Concepts (final)

January 31, 2023

This document gathers together some of the concepts students of B.EP.301/B.Eng.602 will be expected to be able to discuss on the exam. It offers no guarantee that other concepts introduced in your readings or in class won't make an appearance, but it is intended to be reasonably comprehensive. Basic Python concepts, such as lists and dictionaries, are not here covered but are nevertheless required knowledge.

Arousal A scale of subjectively experienced intensity provoked by a word; one among three dimensions on which words are commonly scored for affective meaning (sentiment).

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**.

Bilingual lexical induction The task of determining the optimal translations in a target language for words from a source language. This usually involves comparing **word embeddings**.

Bitext A pair of documents representing the same text in two languages, usually because one is a translation of the other.

Case folding Reducing all characters in your document to the same type case, usually lowercase; thus e.g. `deirdre johnson drank dr. pepper`.

Distributional hypothesis The assumption that words with similar meanings tend to occur in similar contexts.

Document frequency The number of documents (in a given corpus) in which a term occurs. See **TF-IDF**.

Dominance A scale of subjectively experienced "control" exerted by a word; one among three dimensions on which words are commonly scored for affective meaning (sentiment).

Dot product A single value arrived at by multiplying the values of each column between two vectors, then adding up the results. This value represents the cosine of the angle between the vectors, and as such is a measure of their similarity.

General AI AI systems that are good at a wide range of tasks. Contrast **narrow AI**, and compare **strong AI**.

Gloss The definition of a word, e.g. in a dictionary.

Latent Dirichlet allocation (LDiA) A variant of **LSA** (which see) using a different set of probabilistic assumptions. (Will not appear on the exam.)

Latent semantic analysis (LSA) Using SVD to improve the spread of terms that co-occur in BoW or TF-IDF vectors, thus determining word and document topics algorithmically. See also **single value decomposition (SVD)**, **topic vector**.

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.

Lesk algorithm Approach to **word sense disambiguation** (WSD) that adduces its dictionary **gloss** (i.e. definition) and assigns the token to the lemma whose gloss has the greatest overlap with the token context.

Lexical diversity (You should be able to explain how Bird et al. arrive at this measure, and what the strengths and weaknesses of the concept as there defined are.)

Linear discriminant analysis (LDA) An algebraic operation to determine the vector between the centroids of two clusters (of TF-IDF vectors, when used for topic analysis, exactly one of which clusters matches the topic you're interested in). You can then decide on a threshold between the two poles and measure new data vectors against the discriminant vector to classify them.

Narrow AI AI systems that are good at a small selection of tasks, or at one thing only. Contrast **general AI**, and compare **weak AI**.

***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.)

One-hot encoding A document model that encodes each token as a binary string with exactly one `1` indicating the **token**'s position in the document. This has the advantage that no information about the document is lost (it can be easily reconstructed), but it requires a great deal of memory, and tokens cannot be compared across documents, as each document has its own encoding.

Overfitting Training your algorithm on your training data so closely that its ability to take on new data is adversely affected. A typical cause is using too many training epochs, but it could also be training on a small or skewed data set. An example is when I trained a classifier to discern between male and female Old English names, but 94 percent of surviving names are male, and the classifier didn't correct for the uneven distribution, so it inferred almost every new name was going to be male.

Pointwise mutual information (PMI) A measure of how much more frequently two terms co-occur in a corpus than expected by chance. For most corpora only positive PMI (PPMI) values are informative, so negative scores are usually replaced with zeroes.

Precision A measure of how well your algorithm does at avoiding false positives. For instance, if a classifier trained to identify Latin words in an English corpus is evaluated to have low precision, then it identifies many words as Latin that aren't.

Principal component analysis (PCA) See **singular value decomposition (SVD)**.

Recall A measure of how well your algorithm does at avoiding false negatives. For instance, if a classifier trained to identify Latin words in an English corpus is evaluated to have low recall, then it overlooks many Latin words in the corpus.

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.

Singular value decomposition (SVD) Decomposing a matrix into three factor matrices, either for dimension reduction or to discard irrelevant data and thus achieve a greater spread of data; see also **latent semantic analysis (LSA)**.

Skip-gram See **word2vec**.

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.

Strong AI A school of thought holding that AI is, or can be, indiscernable from a human mind either in its capabilities or in terms of intentionality. In the former sense it is also known as **general AI**. Contrast **weak AI**.

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, and consequently also the basis of **TF-IDF** calculations. The phrase "the tortoise and the hare" counts five tokens, but four terms or types, one of which has an absolute count of two.

Term-document matrix A table in which each row represents a word in the vocabulary, while each column represents a document in the corpus; the cells may contain raw word counts. Columns can be compared (as vectors) as a way of comparing the similarity of documents, and rows can be compared to determine how similar terms are in terms of the documents in which they occur.

TF-IDF (See readings and slides. You should be able to explain both parts of this equation, and why it yields statistical importance but not semantic meaning, and the relevance of Zipf's Law.)

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.)

Topic vector A TF-IDF or BoW vector that has been modified using LSA to reflect a grouping of terms it co-occurs with. This happens on the level of the word (word topic vector) and the document (document topic vector); the document topic vector is simply the sum of its word topic vectors. See **latent semantic analysis (LSA)**.

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

Type See **term**.

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

Valence A scale between positive and negative subjective assessment ("the pleasantness of the stimulus"); one among three dimensions on which words are commonly scored for affective meaning (**sentiment**).

Vector space model A model relying on vectors, i.e. ordered sequences of numbers, that are of the same magnitude and can thus be subjected to calculation and comparison among themselves. When used to encode term frequencies, each vector represents a document, and each number in the vector encodes the frequency of one term in that document. This requires that (1) term frequencies are normalized by document length (i.e. divided by the number of tokens in the document) and (2) the length of the vectors, i.e. their lexicon, is the same (which can be achieved by defining your lexicon as the union of the terms in all your documents). The dot product of two vectors then tells us how similar they are in terms of the cosine of the angle between them in vector space. This model is the standard for querying very large corpora, such as the web, where more sophisticated models are unfeasible. The search query is encoded as a vector, whose similarity to the vectors of documents in the corpus informs the ranking of search results.

Weak AI An school of thought holding that AI is only able to turn computers into powerful tools. Contrast **strong AI**, but also compare **narrow AI**.

word2vec A learning algorithm trained on to predict the co-occurrence of all pairs of terms in the vocabulary within a narrow window of tokens. The weights it arrives at for those matches are then used as **word embeddings** expressing the terms' similarity; the algorithm is never actually used to make predictions.

Word alignment Given a sentence-aligned **bitext**, i.e. pairs of matching sentences in two languages, word alignment is the task of identifying which tokens correspond across the language barrier.

Word embedding A representation of a word's meaning, inferred from the contexts in which it occurs.

WordNet A database documenting the relationships between words, among other things by grouping synonyms together in a synset. The original WordNet concerns English, but versions exist for other living languages, as well as for Latin.

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**.

Word-word matrix A table in which rows and columns both represent the vocabulary in a corpus (often excluding rare terms), and the cells count how often the matched terms co-occur in a predefined context. That context may be a whole document or a token window.

WSD See **word sense disambiguation**.

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 e.g. when using **TF-IDF** to determine the statistical importance of their distribution across documents in a corpus.