applications made possible by NLP such as translation tools, speech recognition and transcription softwares, speech to text and text to speech applications, corpus tools, predictive text applications, and many more. These in return have contributed to the increase in the popularity of NLP itself, not only among academics but also among wider audiences who have benefited from it.

This book offers a highly accessible introduction to the study of NLP (p. ix). The intended readers are those who want to know about the fundamental concepts and applications of NLP. Since no prior programming experience is required, the book is suitable for linguistic researchers and students who want to learn how to work with large amounts of linguistic data. What makes the book interesting is that it takes practical, principled, pragmatic and pleasurable approaches (p. x). Practicality is realized by providing real examples and giving the readers the opportunity to write their own programs and to see how an idea is implemented and tested. The principled approach is carried out through a coverage of both theoretical foundations and careful linguistic and computational analysis. The pragmatic approach is implemented through the balanced discussions on both theoretical aspects and practical aspects. Since learning programming will involve a significant amount of time for coding (writing a code), the book has excelled itself by taking a pleasurable approach through the provision of many examples of interesting and entertaining real applications of NLP and the real codes from which the readers can learn and play. Moreover, it also provides graded exercises at the end of every chapter to help readers practice what they have learned from the respective chapter.

The book consists of eleven chapters, each of which develops on the basis of what has come before. The discussion starts with an introduction to the general concept of NLP. It then proceeds with a gentle introduction to Python, the default programming language implemented, and its famous Natural Language Toolkit (NLTK) package that provides hundreds of NLP functionalities. The next topics discussed are fundamentals of NLP such as tagging, classification and information extraction. Parsing and identification of sentence structures are the next topics of discussion, followed by a discussion on constructing representations of meaning. The last topics are on linguistic data and how to manage them.

Chapter 1 starts with an instruction on how to check if Python has been installed in a computer, followed by instruction on how to install NLTK and have the data that accompany it installed in the system. The chapter continues with some demonstrations of text analysis using Python such as counting the number of words in a text, finding concordance of certain words as well as generating simple statistics such as frequency distribution and conditional frequency distribution of a text.

Chapter 2 introduces to the readers some corpora that come with NLTK and shows how Python can be used to mine relevant information from the corpora. Some of the corpora are Brown corpus, Gutenberg corpus, Reuters Corpus, Inaugural Address Corpus (Inaugural speeches of American presidents). In Chapter 3, the readers are given the opportunity to do text analysis using their own texts that they have

in their local disks or any on-line texts available in the web pages of their interest. This chapter also introduces ways to normalize text through Stemming (finding out the stem of a word) and Lemmatization using off-the-shelf stemmers, such as PorterStemmer and LancasterStemmer and lemmatizers such as WordNet lemmatizer.

Chapter 4 is about writing structured programs using Python. Some features of the language such as assignment, conditionals and sequences are discussed briefly. Most of the chapter however is dedicated to the discussion of function since it is a fundamental aspect of any programming language with which one can write a reusable structured program.

In chapter 5, the process of categorizing and tagging words is explained. Tagging is the process of labeling words using some predefined tagset, usually the Part of Speech (POS) tags. NLTK comes with some POS taggers that can automatically assign tags such as Noun, Verb and Adjective to the words in a given text. Ways of measuring the performance of a tagger in terms of its accuracy level are provided to let the readers choose the tagger that best suits their needs.

Chapter 6 is dedicated to a discussion of how to do text classification based on the distinctive features of the texts using supervised classification. Examples of the classification, among others, are gender identification based on names, deciding on the topic of a given article and deciding whether or not an incoming email is a Spam. Some models of classification and ways of evaluating them are explained here.

Chapter 7 deals with ways of extracting pieces of relevant information from texts. If you are always wondering how an intelligent software can detect names of people, places, time and dates as well as other important information from a text then you should know that it is name entity recognition (NER) that does the job. Once the NER is done, it is then possible to analyze relations between the entities.

Chapter 8 and 9 are devoted to a discussion of the analysis of sentence structure. The main issue here is how a grammar model can correctly solve ambiguities in natural language sentences to discover the intended meaning. In chapter 8, some models of formal grammar such as Context-Free Grammar (CFG) and Dependency Grammar (DG) are presented as frameworks for the analysis. Examples of parsing using these models are given to make the idea clear. Chapter 9 provides a way to improve the models given in chapter 8 in order to be able to handle extensive grammatical constructions that are hard to handle using only the models. This is carried out using the concept of grammatical features (Feature-Based Grammar).

Chapter 10 tackles the issue of meaning both in individual sentences and in a discourse as a sequence of sentences. Logics such as Propositional Logic and First Order Logic is used here as a tool to get the "truth" out of sentences. Discourse Representation Theory (DRT) as a way to find out meaning of a discourse is explained together with examples of analysis using the theory.

Chapter 11 covers the management of linguistic data. The questions on how to design a well-structured corpus that is balanced and supports a wide range of uses, how to convert existing data with wrong format to a suitable one and how to manage existing resources and enable others to easily find them constitute the driving forces of the chapter. Information on how to legally obtain required data is also presented to help readers who may want to create a corpus of their own. Finally, Extensible Markup Language (XML), a popular markup language, is introduced as a framework for representing annotated text and for lexical resources. XML has been chosen because, unlike HTML, it allows its users to define their own tags rather than using rigid predefined tags.

The book, in my opinion, serves its purposes very well. I would recommend the book to anyone seriously interested in any linguistic research that uses corpus linguistics as the methodology (Meyer, 2002). Both the concepts and practices introduced in the book provide powerful tools for linguistic research.

References

Meyer, C.F. (2002) English corpus linguistics: An introduction. Cambridge: Cambridge University Press.

Liddy, E.D. (2001). Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.