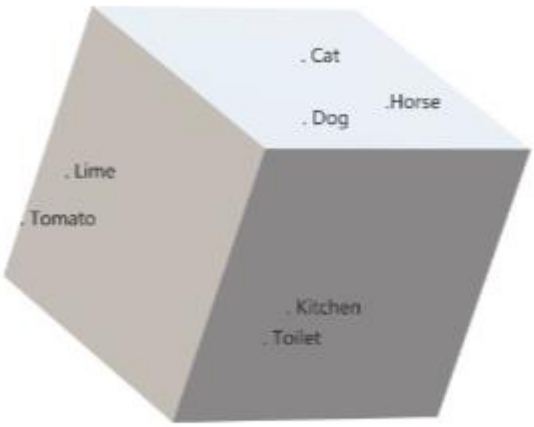
Applications of NLP

i) Processing Text for Computer understanding:

Text Preprocessing is one of the initial and most important steps in NLP. We perform text preprocessing in order to transform text into a more digestible form so that machine learning algorithms can perform better. Following are the three basic steps in text preprocessing

a) ***feature extraction***: We used Bag of Words or TF-IDF (See Appendix) this. This counts the number of words in a sentence so that our algorithm can do further works, like
b) ***tokenization***: Tokenization is about splitting a string into smaller pieces. The algorithms that I am using tokenize the strings into words.
c) ***stopwords removal***: Removing frequent common words like a, an, the , and so that the algorithm doesn't get biased.
d) ***stemming/lemmatization***: Stemming and Lemmatization are used to convert the words to their base forms. Either of the process could be used to make learning of words better.

ii) word embeddings



Each word is mapped to one vector and the vector values are learned in a way that resembles a neural network. Hence we will get a real-valued vector of fixed dimension, depending on the model we are using.

iii) Sentiment Analysis

Sentiment analysis is often done by companies to know the value of their products, review the feedbacks by their customers. It can be

iv) Part of Speech Tagging for product recommendation

A simple idea of finding out nouns (name of objects/products) to recommend people using social media the products.

v) Text Classification

We used Naïve Bayes Classifier, which is a probabilistic model that classifies based on probabilities of events. We use the probability of every individual word in a sentence to perform a binary classification. Sentiment analysis is also a type of binary text classification.

vi) Clustering

Performing Clustering to group similar groups together. Simple distance measure can be used for measuring the proximity of one group with another.

vii) Speech Recognition

Speech recognition can be easily done using third party libraries.