本次实验环境为 ubuntu16.04 python 版本为 python3.6

本次实验采用 navie bayes 对 20newsgroups 进行分类，模型构造直接使用 sklearn 库

1、获取训练数据和测试数据
train_d = fetch_20newsgroups(subset = 'train',categories = categories);
test_d = fetch_20newsgroups(subset = 'test',categories = categories);

2、获取数据集特征
vectorizer = HashingVectorizer(stop_words = 'english',non_negative = True,n_features = 10000)
fea_train = vectorizer.fit_transform(train_d.data)
fea_test = vectorizer.fit_transform(test_d.data);

3、构造 naivebayes 多项式模型
clf = MultinomialNB(alpha = a)
alpha 为平滑参数 默认 1.0

4、训练集合上进行训练， 估计参数
clf.fit(fea_train,train_d.target);

5、对测试集合进行预测 保存预测结果
pred = clf.predict(fea_test);

结果为
alpha= 0.01 precision= 0.8005366715683742
alpha= 0.05 precision= 0.808135994679238
alpha= 0.1 precision= 0.8093082169779041
alpha= 0.15 precision= 0.809488280285531
alpha= 0.2 precision= 0.8094627780947201

alpha= 0.15 时 准确率最高 构造模型选用 alpha= 0.15