学习笔记:

Task 1:

赛题评测标准：F1-score 均值越高越好

解题思路：

1. TF-IDF + ML分类 (SVM, LR or XGboost)

2. FastText

3. WordVec +DL分类 (TextCNN, TextRNN, BiLSTM)

4. Bert 词向量

* WordNet: contain synonym sets and hypernyms



* One-hot vector

Vector dimension = number of words in vocabulary

Above have failed due to disadvantage such as not correct synonym meaning or similar words have same one-hot vector. Thus, encode word similarity.

* Word vectors (word embeddings or word representations): distributed representation
* Word2vec: a framework for learning word vectors
  + A large corpus of text
  + Every foxed word is a vector
  + Go through each position t in the text, which has a center word c and context word o (“outside”)
  + Use similarity of word vectors for c and o to calculate the probability of o given c
  + Adjusting word vectors to maximize the probability

