**NLP**

1. **Data cleaning**
   1. Remove stopwords - the, and etc.
   2. Remove punctuation
   3. Use regular expressions
   4. Tokenize
   5. Stem – stopping and stop becomes stop
      1. Porterstemmer
      2. Many other stemmers available
   6. Lemmatize data – a different form of stemming based on finding the root of the word
2. **Vectors** 
   1. Count vectors
   2. Ngram vectors
   3. Tf-idf vectors and many others..
   4. Build the vectors with the training data only, don’t use the whole data to build the vectors. When adding the vectors from test data as we use the training data to fit the vector only the common words from train and test will be used, other words which are exclusively present only in test will be ignored.
3. **Convert the text to vectors**
4. **Then add other features etc. punctuation%, length of string etc.**
5. **Create the features vectors by combining the new features + the vectors**
6. **Use random forest, gradient boosting etc.**
   1. **Identify the best model parameters by performing – cross validation for ex. 5 fold validation, hold out test. Cross validation is better than hold out test.**
   2. **Perform grid search by creating multiple models with different values for parameters, identifying the model with least mean error, other measurement criteria etc.**
   3. **Precision, recall, accuracy – various measurement criteria. The criteria used to evaluate the model will depend on which is more costly false positive or negative.**
      1. **Recall – higher cost on false negative. Anti-virus, anti-hacking software.**
      2. **Precision – higher cost on false positive. Spam filter, some spam in inbox is okay, vice versa is more expensive missing an important email.**
   4. **Log the running time for fit, predict etc. these could also be critical parameters on the final decision**
7. **Build the model using the best model parameters and model identified in the steps above**
8. **As we get more data, we may have to re-evaluate our decisions and identify a different model**
9. **Consider business context – make the trade off**