Assignment2：实现LSH算法

作业要求

1. 实现LSH算法（从Shingling到Locality-Sensitive Hashing完整步骤）
2. 文档数据集对LSH的效果进行测试，设置k=10，相似性阈值0.8，通过调整M、b、r分析LSH的假阴性和假阳性

作业过程

一、 实现LSH算法

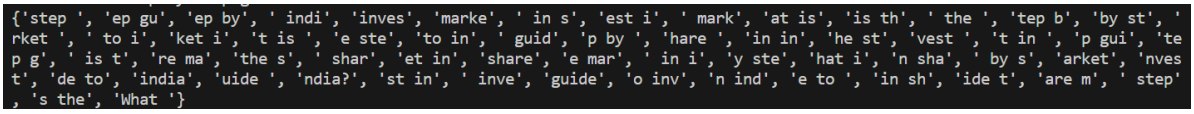
1. Shingling

编写代码将输入文本转化为长度为k的字符串组，使用set类型存储，去除重复的子串

```
def shingle(text, k):
    shingle_set = []
    for i in range(len(text) - k+1):
        shingle_set.append(text[i:i+k])
    return set(shingle_set)
```

输入句子`What is the step by step guide to invest in share market in india?`

输出结果如下：



2. Min Hashing

在进行Min Hashing之前，将shingling转化为0-1向量，过程：1.构建词库；2.0-1向量每个位置代表词库中的每个字符串，如果shingling存在该位置对应的字符串，则将该位置置为一。具体见函数`build_vocabulary`以及`shingles_to_onehot`，报告中不过多阐述，具体介绍Min Hashing，见代码注释：

```
def minhash(one_hot_vector, num_hashes):
    signatures = [[],[]]    # 数据集只对比两条文本
    for _ in range(num_hashes):
        # 生成1~[词库数量]的随机序列
        hash_ex = list(range(1, len(one_hot_vector[0])+1))
        random.shuffle(hash_ex)
        for j in range(len(one_hot_vector)): # len(one_hot_vector)其实就是2
            for i in range(1, len(one_hot_vector[j])+1):    # 从最小索引开始查找
                idx = hash_ex.index(i)
                if one_hot_vector[j][idx] == 1:          # 直到找到第一个对应值为1的0-
1向量值
                    signatures[j].append(idx)            # 为对应文本添加对应的索引
                    break
    return signatures
```

以下是输入样本，得到的0-1向量，以及最后通过min Hashing得到的signatures:

```
first text:What is the step by step guide to invest in share market in india?
second text:What is the step by step guide to invest in share market in india?
one-hot
[array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]), array([1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0,
       1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1,
       1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1])]
signatures:
[[13, 18, 23, 31, 57, 19, 54, 6, 9, 45], [13, 18, 16, 31, 57, 19, 54, 6, 9, 45]]
```

3. Locality-Sensitive Hashing

对前面得到的Signature进行比对前的哈希操作，主要操作是对signature进行分成bands个段，每个段共rows行

```
def locality_sensitive_hashing(signatures, bands, rows):
    buckets = []
    for i in range(0, len(signatures), rows):
        band = tuple(signatures[i:i + rows])
        hash_band = hash(band) % (10**9 + 7)
        buckets.append(hash_band)
    return buckets
```

设置bands=5，rows=2，结果如下：

```
first text:What is the step by step guide to invest in share market in india?
second text:What is the step by step guide to invest in share market in india?
signatures:
[[6, 53, 14, 15, 17, 22, 28, 47, 22, 42], [6, 1, 14, 15, 17, 22, 28, 47, 22, 42]]
bucket:
[255938783, 625574806, 916026279, 220760504, 882680439]
[494200224, 625574806, 916026279, 220760504, 882680439]
```

通过以上实现函数即可得到两个text的bucket，并进行比对，存在bucket_1[i]==bucket_2[i]即可将两个text作为一个Candidate pair。

二、 测试分析

使用同学热心提供的`questions.csv`作为测试文档数据集，设置k=10，相似性阈值0.8，调整M、b、r分析LSH的假阴性和假阳性

定义函数`compute_similarity`计算Jaccard distance得到`sim(C1,C2)`,如果一个Candidate pair的`sim(C1,C2)`低于阈值0.8，则视为false_positive，如果一个非Candidate pair的`sim(C1,C2)`超过阈值0.8，则视为false_negative，数据集展示如下，只使用到question1和question2两列

