实验报告

学号：201814817　　姓名：连艺翔　　班级：2018 级学硕

一、 简介

本实验是基于 KNN 分类器在数据集 20news-18828 上的实现实验，本实验任务分为两部分：1）将 20news-18828 上的数据进行文本处理，转化为 VSM 模型；2）实现 KNN 算法，第 18828 中的文档进行分类。

本实验分为两个部分：1）实现从 20nwes-18828 中读取数据并通过文档预处理将其转换为 BOW 模型再转换为 VSM 模型，主要使用的 socket 有 NLTK、collection、Textblob 等。最后结果为 18828*N 维的矩阵 2）计算测试数据集中每个向量与训练数据集中每个向量的 Cosine Distance 以衡量文档之间的相似程度，选最相似的 K 个。

二、 数据集

20news-18828 数据集一共包含 20 个类，共 18828 个文档，均来自各个不同的新闻评论，各类数据分布平衡，文档编码少部分采用 ISO 格式，大部分采用 ASCII 编码。文档平均长度在 1000 词左右，因为取自新闻有部分网页格式。

三、 方法步骤

3.1VSM

VSM 是在词表统计的基础上，将词表统计结果投影到向量空间的方法，其关键步骤包括：数据集预处理、词典统计、

VSM 向量计算。

3.1.1 数据集预处理

数据集预处理分为：1）去符号 2）分词 3）词形还原 4）大小写转换 5）去停用词这五个部分组成。使用 str 类的方法 maketrans 对文档进行去符号，这里我们除去所有非英文字母的符号。使用 textblob 对文档进行分词，使用 nltk 套件中的 Snowballstemmer 进行词形还原，使用 nltk 中的 stopwords 作为停用词表。

3.1.2 统计词频

对所有文档进行一次词频统计，取词频小于 1000，大于 10 的词作为词表。

对每个文档以词表为标准进行词频统计，将每篇文档以词表长度 N 表示为向量。

3.1.3VSM 向量计算

VSM 通过计算所有文档的权重值 IDF 与每个文档的权重值 TF,使 IDF 与 TF 做 element product 得到同维度的向量，其计算公式如下：

TF：对于一篇文章而言某个维度上的 TF 值越高说明特征越突出。

$$- \quad tf(t,d) = \begin{cases} 1 + \log c(t,d), \, if \; c(t,d) > 0 \\ 0, \, otherwise \end{cases}$$

IDF：对于文档，某个词的 IDF 越高，越说明其重要性低

$$IDF(t) = \log(\frac{N}{df(t)})$$

VSM：

$$\mathrm{VSM} = \mathrm{IDF} \times \mathrm{TF}$$

3.2KNN

1）计算文档之间的 Cosine Distence：

$$cosine(d_i, d_j) = \frac{V_{d_i}^T V_{d_j}}{\left|V_{d_i}\right|_2 \times \left|V_{d_j}\right|_2}$$

2）挑选值最大的前 K 个训练集文档，统计他们的类别，对测试数据集的文档进行类别预测。

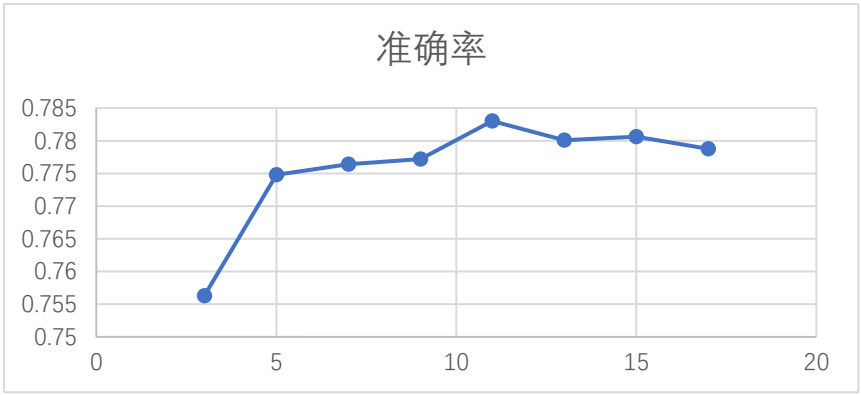
四、评价与结果

我们使用 precision accuracy 对结果进行评价：

$$\mathrm{accuracy} = \frac{\mathrm{N}}{\mathrm{M}}$$

其中的 N 是测试集中预测正确的个数，M 是测试集中所有文档的个数。

根据预测结果，我们发现随着 K 值增大，准确率变化如图所示：


准确率
0.785
0.78
0.775
0.77
0.765
0.76
0.755
0.75
0
5
10
15
20