PGR 210 - Natural Language Processing Part

Kristiania University College

By Huamin Ren

Huamin.ren@kristiania.no

Høyskolen Kristiania

Outline of week 41

- Text processing
- Visualization and analysis
- ***Word representation and Bag-of-Words***

Why we represent words as input?

sentence = "The brown fox is quick and he is jumping over the lazy dog"

The grammar and ordering of words definitely gives meaning to a sentence. What if we jumbled up the words? Would the sentence still make sense?

Words: ['quick', 'is', 'fox', 'brown', 'The', 'and', 'the', 'is', 'he', 'dog', 'lazy', 'jumping', 'over']

Høyskolen Kristiania

- ***Lexical units***: represented by morphemes, the smallest meaningful and syntactically correct unit of a language. Words are inherently a subset of these morphemes.
- ***lexicon*** is a complete vocabulary of these lexical units.
- A rich lexical corpus and database called WordNet, which has an exhaustive list of different lexical entities that are grouped into synsets based on semantic similarity (e.g., synonyms).

- The most popular tokenization techniques include sentence and word tokenization, which are used to break down a text document (or corpus) into sentences and each sentence into words.

Why we represent words as input?

- Arguably most important common denominator across all NLP tasks
- To perform well on most NLP tasks, we first need to have some notion of similarity and difference between words
- With word vectors, we can quite easily encode this ability in the vectors themselves

Høyskolen Kristiania

- English language: an estimated 13 million tokens
- Norsk?
- https://bora.uib.no/bora-xmlui/bitstream/handle/1956/20906/drthesis_BjarteJohansen_2019.pdf?sequence=1&isAllowed=y
- https://www.duo.uio.no/bitstream/handle/10852/59276/11/Teaching_NLTK_Norwegian.pdf

How we represent words as input?

- BoW model

Introduction to Bag-of-Words model

- The most simple vector space representational model for unstructured text?
- The Bag of Words model represents each text document as a numeric vector - each dimension is a specific word from the corpus and the value could be its frequency in the document, occurrence (denoted by 1 or 0), or even weighted values.
- Text1 = [‘ a ’, ‘aa’]
- Text2 =[‘a’, ‘b’]
- Word_list = [‘a’, ‘aa’, ‘b’]   bow_1 = [1 1 0] bow_2=[1 0 1]

$$tf(w,D) = f_{w_D}$$

where f_{wD} denoted frequency for word **w** in document **D**, which becomes the term frequency (tf).

- TF: term frequency
- IDF: inverse document frequency

Høyskolen Kristiania