|  |
| --- |
| [Company name] |
| Language Model |
| ICS2203 Natural Language Processing: Methods and Tools Project Task 1 |

|  |
| --- |
| Luke Ellul  [Date] |

Contents

Corpus Processing

The English corpus[[1]](#footnote-1) was used for this language model. The directory **corpus-cleaner** contains **cleaner.js** which includes the code responsible for parsing the xml files of the corpus and deriving the sentences. The derived sentences are saved as plain text files in the directory **cleanTexts** which are then processed by the language modeler.

To derive the appropriate text data from the xml files **cheerio[[2]](#footnote-2)** was used. Cheerio is a Node.js module which is primarily used to scrape content from html pages but can also be used for xml.

1. <http://ota.ox.ac.uk/desc/2554> [↑](#footnote-ref-1)
2. https://www.npmjs.com/package/cheerio [↑](#footnote-ref-2)