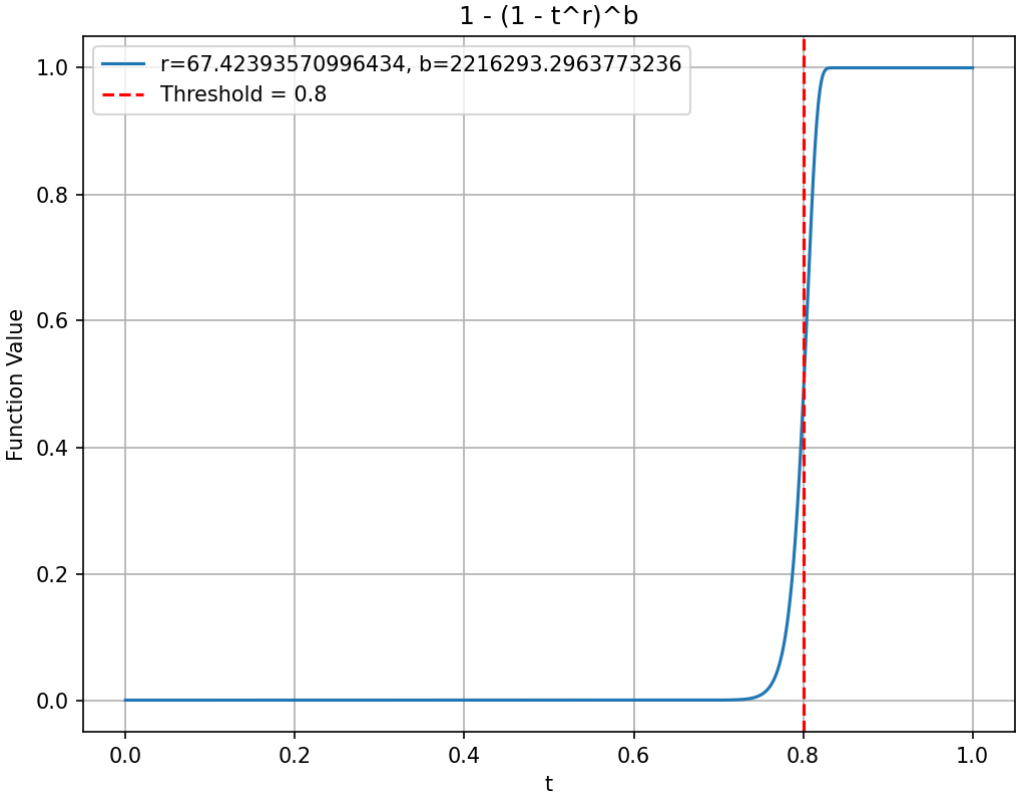
id	qid1	qid2	question1	question2	is_du
0	1	2	What is the step by step guide to invest in share marke	What is the step by step guide to invest in share market?	
1	3	4	What is the story of Kohinoor (Koh-i-Noor) Diamond?	What would happen if the Indian government stole the Kohinoor (Koh-i-Noor) di	
2	5	6	How can I increase the speed of my internet connection	How can Internet speed be increased by hacking through DNS?	
3	7	8	Why am I mentally very lonely? How can I solve it?	Find the remainder when [math]23^{24}[/math] is divided by 24,23?	
4	9	10	Which one dissolve in water quikly sugar, salt, methane	Which fish would survive in salt water?	
5	11	12	Astrology: I am a Capricorn Sun Cap moon and cap rising	I'm a triple Capricorn (Sun, Moon and ascendant in Capricorn) What does this	
6	13	14	Should I buy tiago?	What keeps childern active and far from phone and video games?	
7	15	16	How can I be a good geologist?	What should I do to be a great geologist?	

对样本进行测试，调整M、b、r，只使用到数据集的10000个样本，由图表可得，随着rows的增加，false_negative会随之增加，false_positive变少；随着bands的增加，false_negative减少，false_positive增加

bands	rows	false_negative	false_positive
5	10	48	71
5	15	76	12
5	20	87	6

bands	rows	false_negative	false_positive
10	15	52	35
15	15	41	45
20	15	27	53

通过优化最接近阈值s=0.8对应$1-(1-t^r)^b$图像中的b、r值如下图，数值过大，反而引入了更大的计算量，因此b、r的优化也需要考虑与此引起的数据量均衡



总结

本次作业通过代码实现了LSH算法从shingling到Locality-Sensitive-Hashing的总体过程，对加深LSH算法的印象和理解有很大作用，一方面，LSH算法减少了大量数据的比对，另一方面又引入了false_negative和false_positive的问题，通过调整bands和rows，能在一定程度上减少false_negative和false_positive出现的概率。