话题分类

数据集一:八个类别，房产，IT，汽车，体育，旅游，健康，教育，军事。

每个类别随机选取200篇文章。经过分词，停止词过滤等，平均每篇文章250个单词。

实体识别

CRF: **在linux上进行，window格式有很多问题。**

1. 生成配置文件，字符序列，对应标记序列，不要有多余句号。其次CRF++使用了linux格式文件，在windows生成的要更改句子结尾标记。可以在nodepad++>edit>EOL Conversion, 先转换成Mac format，再转换成linux format，不然会出现bad标记。
2. 学习：

./crf\_learn ./example/CPT/template.txt ./example/CPT/cpt.train ./example/CPT/model

1. 预测：

./crf\_learn ./example/CPT/template.txt ./example/CPT/cpt.train ./example/CPT/model

1. 评估结果，因为CRF++使用制表符分割，所以需要指明。

./conlleval.pl -d "\t" < ./example/CPT/predict.txt