Chapter 6 : learning to classify texts

Gender identification: names are not gender, unethical to predict

Choosing the right features: length of vector is no of unique words and has no of times a word appears

Exploiting context: to know where the word is in sentence rather than just its count

Sequence classification: named entity tagging and coreference resolution

HMM – hidden markov model : everything's probability depends on probability that comes just before that

RNN : takes care of a sequence – probability of a word depends on everything that's before that – sentences can be really big – vanishing gradient problem

LSTM – captures long term dependencies

Coreference resolution – combining named entity detection and clustering them

Evaluation : accuracy tells you nothing

- Precision and recall break down problem into 4 parts, True and false positives and negatives
- Precision indicates how many of the items that we identified were relevant TP/(TP+FP)
- Recall indicates how many of the relevant items that we identified TP/(TP+FN)
- F – score is harmonic mean of precision and recall

For multi class classification we use confusion matrix

Chapter 8: Analyzing sentence structure

Context free grammar – no subtree in a grammar tree is connected to any other subtree – simplified but not correct

---

Topic Model

Doesn't care about order of words, it's a statistical model, unsupervised machine learning

Finds list of unique words in the document – vocabulary, makes probability distributions

To find themes in a book or in a list of documents

Code for topic model : https://mimno.github.io/Mallet/topics.html