词汇相似度实验报告

黄梓铭 1501214385

1. **实验目的**

目的：运用所学知识实现5种词汇相似度计算方法，尽量保证方法的多样性。

数据：wordsim353

评价标准：Spearman’s rank correlation coefficient

1. **实验工具**

Nltk工具包中的Wordnet、gensim、BeautifulSoup、GoogleNews-vectors-negative300.bin、scikit-learn的SVM包

HuangModel：基于GoogleNews-Vector训练好的语料编译生成共175M、每个词向量共300维的词向量集合。

1. **实验方法**

本次实验实现了5种词汇相似度计算方法，其中包括：

1. 计算向量的cosin相似度
   1. Cosin-based
2. 基于路径的相似度：
   1. Path-based
   2. Wu-Palmer-based
   3. Leacock-Chodorow-based
3. 基于互信息的相似度计算：
   1. Jiang-Conrath-based
   2. Lin’s-based
   3. Resink-based
4. 基于google检索返回的页面数量的方法
   1. Page-count-based
5. 有监督的SVM方法计算词汇相似度
   1. SVM-based

上述的5种相似度计算方法，主要可归类为4种基本思想：

1. 基于语料统计的词汇相似度计算

基本思想：基于word2vec项目给出的googleNews语料得到词汇向量。调用gensim的Word2vec 模型，利用googleNews向量得到词汇向量模型，进而计算词汇cosin相似度。

1. 基于语义词典的词汇相似度计算

基本思想：调用nltk工具包，获取wordnet词典，并分为基于路径计算相似度以及基于互信息计算相似度。

1. 基于检索页面数量计算相似度

基本思想：根据google搜索返回的页面数量，利用WebJaccard计算相似度：

1. 基于有监督计算相似度

基本思想：有监督的SVM算法，用set1作为训练集得出set2的预测结果，同时用set2作为训练集得set1的预测结果。SVM算法中统计每个词对中相同词义的数量以及不同词义数量作为特征，并且将基于wordnet获得最大的相似度作为输入，进行SVM回归拟合，最后依据SVM预测得出词对间相似度的度量。

1. **实验结果比较与分析**

**4.1 实验结果**

表-1 5种词汇相似度spearman结果展示

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | method | set1 | set2 | combined |
| 1-WordNet-Path | Path-based | 0.326779071 | 0.241554121 | 0.299440209 |
| Wu-palmer-based | 0.246995375 | 0.212842164 | 0.25897806 |
| LC-based | 0.330724598 | 0.240651359 | 0.301191532 |
| 2-WordNet-IC | Resnik-based | 0.376714224 | 0.234239314 | 0.33087501 |
| J&C-based | 0.364555329 | 0.158689775 | 0.283353797 |
| Lin-based | 0.358463143 | 0.196707109 | 0.298189066 |
| Word2vec | 3-GoogleNews | 0.665138987 | 0.658314845 | 0.700016649 |
| Search Result | 4-GoogleSearch | 0.268630561 | 0.04013975 | 0.140247016 |
| Regression | 5-SVM | -0.19723532 | -0.02847558 | 不要求实现 |

图-1 不同方法的spearman差异度

**4.2实验结果分析**

基于WordNet语义关系计算词汇相似度的方法比较简单， 可以利用词汇的同义词关系、语义解释、例句等信息。但存在一些缺点，（1）有些词不被语义词典所包含：实体、新词等。（2）大部分方法依赖于上下位层次关系：限于名词，对于形容词和动词并不完善。（3）不词性不能通过wordnet计算相似度。

基于语料统计计算词汇相似度的方法比较多，本文用了GoogleNews该方法，从图1可以看到，基于语料统计的方法的计算结果明显高于基于语义词典的方法，而且基于googlenews的方法远远高于基于语义词典的方法获得的最大值，其主要原因是使用的googlenews是使用了300G原始语料运用300维度计算所得的模型，故相似度计算准确。GoogleNews模型在65%~70%之间。

基于google检索结果计算词汇相似度，最高只有26%的相关系数。这种方法过于简单单，没有考虑到词汇之间的层次关系和语义关系，只是根据返回检索数量进行计算，显然结果准确度较低。

最后一种方法是基于有监督的方法计算相似度，从表-1结果可以看出，通过SVM训练得到的模型计算的结果，显然准确低很低，当然这和我们训练集中的数据很少有关，还需要注意一点的是，用SVM计算出来的结果和原始数据是呈负相关的关系。