一、概述：分别基于SVD分解以及基于SGNS两种方法构建词向量并进行评测。

二、具体说明：

1. 语料：词向量学习语料采用text8.txt
2. 评测：

采用wordsim353\_agreed.txt

对wordsim353\_agreed.txt 每行中的两个英语单词(即每行中的第二、三列的串，第一和第四列的内容不用管)，利用所学习到的词表示进行相似度计算，采用余弦相似度计算两个词向量的相似度。如果没有获得某个词的词表示，则相似度为0.

3、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wordsim353\_agreed.txt中同一行中两个子词的余弦相似度sim\_svd。当wordsim353\_agreed.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。

4、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wordsim353\_agreed.txt中同一行中两个词的余弦相似度sim\_sgns。当wordsim353\_agreed.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。

5、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)：

5.1 保持wordsim353\_agreed.txt编码不变，保持原文行序不变

5.2 每行在行末加一个tab符之后写入该行两个词的sim\_svd，再加一个tab符之后写入该行两个词的sim\_sgns。

5.3 输出文件命名方式：学号。

6算法采用Python (3.0以上版本) 实现

7计算资源：Google Colab 的GPU 可以考虑：免费的以及  9.9美元包月的

三、作业提交：

1、提交时间：5月31日24:00之前提交作业(以收到邮件时间为准)。

2、提交方式：云邮教学空间。

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。