1. Introduction to natural language processing

Subject: Task: Common Problems:

Use Cases:

1. Assigning subject categories, topics, or genres
2. Spam detection
3. Age/Gender detection
4. Language Identification
5. Smart assistant
6. Language translation

1.1. Text classification

Subject: document d, fixed set of classes $C = \{c_1, c_2, ...\}$ Task: predicted class c

Common Problems:

1. How to classify

Hand-coded rules. Use the human defined combination of words, phrases. Example: (spam: black-list-address OR ("dollars" AND "have been selected"))

Pros: may have high accuracy Cons: took too long time for build and maintain, requires an expert

Supervised ML. Use classifier to train model on data.

Pros: may have high accuracy. took less time. easy to maintain. Cons: may have less accuracy, need more engineering resource.

1. Get training set of data: (d_i, c_j)
2. Configure and train model