实验 5：基于同义词词林的词语相似度计算

李鹏，*10175501102@stu.ecnu.edu.cn*

January 2, 2020

简介

本次实验基于哈工大信息检索研究中心提供的《同义词词林》扩展板，参照《基于路径与深度的同义词词林词语相似度计算》一文提出的相似度计算方法，实验完成了基于同义词词林的词语相似度计算，并在提供的测试集上进行了测试。

距离计算公式

根据原论文，两个义项 s_1、s_2 的相似度计算公式为：

$$sim(s_1, s_2) = \frac{Depth(LCP(s_1, s_2)) + \alpha}{Depth(LCP(s_1, s_2)) + \alpha + Path(s_1, s_2) + \beta}$$

其中 $Depth(LCP(s_1, s_2))$ 表示两个义项公共父节点所在的层数到根节点的权重值相加，层数越深，表明两个义项离公共父节点越近，两个义项越相似；$Path(s_1, s_2)$ 表示两个义项到公共父节点的距离，距离越远，两个义项差异性越大。其中 α 的值设置为根节点的深度值，经实验测试最优值为 0.9，而 β 的含义如下：如下：

$$\beta = \frac{K}{N} * Weight(i)$$

其中 N 表示公共父节点有多少子节点，而 K 表示两个义项对应的父节点的公共子节点之间的距离。K 越大，两个义项差异性越大，N 越大，两个义项相对差异性越小。而 $Weight(i)$ 表示公共父节点到下面一层节点的边的权重，权重越大，两个义项的差异性越大。其他一些边界情况可以参见原论文。

实验结果

本实验最后在提供的测试样本上测试了整个相似度计算方式，最终结果如下：

（备注：代码中还提供了基于词向量的词汇相似度计算方式，从另一个角度计算词汇的相似度。虽然目前基于词向量的词汇表示被广泛应用，但是相对于基于词林的计算方式，其相似度的含义还有待考虑，不同距离计算方式也会影响到相似度的含义。）

义项 1	义项 2	相似度	义项 1	义项 2	相似度
轿车	汽车	0.936086529006883	鸟	公鸡	0.5468326152864209
宝石	宝物	0.9376231122783979	鸟	鹤	0.5503452705957925
旅游	游历	1	工具	器械	0.494195688225539
男孩子	小伙子	0.9345549738219896	兄弟	和尚	0.25397472482674277
海岸	海滨	0.9477351916376306	起重机	器械	0.494195688225539
庇护所	精神病院	0.9500998003992015	小伙子	兄弟	0.267038148306901
魔术师	巫师	0.8120300751879699	旅行	轿车	0.02030293264582662
中午	正午	1	和尚	圣贤	0.2636479052052476
火炉	炉灶	0.951048951048951	墓地	林地	0.5252643948296123
食物	水果	0.24563017479300828	食物	公鸡	0.24791086350974934
海岸	丘陵	0.5437956204379563	水果	火炉	0.25260170293282874
森林	墓地	0.021701687909059596	署名	海滨	0.02030293264582662
岸边	林地	0.24791086350974934	汽车	巫师	0.021701687909059596
和尚	奴隶	0.24791086350974934	高地	火炉	0.021406727828746176
海岸	森林	0.2625368731563422	大笑	器械	0.0200445434298441
小伙子	巫师	0.26034266610948603	庇护所	水果	0.021701687909059596
琴弦	微笑	0.0200445434298441	庇护所	和尚	0.021406727828746176
玻璃	魔术师	0.021701687909059596	墓地	精神病院	0.021701687909059596
中午	绳子	0.021701687909059596	男孩子	公鸡	0.021701687909059596
公鸡	航行	0.02030293264582662	垫子	宝物	0.23482849604221637
庇护所	墓地	0.5277449822904369	墓地	坟堆	0.021701687909059596
大笑	小伙子	0.01979264844486334	玻璃	珠宝	0.2625368731563422
男孩	圣人	0.021119678176332552	魔术师	圣贤	0.25090616190092635
汽车	垫子	0.2678034102306921	圣人	巫师	0.25090616190092635
护堤	海滨	0.24791086350974934	圣贤	圣人	1
海滨	航行	0.02030293264582662	山岗	斜坡	0.7991543340380549
鸟	树林	0.2678034102306921	绳索	绳子	1
火炉	器械	0.5340501792114696	玻璃	杯子	0.2625368731563422
鹤	公鸡	0.5503452705957925	大笑	微笑	0.951682772409197
山岗	树林	0.2625368731563422	农奴	奴隶	1
署名	签名	1	男性	户口	0.5491003685237373
森林	树林	1	优点	先天不足	0.7991543340380549
雄鸡	公鸡	1	长期性	防御性	0.9501970619849516
靠枕	枕头	0.9422632794457275	双喜	有事	0.9513872421152478
墓地	墓园	1	先例	前例	1
叛徒	工作狂	0.26034266610948603	狂潮	工人运动	0.9504950495049505
武力	药力	0.9407952137604388	遭遇	风浪	0.951048951048951
特长	能力	0.8120300751879699	命运	苦命	0.9488372093023255
抄袭	克隆	0.2636479052052476	天灾	水害	0.9517884914463451
成年人	市民	0.2636479052052476	八卦	两仪	0