汉语词向量学习作业

1、使用模型：Skip-Gram with Negative Sampling (SGNS)

1.1.尽可能自己完整写代码，不行的话可以用库。

1.2模型的所有超参数可以自定，需要在说明文档中完整说明

2、模型训练语料：

2.1.小规模已切分数据：

train.txt

2.2.大规模未切分数据：

2.2.1 wiki数据：https://dumps.wikimedia.org/backup-index.html

2.2.2搜狗数据：http://www.sogou.com/labs/resource/cs.php

请依据计算能力适量选择采用数据。

3、词相似度计算：

基于学习到的词向量计算pku\_sim\_test.txt中同一行中两个词的相似度。对于pku\_sim\_test.txt中同一行的两个词wi和wj，设两个词的词向量分别为Vi和Vj。

3.1先计算Vi和Vj的欧式距离，设为dij，设max和min为所有同行两个词间dij的最大值和最小值，

3.2如下计算同行两个词间的相似度：

sim(wi,wj)= (10\*max-min -9\*dij) / (max-min)

4、结果输出要求(因为是机器判定，请一定按如下格式输出)：

4.1 输出文件的编码：utf-8 (注意，上述train.txt、pku\_sim\_test.txt是GBK编码的)

4.2 输出格式：每行一组词及其相似度，词与词之间、词与相似度值之间用一个tab符分开，如下例：

没戏 没辙 4.3

4.3 不要打乱pku\_sim\_test.txt中原来词对的行序

4.4 当pku\_sim\_test.txt中某行中存在任一个词没有获得词向量时，对应的该行的词间相似度为1，即，如果下面一行“没辙”或“没戏”两个词中任一个词没有词向量，则输出：

没戏 没辙 1

4.5 输出文件以自己的学号命名。