CS 6140: Data Mining Assignment 2

Document Similarity and Hashing

Yash Gangrade (u1143811), MS in Computing 30^{th} January 2019

Contents

1 Creating k-grams

1.1 Distinct k-grams for each document

Ans: The results are summarized in form of a table below:

Document	G1	G2	G3
D1	266	815	309
D2	265	804	308
D3	258	757	295
D4	259	771	261

1.2 Jaccard Similarity between pair of documents

Ans: We know that $JS(A, B) = \frac{|A \cap B|}{|A \cup B|}$. The results are summarized in form of a table below:

Document	G1	G2	G3
D1 & D2	0.9962	0.9648	0.7679
D1 & D3	0.9124	0.7313	0.2797
D1 & D4	0.7327	0.3556	0.0124
D2 & D3	0.9157	0.7402	0.3052
D2 & D4	0.7294	0.3496	0.0161
D3 & D4	0.6951	0.3510	0.0146

2 Min Hashing

2.1 Reporting JS for multiple values of t

Ans: The Jaccard Similarity results for Document 1 and Document 2 using Min Hashing for different values of t along with the the time taken are summarized below:

t	Jaccard Similarity	Time
20	1.0	0.0503
60	0.9833	0.1032
150	0.9667	0.2829
300	0.9633	0.5388
600	0.965	1.1282

2.2 Finding Optimal Value of t

Here, we want to find an optimal value of t which would get you both good accuracy and low running time. To find the optimal value of t, we are running more experiments with multiple values of t by fine graining it. We are then plotting three plots which are, first is Jaccard Similarity vs Number of Hash Functions (t), second is Jaccard Similarity vs Time Taken in the process, and the third is Time Taken vs Number of Hash Functions (t). The graphical results are attached in the next page. Also, the good approximation of t should be close to the t value calculated in part 1b i.e. 0.9648

From these plots, after careful observation, we find that the t value of **300** is the best value for a good approximation. Since, it provides us the Jaccard Similarity of 0.9667 and the time taken is only 0.646 secs which is optimal.

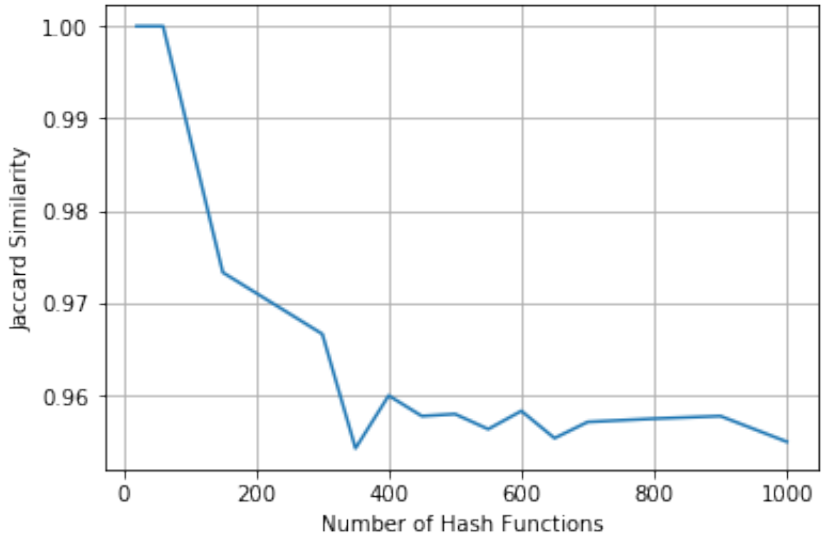

Figure 1: Jaccard Similarity vs Number of Hash Functions (t)

