School of Computing and Information Systems
The University of Melbourne
COMP90042
NATURAL LANGUAGE PROCESSING (Semester 1, 2022)

Workshop exercises: Week 2

Discussion

1. Give some examples of text processing applications that you use on a daily basis.
2. What is **tokenisation** and why is it important?
   (a) What are **stemming** and **lemmatisation**, and how are they different? Give examples from the `01-preprocessing` iPython notebook.

Programming

1. Make sure that you have a Python environment where you can run the given iPython notebooks. In particular, ensure that the `numpy`, `sklearn` and `nltk` packages are installed (i.e. you can `import` them).
2. Adapt the `01-preprocessing` iPython notebook into a program which tokenises an input file based on the five–step model given in the lectures.
3. Complete the BPE tokenisation algorithm in the `02-bpe` iPython notebook.

Catch-up

- Revise the following terms, as they are used in a text processing context: "corpus"; "document"; "term"; "token".
- Revise "stop words", and why they are often removed from a text in a text processing/information retrieval context. Use the Web to find a list of stop words for English — are there any words in the list that you might consider not to be a stop word? Are there any words that you consider to be stop words that are missing from the list?
- Recall the most common regular expression operators; practice writing some regular expressions to solve common text processing problems.
- (Re-)familiarise yourself with Python, if you haven't used it recently. In particular, focus on string and array processing, including regular expressions. Also revise functions and mapping mathematical formulae to Python syntax (including the `numpy` package).
- Familiarise yourself with the Natural Language Toolkit (NLTK). You might like to use the e-book `http://nltk.org/book` as a resource; it also covers some of the basics of Python, if you've never used the language before.

Get ahead

- (Extension) Identify some tokenisation issues in a language (other than English) of your choice. How much alteration would need to be made to the tokenisation strategy from the lectures to account for these issues?