COMP90042 - Natural Language Processing

Workshop Week 2

Biaoyan Fang

9 March 2020

Contract

- Canvas - Discussions
  - https://canvas.lms.unimelb.edu.au
- Subject Co-ordinator
  - Jey Han Lau
  - laujh@unimelb.edu.au
- Me
  - Biaoyan Fang
  - biaoyan@unimelb.edu.au

Programming

- Python 3
- Jupyter notebook (Anaconda 3)
- Packages
    1. NLTK, gensim
    2. Matplotlib, Numpy, Scipy
    3. Scikit-learn

Outline

1. Introduction
2. Pre-processing
3. Byte-Pair Encoding

Introduction

Applications of NLP:

- Search engines
- Translation
- Speech-to-text systems
- Spelling correction
- ...

Definitions

- Corpus: a collection of documents.
- Document: one or more sentences.
- Sentence
  - “The student is enrolled at the University of Melbourne.”
- Words
  - Sequence of characters with a meaning and/or function
- Word token: each instance of “the” in the sentence above.
  - 9 word tokens in the example sentence.
- Word type: the distinct word “the”.
  - Lexicon (“dictionary”): a group of word types.
  - 8 word types in the example sentence.

Pre-processing

Pipeline (notebooks)

- Formatting
- Sentence Segmentation
- Tokenisation
- Normalisation
  - Lemmatisation
  - Stemming
- Remove Stopwords

Lemmatisation & Stemming

Inflectional Morphology

- Grammatical variants:
  airline -> airlines
  speak -> speaking
  old -> older

Derivational Morphology

- Another word with different meaning:
  write -> writer
  write -> rewrite

Lemmatisation

Remove all inflections

Matches with lexicons

Product: Lemma

Stemming

Remove all suffixes

No matching required

Product: Stem

Byte-Pair Encoding (BPE)

- Subword Tokenisation
  - *Colourless green ideas sleep furiously ->*
    [colour] [less] [green] [idea] [s] [sleep] [furious] [ly]
- Core idea: iteratively merge frequent pairs of characters
- Advantages:
  - Data-informed tokenisation
  - Works for different languages
  - Deals better with unknown words

BPE

- Dictionary
  - [5] l o w _
  - [2] l o w e r _
  - [6] n e w e s t _
  - [3] w i d e s t _
- Vocabulary (characters)
  - l, o, w, _, e, r, n, s, t, i, d
- Vocabulary (iterations)
  - l, o, w, _, e, r, n, s, t, i, d, es, est, est_, lo, low, ne, new, newest_, low_

Discussion

Questions