Concepts in English Linguistics

Session 13: Natural Language Processing

Bag-of-words (BoW) model A document model that encodes the **terms** (i.e. distinct forms) of the document along with their frequency. A disadvantage is that this abandons word order, but an advantage is that it requires fewer dimensions than keeping track of all individual **tokens**, reducing the memory burden. Additionally, the members of this model may be straightforwardly sorted alphabetically or by frequency, and words irrelevant to a task (e.g. **stop words**) may be easily discarded.

Bigram A sequence of two tokens treated as one; see ***n*-gram**.

Lemma Dictionary headword form of a word. In **lemmatization**, this value is used to identify the token form as an instance of that headword. In some NLP tasks, it may simply replace the inflected form in the list of tokens, but in others it is kept alongside the token form, e.g. as a tuple `('walked', 'walk')`.

Lemmatization Assigning individual **tokens** a lemma or headword reference so all forms carry a reference to the word they represent. For instance, consider how distinct the various forms of the English verb *to be* are, and that software needs to be told that they all go together, unless you want it to find out using a learning algorithm.

***n*-gram** A sequence of *n* **tokens**, treated as a single token. For instance, we can divide a document up into **bigrams** by populating a list with each sequence of two tokens, and then by counting which sequences are particularly frequent we can decide to treat these popular sequences as single tokens in our regular tokenized document, so that we can more accurately model the meaning of "Mother Theresa" than we would if we treated each word in this sequence as carrying its own meaning. **Bigrams** and **trigrams** are the most frequently used *n*-grams.

Normalization (You should be able to define this task and explain some of the common aspects it involves, particularly with reference to premodern corpora. You should also have an understanding of at least one way of achieving normalization in Python. See Lane et al. §2.2.5.)

Sentiment A measure of the attitude expressed in a document, most commonly involving **valence** (i.e. a scale from positive to negative evaluation, as in a product review). In present-day corpora, sentiment may be inferred from self-labelled data sets such as product reviews (with stars) or social media posts (with emojis).

Sentiment lexicon An index scoring terms for their sentiment scores along one or more axes. With

the help of a sentiment lexicon for a given language, we can calculate the sentiment value of a document in that language by adding up the scores for the document's tokens.

Stemming In NLP, removing inflectional (and sometimes lexical) information to reduce a token to a string approximating its lexical stem, or even root, e.g. by removing Germanic genitives like "-es." This is a complex task that can hardly accommodate all word forms in a language, and it is accordingly often imprecise. The resulting "stem" is not always identical with what linguists might consider the stem.

Stop words A list of the **terms** in a document or corpus to treat differently, usually by discarding them from the list of terms. For a variety of NLP tasks it is customary to discard function words, i.e. words that serve grammatical functions but do not themselves carry lexical meaning, because these would top the frequency rankings, which is undesirable e.g. in topic modelling, where with the stop words included it would seem that the documents are primarily "about" these function words.

Term Also known as a word form or type, a term is a spelling as it occurs (or does not occur) in a document. It is typically used in the **bag-of-words model** and associated with counts. The phrase "the tortoise and the hare" counts five tokens, but four terms or types, one of which has an absolute count of two.

Token Individual occurrence of a word in a document. After tokenization, the document may be reconstructed by lining up all of its tokens in order. The phrase "the tortoise and the hare" counts five tokens. Compare **term**.

Tokenization (You should be able to define and discuss this task, with only the most basic issues one might run into. You should also be aware of at least one Python method used to this end.)

Trigram A sequence of three tokens treated as one; see ***n*-gram**.

Type See **term**.

Unigram An ***n*-gram** consisting of a single token.

Word sense disambiguation The undertaking of determining to which of two or more matching lemmas (dictionary headwords) a token form belongs. Usually relies on one or more features of the token context, among other things. When such approaches fail, an algorithm may fall back on such heuristics as "one sense per discourse" and assigning the most frequent sense.

Word type See **term**.

Zipf's Law A statistical phenomenon that states that given a set in a range of phenomena in the natural world or that of humans, the absolute frequency of each member is approximately one over its rank in the frequency table, multiplied by the frequency of the most frequent member. In other words, the most frequent member occurs about twice as often as the second most frequent member, and three times as often as the third-ranking member. This means the distribution of members resembles an exponential function. The relevance of this to the study of term frequency is that we can work in logarithmic space for a more linear account of word distribution,

which makes it easier to do arithmetic on words counts.