词向量编程作业：汉语词向量

作业内容：
一、概述：分别基于 SVD 分解以及基于 SGNS 两种方法构建汉语词向量并进行评测。

二、具体说明：
1、语料：training.txt。
2、基于 SVD 分解的方法：获取高维 distributional 表示时 K=5，SVD 降维后的维数自定，获得子词向量 vec_sta。之后基于该向量计算 pku_sim_test.txt 中同一行中两个子词的余弦相似度 sim_svd。当 pku_sim_test.txt 中某一个词没有获得向量时(该词未出现在该语料中)，令其所在行的两个词之间的 sim_svd=0。
3、基于 SGNS 的方法：SGNS 方法中窗口 K=2，子词向量维数自定，获得向量 vec_sgns。之后基于该子词向量计算 pku_sim_test.txt 中同一行中两个词的余弦相似度 sim_sgns。当 pku_sim_test.txt 中某一个词没有获得向量时(该词未出现在该语料中)，令其所在行的两个词之间的 sim_sgns=0。
4、两种方法的结果输出要求(因为是机器判定，请一定按如下格式输出)：
4.1 保持 pku_sim_test.txt 编码(utf-8)不变，保持原文行序不变
4.2 每行在行末加一个 tab 符之后写入该行两个词的 sim_sv，再加一个 tab 符之后写入该行两个词的 sim_sgns。
4.3 输出文件命名方式：学号。
5、所有输出文本均采用 Unicode(UTF-8)编码
6、算法采用 Python (3.0 以上版本) 实现

三、作业提交：
1、提交方式：学校教学平台
2、提交时间：见平台上的时间要求
3、提交内容：
3.1、算法说明文件

- 提交 doc(或 pdf)文件，文件命名方式：学号；说明中分别对两个方法的模型参数和执行细节进行说明。模型参数和执行细节应至少包含：
  - 对于 SVD 方法：总共有多少个非零奇异值，选取了多少个奇异值，选取的奇异值之和、全部奇异值之和以及二者的比例；SVD 分解的算法详述。
  - 对于 SGNS 方法：所用初始词向量来源、词向量维数、训练算法的学习率、训练批次大小、训练轮数等；

这个说明文档的主要目的至少包含两个方面：其一增强读者对你程序设计思路的认识，从而帮助读者对你代码的理解，其二表明你完全了解代码的设计思路和实现过程，你对算法的代码实现做了预先的设计。
3.2、完整的实现代码，其中关键部分需要进行注释说明：与文本说明中的参数和执行细节对应。
3.3、相似度输出文件
提交包含基于两种方法所计算的相似度的 txt 文件，文件命名方式：学号。

参考资料：

1、论文：

[Mikolov2013ICLRworkshop]Tomas Mikolov, Greg Corrado, Kai Chen, Jeffrey Dean, Efficient Estimation of Word Representation in Vector Space, ICLR2013 workshop.

[Mikolov2013NIPS]Tomas Mikolov, Ilya Sutskever, Kai Chen. Distributed Representations of Words and Phrases and their Compositionality, Advances in Neural Information Processing Systems. 2013.

2、代码

SGNG 有很多版本的实现代码：

1、来自:https://paperswithcode.com/

https://paperswithcode.com/search?q_meta=&q=distributed-representations-of-words-and-phrases-and-their-compositionality

https://github.com/theeluwin/pytorch-sgns

2、https://github.com/fanglanting/skip-gram-pytorch

A complete pytorch implementation of skipgram model (with subsampling and negative sampling).

3、gensim

Gensim 库有 Word2Vec：SGNS

建议自己先尽力尝试编码实现。