1 What is Word2Vec

If somebody wants to train a machine learning model on textual input or if they need to find relations between words, they first need to convert text to vectors, these are called Word Embeddings. There are many ways of doing this, one of them being taking a large sample of textual data and finding all unique words and assigning an Id to them. That works great but how how do you cluster similar or related words together? You need to run another clustering algorithms over the data to group related words.

Word2Vec[1] is a machine learning model used to generate Word Embeddings with words which are similar to each other are in close proximity in vector space. Word2Vec is developed by team led by Mikolov et al while he was at google.

Word2Vec can automatically capture the relations between words like **Paris, Beijing, Tokyo, Delhi, New York** are all clustered together in vector space. Similarly **Cat, Dog, Rat, Duck** are all clustered together in vector space.

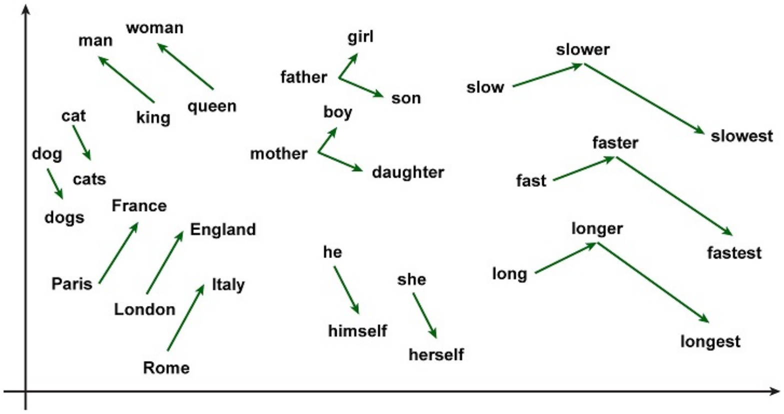

Figure 1

This also helps us finding interesting relations between many words. What if you remove **Man** from **King** and add **Women**? You get **Queen**

King - Man + Women = Queen

Word2Vec can be used to perform the above relational operations. It can also extract and provide the most similar words, how similar are two words, pick the odd word out of group of words, if the model is trained with large enough corpus of data.

References:

[1] Word2Vec: https://code.google.com/archive/p/word2vec/