浙江工业大学

文本分析与挖掘

上机实验

计算机科学与技术学院

词嵌入实现与应用

一、 实验目的

1. 实现两种基本词嵌入算法 skipgram 和 CBOW，并对结果进行对比和评估。
2. 掌握词嵌入在分类和聚类中的基本应用，了解预训练词嵌入的适用范围。

二、实验内容

1.调用 gensim 得到词嵌入向量。

a. 对 20newsgroups 子集 multi3 和 multi5 调用 gensim 得到 skipgram 词嵌入向量。假设维度 100， 窗口大小：5。

b. 通过所含词的词向量取平均得到文档向量。

c. 基于以上文档向量，计算所有类每两个文档之间的欧式距离的平方做为该表示下类别的紧凑度 compactness，具体计算如下：

$$Comp = \sum_{c=1}^{C} \sum_{x_i, x_j \in \chi_c} \left\| x_i - x_j \right\|^2$$

d. 改变参数,对比 a 中得到 compactness。具体为:改变维度=200，窗口大小保持 5，重复上面实验;改变窗口大小 10，维度保持为 100，重复上面 skipgram 实验。

e. 按照 a 中参数设置，调用 gensim 实现 CBOW 词向量嵌入，计算 Compactness,对比同样参数设置下 skipgram 的结果。

f. （选做）基于 keras，参照课件中代码和步骤，逐步实现两种方法的词嵌入。

g. 基于 sklearn 中的 TSNE 工具，进行可视化。具体用法参考课件例子。基于 skip-learn 和 TFIDF 用 TSNE 画出 2 维词嵌入结果并讨论。

2.基于词嵌入向量对 multi3 和 multi5 进行分类。

a. 基于词嵌入向量（设定适当参数，文档向量为其所含词的词嵌入向量的平均），用 SVM 进行分类，得到 accuracy。

b. 基于 TFIDF 词袋模型用 SVM 对上面两个数据集进行分类，对比 a 中结果并讨论。

c. 把两种表示拼接起来得到新的文档表示后进行 SVM 分类，对比结果并讨论。

d. 以 TFIDF 为权重，对每个文档所含词向量进行加权平均得到文档向量，进行 SVM 分类，对比前面结果。

2.预训练词向量

基于 Glove 预训练词向量，重复上面 a,c,d 中的实验，根据结果讨论预训练词向量的适用范围。