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

Lib: https://habr.com/ru/companies/macloud/articles/558760/ https://habr.com/ru/companies/macloud/articles/560062/ https://ru.wikipedia.org/wiki/%D0%90%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC_%D0%9A%D0%BE%D0%BA%D0%B0_%E2%80%94_%D0%AF%D0%BD%D0%B3%D0%B5%D1%80%D0%B0_%E2%80%94_%D0%9A%D0%B0%D1%81%D0%B0%D0%BC%D0%B8 https://en.wikipedia.org/wiki/CYK_algorithm

https://cdn.openai.com/papers/gpt-4.pdf