lab2 locality-sensetive hashing

20337025 崔璨明

数据集见`data.txt`，代码见`lsh.py`，直接运行即可。

数据集

采用的数据集合是人工智能课程上使用的情感文本数据集，去除了情感标签，只留下句子，总共有1222个句子。

实验结果

实验目的为找到数据集中相似的文本对，打印相似度前十的文本对并输出相似度大于0.65的文本对：

```
10 most similar text pairs:

1. NO.180 and NO.199, similarity:0.8508928921004846
toddler di from taint spinach  <---->  toddler di from e coli taint spinach

2. NO.91 and NO.92, similarity:0.7750597222706329
photograph kidnap condemn  <---->  photograph kidnap in gaza

3. NO.2 and NO.6, similarity:0.7156006318835887
nigeria hostag fear dead is freed  <---->  kate is marri doherti

4. NO.11 and NO.25, similarity:0.6946338007398866
nicol kidman ask dad to help stop husband s drink  <---->  hacker unlock appl music download protect

5. NO.32 and NO.62, similarity:0.6784542481569009
more human remain found at ground zero  <---->  tumor type mai explain surviv rate for cancer

6. NO.0 and NO.62, similarity:0.6780609400935718
mortar assault leav at least dead  <---->  tumor type mai explain surviv rate for cancer

7. NO.1 and NO.3, similarity:0.6739154761568333
goal delight for sheva  <---->  bomber kill shopper

8. NO.1 and NO.8, similarity:0.6727049923337955
goal delight for sheva  <---->  happi birthdai ipod

9. NO.3 and NO.8, similarity:0.665131769417656
bomber kill shopper  <---->  happi birthdai ipod

10. NO.0 and NO.2, similarity:0.6617816010587484
mortar assault leav at least dead  <---->  nigeria hostag fear dead is freed
```

```
All pairs that similarity>0.65(totally 13):

(180, 199) (91, 92) (2, 6) (11, 25) (32, 62) (0, 62) (1, 3) (1, 8) (3, 8) (0, 2) (7, 8) (0, 1) (10, 22)
```