* Regular Expressions are text strings which are used as pattern for matching strings. The language accepted by finite automata can be described by Regular Expressions
* Clustering is a type of unsupervised learning method. It is the task of grouping unlabeled data. With the help of clustering, unlabeled objects with a similarity between them can be grouped together
* K-Means  
  Mean Shift  
  Hierarchical Clustering
* Clustering can be used on a group of documents too. By clustering words, we can group together words with same meaning or same contexts. The semantic clustering and feature selection method is proposed to improve the clustering and feature selection mechanism with semantic relations of the text documents. We can also use traditional vector space model, which treats documents as group of words. Soft clustering using k-means algorithms should provide good results
* <https://towardsdatascience.com/a-friendly-introduction-to-text-clustering-fa996bcefd04>  
  <https://ieeexplore.ieee.org/abstract/document/7881515>  
  <https://dc.uwm.edu/cgi/viewcontent.cgi?article=3354&context=etd>  
  <https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a>  
  <https://devopedia.org/text-clustering>
* Facet groups all the values that appear in a column, and lists each value with the number of records it appears in. It also allows to filter the data by these values and edit values across many records at the same time.
* When the hash value of a key maps to an already occupied bucket of the hash table, it is called a Collision. This technique is applied in Key Collision Methods. With this method, it is possible to cluster different strings based on the fact that their key is the same
* The fingerprint method generates the key from a string value by following a set of operations in an orderly fashion. Some of the operations are removing the whitespaces, converting characters to lowercase representations, removing the duplicates and so on.
* The N-gram fingerprint method creates a key comprising the substrings of length N of a string. Though this method can produce false positives but using this method, such clusters can be found which the previous normal fingerprint method was unable to find.
* Phonetic fingerprint can cluster similar sounding words together. This is useful when people doesn’t know the spelling of a word or misunderstand it. It is based on the idea of a pronunciation of a word. This method can find clusters which is not possible to find with regular and n-gram fingerprint methods
* Soundex code consists of a letter at the beginning followed by 3 numbers. According to the Soundex coding guide, most of the letters are assigned to a number and the rest of them are discarded.
* Metaphone codes use the 16 consonant symbols. What it basically does it, drop some letters according to their position in the string and also transforms a single letter to a pair of letters or transform a pair of letters to a single letter. Moreover, it encodes voiced and unvoiced consonants differently
* The Cologne phonetics matches each letter of a word to a digit between "0" and "8". And sometimes to select the appropriate digit, at most one adjacent letter can also be used a context.
* Daitch–Mokotoff Soundex is an in improved and refined version of the Soundex algorithm. Here the coded names are 6 digits long. Also, this algorithm encodes multiple character n-grams as single digits.
* The steps of Beider-Morse algorithm are as follows. It first identifies the language. Then it calculates the exact phonetic value, approximate phonetic value and Hebrew phonetic value. And then finally it searches for matches.
* In Nearest Neighbour method, any pair of strings that is closer than a certain value will be clustered together. That certain value is termed as ‘radius’ which simply refers to the distance threshold.
* Key-Collision methods are fast but we have no control over the similarity index between strings. Whereas Nearest Neighbour method provide us with a parameter which we can tune to specify the amount of difference between strings up to which we will consider.
* Edit distance is a way of differentiating 2 strings based on the number of operations that are required to convert a string to another.
* Levenshtein distance  
  Longest common subsequence
* Text compressors work by estimating the information content of a string. So the concatenation of 2 similar strings after compression, should vary a little. But the concatenation of 2 very different strings should vary a lot in length. Based on this idea, PPM works.
* Dynamic Markov Compression  
  Huffman encoding
* Supervised ML means to train a machine with dataset which is well labelled.
* Semi-supervised ML means to train a machine using data which is partially classified.
* Unsupervised ML means to train a machine using information that is not labelled.
* Reinforcement learning means to learn through experience. Here, the machine tries to make the best decision in a particular situation so that it can maximizes the total reward in the end.
* Feature learning means to extract or derive features from the given data for the sake of a more accurate prediction or classification.
* Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks

Medical Image Processing and Data Mining  
Optical character recognition