## Bag of words(BOW)

Bag of words is a simple and popular technique for feature extraction from text. **Bag of word model processes the text to find how many times each word appeared in the sentence. This is also called as vectorization.**

Steps for creating BOW

* Tokenize the text into sentences
* Tokenize sentences into words
* Remove punctuation or stop words
* Convert the words to lower text
* Create the frequency distribution of words

***What is the problem with bag of words?***

In the bag of words model, each document is represented as a word-count vector. These counts can be binary counts, a word may occur in the text or not or will have absolute counts. The size of the vector is equal to the number of elements in the vocabulary. If most of the elements are zero then the bag of words will be a sparse matrix.

In deep learning, we would have sparse matrix as we will be working with huge amount of training data. Sparse representations are harder to model both for computational reasons as well as for informational reasons.

**Huge amount of weights:**Huge input vectors means a huge number of weights for a neural network.

**Computationally intensive:**More weights means more computation required to train and predict.

**Lack of meaningful relations and no consideration for order of words:**BOW is a collection of words that appear in the text or sentences with the word counts. Bag of words does not take into consideration the order in which they appear.

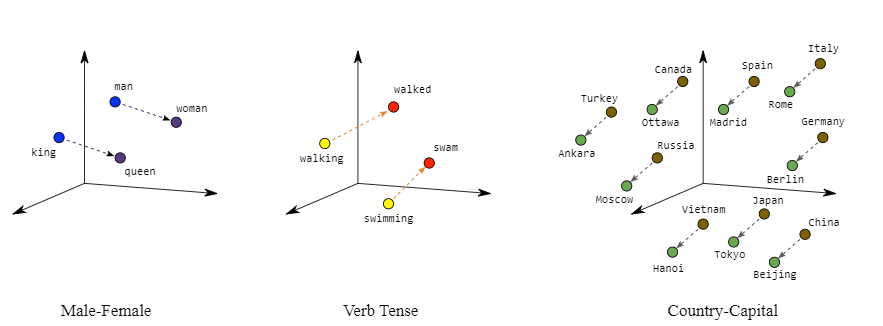
Word Embedding is solution to these problems

**Embeddings translate large sparse vectors into a lower-dimensional space that preserves syntactic and semantic relationships**.

Word embeddings is a technique where individual words of a domain or language are represented as real-valued vectors in a lower dimensional space.

**Sparse Matrix problem with BOW is solved by mapping high-dimensional data into a lower-dimensional space.**

Lack of meaningful relationship issue of BOW is solved by placing **vectors of semantically similar items close to each other**. This way words that have similar meaning have similar distances in the vector space as shown below.



“king is to queen as man is to woman” encoded in the vector space as well as verb Tense and Country and their capitals are encoded in low dimensional space preserving the semantic relationships.

## **Word2Vec**

Word2vec is an algorithm invented at Google for training word embeddings.

The distributional hypothesis states that words which, often have the same neighboring words tend to be semantically similar.

This helps **to map semantically similar words to geometrically close embedding vectors.**

Distributional hypothesis uses continuous bag of words(CBOW) or skip grams.

**word2vec models are shallow neural network with an input layer, a projection layer and an output layer.**

**It is trained to reconstruct linguistic contexts of words.**

Input layer for Word2vec neural network takes a larger corpus of text to produce a vector space, typically of several hundred dimensions.

Every unique word in the text corpus is assigned a corresponding vector in the space.

**This architecture is called continuous bag of words CBOW as it uses continuous distributed representation of the context. It considers both the order of words in the history as well as future.**

Two different learning models were introduced that can be used as part of the word2vec approach to learn the word embedding; they are:

* Continuous Bag-of-Words, or CBOW model.
* Continuous Skip-Gram Model.

The CBOW model learns the embedding by predicting the current word based on its context.

The continuous skip-gram model learns by predicting the surrounding words given a current word.

The continuous skip-gram model learns by predicting the surrounding words given a current word.