**1. Explain One-Hot Encoding**

A one hot encoding is a representation of categorical variables as binary vectors.

This first requires that the categorical values be mapped to integer values.

Then, each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1.

**2. Explain Bag of Words**

The Bag of Words (BoW) model is the simplest form of text representation in numbers.

**3. Explain Bag of N-Grams**

A bag-of-n-grams model is a way to represent a document, similar to a [bag-of-words][/terms/bag-of-words/] model.

A bag-of-n-grams model represents a text document as an unordered collection of its n-grams.

For example, let’s use the following phrase and divide it into bi-grams (n=2).

James is the best person ever.

becomes

* <start>James
* James is
* is the
* the best
* best person
* person ever.
* ever.<end>

**4. Explain TF-IDF**

* TF-IDF stands for term frequency-inverse document frequency.
* Term frequency-inverse document frequency is a text vectorizer that transforms the text into a usable vector. It combines 2 concepts, Term Frequency (TF) and Document Frequency (DF).
* The term frequency is the number of occurrences of a specific term in a document. Term frequency indicates how important a specific term in a document. Term frequency represents every text from the data as a matrix whose rows are the number of documents and columns are the number of distinct terms throughout all documents.
* Document frequency is the number of documents containing a specific term. Document frequency indicates how common the term is.

**5. What is OOV problem?**

Out-of-vocabulary (OOV) are terms that are not part of the normal lexicon found in a natural language processing environment.

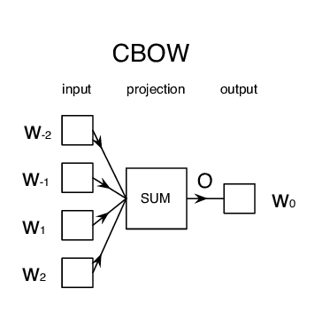
**6. What are word embeddings?**

word embedding is a term used for the representation of words for text analysis, typically in the form of a real-valued vector that encodes the meaning of the word such that the words that are closer in the vector space are expected to be similar in meaning.

**7. Explain Continuous bag of words (CBOW)**

The context of the words and takes this as input. It then tries to predict words that are contextually accurate. Let us consider an example for understanding this. Consider the sentence: ‘It is a pleasant day’ and the word ‘pleasant’ goes as input to the [neural network](https://analyticsindiamag.com/how-neural-network-can-be-trained-to-play-the-snake-game/). We are trying to predict the word ‘day’ here. We will use the one-hot encoding for the input words and measure the error rates with the [one-hot encoded](https://analyticsindiamag.com/comparing-label-encoding-and-one-hot-encoding-with-python-implementation/) target word. Doing this will help us predict the output based on the word with [least error](https://analyticsindiamag.com/decoding-most-used-confused-abused-jargons-in-machine-learning/).

CBOW model architecture

****

**8. Explain SkipGram**

* used to find the most related words for a given word.
* Skip-gram is used to predict the context word for a given target word.
* It's reverse of CBOW algorithm.
* target word is input while context words are output.

**9. Explain Glove Embeddings.**

* GloVe (Global Vectors for Word Representation) is an alternate method to create word embeddings.
* Generating word embeddings by aggregating global word-word co-occurrence matrix from a corpus.
* t is based on matrix factorization techniques on the word-context matrix. A large matrix of co-occurrence information is constructed and you count each “word” (the rows), and how frequently we see this word in some “context” (the columns) in a large corpus.