Refer: <http://www.tfidf.com/>

1. TF-IDF: (Term frequency inverse document frequency)

TF: Term Frequency measures frequency of a word in a document.TF consider all words are equally important

Formula: (frequency of word) / (Total number of words in a document)

IDF: Inverse document frequency measures how important a word in that document. As we know words like “is,am,the” appears multiple times in a document but have lesser importance. Thus we need to think on IDF

Formula: log\_e(Total number of words in a document)/(frequency of that word)

1. Corpus: Body of a text
2. Lexicon: Word and their meanings
3. Countvectorizer: Count the number of times a words appear in a document
4. Data Cleaning methods:
5. Deleting duplicate/redundant or irrelevant values
6. Work on case sensitive(America, america), different pattern(yellow, \_yellow) and combine in same form
7. Handle missing data
8. Remove space, extra characters

f) Chi-Square Test for feature selection: Calculation of Chi-Square between each feature and target feature and select the number of features with best Ch-Square value.

Chi-Square = (observed freq – expected freq)^2/ (Expected freq)

g) Decision Tree vs Random Forest: Decision tree use a single tree structure to train the model. The advantage of simple decision tree is easy to interpret but it leads to overfit

Random forest like a black box and works with numbers of trees (passed with n\_estimator variable) and here we can pass max number of features to be used in each tree. Unlike DT this doesn’t create highly biased model and reduce the variance.

h) Limitation of linear regression:

a) It is limited to linear relationship

b) Is easily effected by outliers

c) Overfitting arises if data has too many parameters

i) Feature Engineering: Creating new features from existing data to improve model performance is feature engineering. This varies with data. Like:

a) Combining two features to create a single one

b) Put range of certain features for filtering the data

j) PCA: (Principle Component Analysis):