Week 8 Glossary

Bag of Words

A simple method that creates a vector of word counts. It counts the number of occurrences that a word (or token) appears in a document. For example, if a word appears four times in a given document, the associated numeric value of that word in the vector would be 4. A count vectorizer is scikit-learn's implementation of a bag-of-words vectorizer.

Binary Term Frequency (BTF) Vectorization

BTF, also referred to as binary encoding, represents documents as binary vectors in which each element is a binary indicator. The presence (1) or absence (0) of a word (or token) in a document is the numeric feature value.

Cosine similarity

A function that computes the similarity between two vectors (or sequences of numbers).

Count vectorizer

A scikit-learn library tool in Python that implements bag-of-words. It counts the number of occurrences that a word (or token) appears in a document. For example, if a word appears four times in a given document, the associated numeric value would be 4.

Data preprocessing

The manipulation or dropping of data before it is used to ensure or enhance performance.

Deep averaging network

A type of sequence-to-sequence model that consists of two components: a word embedding and a traditional neural network (sometimes even a linear classifier, which is a neural network without hidden layers).

Deep learning

A subset of machine learning that uses neural networks with several hidden layers. In deep learning algorithms, such as are used for NLP tasks, it is sometimes the case that different layers are able to learn different types of features in the input data so that by the time the output layer is reached, a comprehensive characterization of the input data is achieved. Deep learning methods are supported by many different software packages, which typically provide high-level support for specifying network architectures and rules for firing.

Deep neural network

A neural network with many more than two hidden layers, each with many more than two nodes. Traditional networks have a few hidden layers, whereas deep learning models can have hundreds of hidden layers. Empirically deep neural networks perform better on certain machine learning tasks, such as NLP tasks.

Generative AI

A type of artificial intelligence that is capable of generating new content. It uses deep neural networks to learn complex patterns in training data in order to generate new content that has similar characteristics.

Large language model (LLM)

A very large deep neural network that is trained on vast amounts of text data and can therefore excel at complex NLP tasks, namely Generative AI tasks.

Lemmatization

The process of identifying the inflected form, or the root meaning, of a word aSnd returning the word to that single form. Lemmatization helps with dimensionality reduction.

N-grams

Combinations of individual word tokens.

Natural language processing (NLP)

A branch of artificial intelligence (AI) that enables machines to understand the human language. NLP interprets raw, arbitrary written text and transforms it into something a computer can understand

Natural language processing (NLP) Pipeline

The NLP pipeline is a sequence of text preprocessing tasks that are required to create an NLP model.

POS (parts-of-speech) tagging

Process by which a word's part of speech is identified (verb, adjective, noun, etc.)

Recurrent neural networks (RNN)

A sequence-to-sequence model that is designed so that a given node's output flows back into the same node. In RNN, the information cycles through a loop. When it makes a decision, it considers the current input and what it has learned from the inputs it has previously received. Recurrent neural networks have had success in modeling sequence data such as text.

Sentiment Analysis

A real-world application of machine learning that predicts the sentiment (negative or positive) of a document.

Sequence-to-sequence models

A class of neural networks have been developed to deal with text data. While there are different types of sequence-to-sequence models, they all typically consist of an encoder and a decoder. An encoder is a neural network that takes in a sequence of words and outputs a vector or a code that can be viewed as a summary of the input sequence. A decoder is a neural network that takes in the vector output of an encoder and turns it into a scalar or sequence of outputs. These can be words represented by word embeddings or other things, depending on the application.

Stemming

The process of reducing a word to its root form by removing or replacing suffixes. Stemming helps with dimensionality reduction.

Stop word

A token that appears very frequently in different examples of text but also adds very little predictive value.

Stop word removal

A data preprocessing step that removes the words that commonly occur. Stop words are usually removed to reduce data size and to speed up computation.

TF-IDF vectorizer

Also known as "term frequency-inverse document frequency"; a process for encoding text that captures the relative importance of a word to a given document.

Text classification

A machine learning technique that assigns a set of predefined categories to open-ended text, categorizing text into organized groups.

Tokenization

The process of parsing text to remove certain words. Tokenization allows you to use textual data for predictive modeling and map every word in training data to a future position.

Vectorization

The process of converting raw text data into a numerical vector representation.

Word embedding

The process of converting raw text data into a numerical vector representation while seeking to capture the meaning of words within the text. It allows words with similar meanings to have an equal representation. In word embedding, each word can be represented by a k-dimensional vector. Those factors are commonly pre-trained and available as a lookup table.