文本表征学习

作业2：Word2vec

EE1513

宋彦

作业2：Word2vec

- 概述：使用训练数据以及word2vec算法训练词向量并对其性能进行评价，对比分析不同word2vec算法的性能，完成并提交1-2页的实验报告

- 使用 enwiki8 (http://mattmahoney.net/dc/enwik8.zip) 语料库，以及下面四种算法训练 word2vec 词向量 (https://code.google.com/archive/p/word2vec/source/default/source)
  - (1) HS + CBOW:(2) HS + SG:(3) NS + CBOW:(4) NS +SG

- 其他超参数
  - 向量维度:200
  - 窗口大小:+/- 5词
  - 词频过滤:5
- 使用Spearman's correlation评估上述四种算法的性能

作业2：Word2vec

- 实验报告至少应包括
  - 四种方法训练得到的词向量的性能结果
  - 结合你对四种方法结果的观察，比较分析HS与NS，以及CBOW和SG
  - 简要描述处理数据和实现模型的过程，其中使用了哪些数据结构，遇到了哪些问题，是如何解决的

- 建议使用PDF作为提交格式
- 作业文件命名：学号 + “_” + 姓名 + “_作业2”
  - 例如：PB12345678_张三_作业2
- 作业提交时间：4月25日00:00之前