evaluation

N gram

Evaluation

Confusion matrix: How many doc in c1 are incorrectly assigned to c2

[Precision-Recall, Accuracy, F score](https://www.youtube.com/watch?v=Wq0taCUCSlA&list=PLLssT5z_DsK8HbD2sPcUIDfQ7zmBarMYv&index=30)

Text classification

Step1. tokenization. Convert raw text to tokens

Step2. Extract features from documents

Step3. Train: Different classification methods

* Naive bayes(max posterior)
* Max entropy
* SVM
* kNN
* Regularized logistic regression
* Decision trees
* Methods for semi-supervised training: bootstrapping, EM, over unlabeled documents

Validation:

Cross validation. Take 1/5 data from trainset, use it as “development testset” (=val set)

Step 4

Test:

Perplexity

evaluation

Sentiment analysis task 举例 polarity classification

Input x: a document. x: a paragraph or a document or a review 一篇评论.

Output y: a class. predicted class y∈{ a positive, a negative review} polarity: does the review/doc/word belongs to positive or negative sentiment?

100篇review docs 对应于 100 annotated classes.

Goal: when we get a new unseen review, estimate is it a positive or negative review?

P(a doc|class “positive”)

P(a doc|class “negative”)