Passage retrieval 段落检索

输入接口：

Question.txt：一个问题句子

Keywords.txt: 一个问题句子中拆出的关键词

Article.txt：根据一个问题搜索得到的wiki文档

输出接口：

Sentence.txt: 筛选出的可能包含答案的一堆句子

Score.txt: sentence.txt中的句子的对应相似度得分

执行过程：

1. 读入文本文件。按照”。？！”作为分句符号进行分句，得到list\_sentence。
2. 筛选出list\_sentence中包含关键字的句子，得到important\_sentence
3. 利用list\_sentence训练tfidf模型/lsi模型/lda模型（在Passage\_Retrieval.py 75-77行设置）
4. 基于sentence2vec，计算importance\_sentence中的句子向量和question的句子向量
5. 计算两个向量的余弦相似度，作为输出句子的相似度的得分

参考资料：

http://blog.csdn.net/wds2006sdo/article/details/51059737

<https://github.com/WenDesi/sentenceSimilarity>

TODO：

现在的模型的训练集是原来的wiki内容+问题，后面改一下训练集？

丁可 2017.12.2