

Instead of loading everything into memory at once, we can deal with text line by line.

Fit into ram, this is common way to deal with this.

Open source data set.

Old data set: Enron

Raw data set -> processed messages.

Label two data sets and put them together into same container, then apply reshuffle.

So the model get well mixed data to train.

Lemmatization : have – has, bring word to its original form.

Reduce the noise.

WordNetLemmatizer()

Remove noises from data set == pre-process the data set.

Remove capital cases, most common works

Bad-of-words:

Words appear frequency.

Image 2d number- pixels.

Feature extraction function apply to BOW.

Tokens = list of words.

Transfer from Collection->dictionaries [ word – occurrences]

Word of tokens

Splitting datasets into training and testing.

It is natural for training accuracy to be higher than unseen data set.

Ensemble model will make combined model more robust and stronger.

The downloaded file Unicode file format is important.

Random state can be reproduced?

POS( parts of speech) tagging

Inform downstream model.

Hashing vectorizer vs count vectorizer

Hash -transform variable length data to fixed length representation

Pros: number represent word, save memory, fast, can remember word

Variable data pass mathematical formula.

Reduce the matrix size:

**singular value decomposition(SVD)**

**matrix decomposition**

[Latent semantic analysis(LSA)](https://en.wikipedia.org/wiki/Latent_semantic_analysis)

Pipeline:

Hashing vectorizer->CountVectorzier -> transformer -> SVD decomposition -> normalize-> naviebayes

SVM(support vector machine)

Support and vector: maximize the distance between vector(separate) and nearest support.

Kernel: no longer need linear data.

Support vector machine: ONLY work for linear separable data set.

Hyper-plane.

Only Support vectors(extremes) are important.

Support curve linear classification

SVR: regression.

TFIDF->SVM

Non-linear : lower dimensional -> high dimensional

Kernel trick-kernel function.

Advantage of SVM: memory efficient, effective in high dimension, #feature < #dimension.

Disadvantage of SVM: #feature > #dimension. poor performance. SVM do not provide probability estimates directly.

TF-IDF : based on the word

LSA : based on the topic