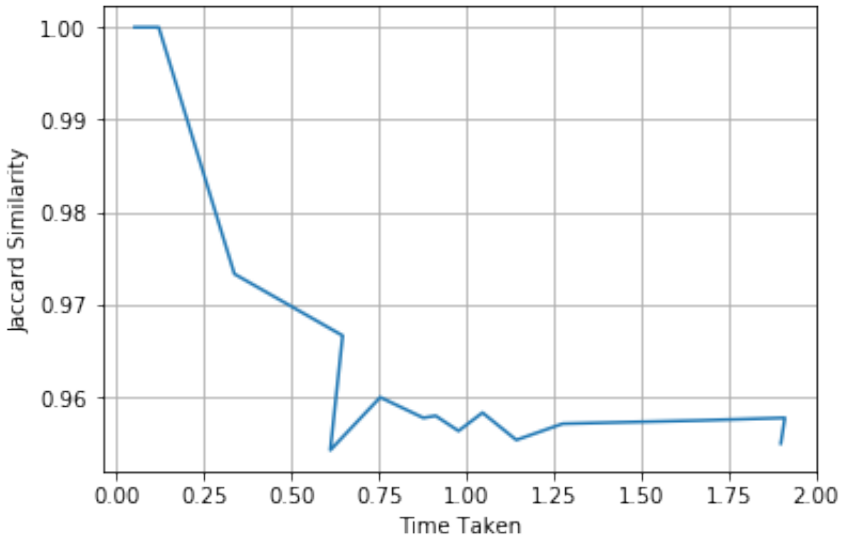

Figure 2: Jaccard Similarity vs Time Taken

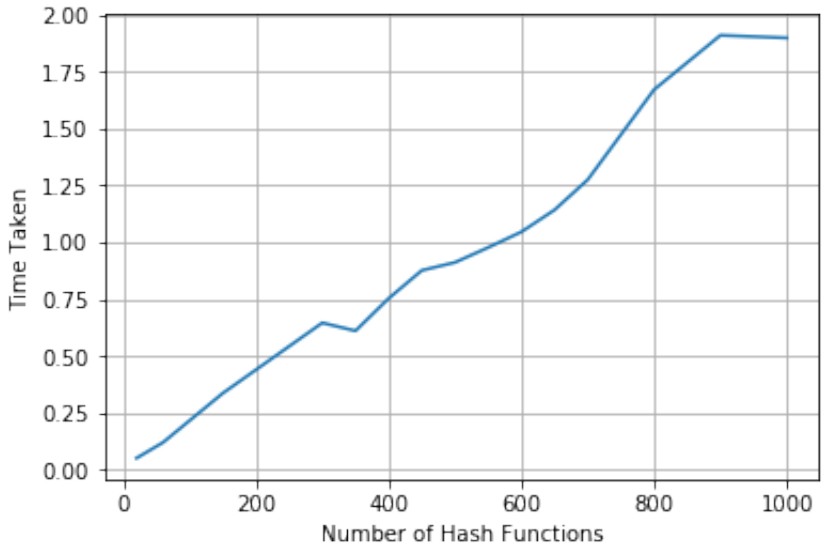

Figure 3: Time Taken vs Number of Hash Functions (t)

3 Bonus

Here, to get an intuition and understanding of Jaccard and Andberg Similarity. We will compare them by extending the experiments in Question 1 to accommodate the Andberg similarity. We are only gonna perform G2 experiments for both type of similarities to get an idea of comparison. The results are summarized in the table below. So we have

$$JS(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

$$Andberg(A,B) = \frac{|A \cap B|}{|A \cup B| + |A \Delta B|}$$

Document	**Jaccard Similarity (G2)**	**Andberg Similarity (G2)**
D1 & D2	0.9648	0.9320
D1 & D3	0.7313	0.5764
D1 & D4	0.3556	0.2162
D2 & D3	0.7402	0.5877
D2 & D4	0.3496	0.2118
D3 & D4	0.3510	0.2129

So from the above table it is clear that the Jaccard Similarity and Andberg Similarity are very different irrespective of the hash functions if the symmetric difference between the sets is at the same level of significance as the union and intersection. So the process like min-hashing will work in the same way for Andberg similarity as it works for Jaccard similarity provided $|A \cup B| >> |A \Delta B|$. If not, such a process cannot be done.