Asignment 2: Document Similarity and Hashing

Christopher Mertin/u1010077
cmertin@cs.utah.edu
February 12, 2017

Overview

In this assignment you will explore the use of k-grams, Jaccard distance, min hashing, and LSH in the context of document similarity.

You will use four text documents for this assignment:

- http://www.cs.utah.edu/~jeffp/teaching/cs5140/A2/D1.txt
- http://www.cs.utah.edu/~jeffp/teaching/cs5140/A2/D2.txt
- http://www.cs.utah.edu/~jeffp/teaching/cs5140/A2/D3.txt
- http://www.cs.utah.edu/~jeffp/teaching/cs5140/A2/D4.txt

As usual, it is highly recommended that you use LaTeX for this assignment. If you do not, you may lose points if your assignment is difficult to read or hard to follow. Find a sample form in this directory: http://www.cs.utah.edu/~jeffp/teaching/latex/

1 Creating k-Grams (40 points)

You will construct several types of k-grams for all documents. All documents only have at most 27 characters: all lower case letters and space. *Yes, the space counts as a character in character k-grams.*

[G1] Construct 2-grams based on characters, for all documents.

[G2] Construct 3-grams based on characters, for all documents.

[G3] Construct 2-grams based on words, for all documents.

Remember, that you should only store each k-gram once, duplicates are ignored.

A: (20 points) How many distinct k-grams are there for each document with each type of k-gram? You should report $4 \times 3 = 12$ different numbers.

File	k_2-Character	k_3-Character	k_2-Word
D1.txt	331	1299	521
D2.txt	361	1516	632
D3.txt	354	1543	841
D4.txt	298	1025	413

B: (20 points) Compute the Jaccard similarity between all pairs of documents for each type of k-gram. You should report $3 \times 6 = 18$ different numbers.

Table 1: k_2-Character Jacard Similarities

	D1.txt	D2.txt	D3.txt	D4.txt
D1.txt	1.000	0.845	0.770	0.705
D2.txt	0.845	1.000	0.761	0.707
D3.txt	0.770	0.761	1.000	0.720
D4.txt	0.705	0.707	0.720	1.000

Table 2: k_3-Character Jacard Similarities

	D1.txt	D2.txt	D3.txt	D4.txt
D1.txt	1.000	0.639	0.460	0.327
D2.txt	0.639	1.000	0.440	0.312
D3.txt	0.460	0.440	1.000	0.362
D4.txt	0.327	0.312	0.362	1.000

Table 3: k_2-Word Jacard Similarities

	D1.txt	D2.txt	D3.txt	D4.txt
D1.txt	1.000	0.257	0.033	0.005
D2.txt	0.257	1.000	0.025	0.006
D3.txt	0.033	0.025	1.000	0.012
D4.txt	0.005	0.006	0.012	1.000

2 Min Hashing (30 points)

We will consider a hash family $\mathcal{H}$ so that any hash function $h \in \mathcal{H}$ maps from $h : \{k\text{-grams}\} \to [m]$ for m large enough (To be extra cautious, I suggest over $m \geq 10{,}000$).

A: (25 points) Using grams G2, build a min-hash signature for document D1 and D2 using $t = \{20, 60, 150, 300, 600\}$ hash functions. For each value of t report the approximate Jaccard similarity between the pair of documents D1 and D2, estimating the Jaccard similarity:

$$\hat{\mathsf{JS}}_t(a,b) = \frac{1}{t}\sum_{i=1}^{t}\begin{cases}1 & \text{if } a_i = b_i\\ 0 & \text{if } a_i \neq b_i.\end{cases}$$

You should report 5 numbers.

t	J_S
20	0.750
60	0.667
150	0.633
300	0.640
600	0.612

B: (5 point) What seems to be a good value for t? You may run more experiments. Justify your answer in terms of both accuracy and time.

From the direct calculation, we know that the Jacard Estimate between documents 1 and 2 using 3-grams is 0.639. Looking at the estimated Jacard Similarities above $t = 150$ seems to be the best bet. To check, the value of t was increased up to 1900 to check the variation of the Jacard Similarity Estimate. As the table below shows, there was limited variation in the values.

t	J_S
500	0.642
700	0.639
900	0.651
1100	0.638
1300	0.632
1500	0.629
1700	0.668
1900	0.635

3 LSH (30 points)

Consider computing an LSH using $t = 160$ hash functions. We want to find all documents pairs which have Jaccard similarity above $\tau = 0.4$.

A: (8 points) Use the trick mentioned in class and the notes to estimate the best values of hash functions b within each of r bands to provide the S-curve

$$f(s) = 1 - (1 - s^b)^r$$

with good separation at τ. Report these values.

The function $f(s) = 1 - \left(1 - s^b\right)^r$ is steepest at the point of inflection $\tau = (1/r)^{1/b}$. If we want to use s as our cut where $\tau = s = \alpha = 1 - \beta$ we can substitute and get $\tau = (b/t)^{1/b}$ which can be approximated as $b \approx -\log_\tau(t)$. Therefore, with $t = 160$ and $\tau = 0.4$, we have $b = 5.539 = -\log_\tau(t) = -\frac{\log_{10}(t)}{\log_{10}(\tau)}$. From $b = t/r$ we have $r = t/b = 160/5.539 = 28.886$. Therefore, the S-curve is given by

$$f(s) = 1 - \left(1 - s^{5.539}\right)^{28.886}$$

B: (24 points) Using your choice of r and b and $f(\cdot)$, what is the probability of each pair of the four documents (using [G2]) for being estimated to having similarity greater that τ? Report 6 numbers.

The probability of documents being estimated for using 3-grams have a similarity greater than $\tau = 0.4$ is given by the above equation, where s is the Jacard Similarity between the two documents.

$f(D_i, D_j)$	Probability/100
f(`D1.txt`, `D2.txt`)	0.999
f(`D1.txt`, `D3.txt`)	0.334
f(`D1.txt`, `D4.txt`)	0.022
f(`D2.txt`, `D3.txt`)	0.141
f(`D2.txt`, `D4.txt`)	0.036
f(`D3.txt`, `D4.txt`)	0.040

4 Bonus (3 points)

Describe a scheme like Min-Hashing for the *Andberg Similarity*, defined $\mathsf{Andb}(A, B) = \frac{|A \cap B|}{|A \cup B| + |A \triangle B|}$. So given two sets A and B and family of hash functions, then $\mathbf{Pr}_{h \in \mathcal{H}}[h(A) = h(B)] = \mathsf{Andb}(A, B)$. Note the only randomness is in the choice of hash function h from the set $\mathcal{H}$, and $h \in \mathcal{H}$ represents the process of choosing a hash function (randomly) from $\mathcal{H}$. The point of this question is to design this process, and show that it has the required property.

Or show that such a process cannot be done.

The similarity is also known as S-Dice Similarity and the equation can be simplified as $S_D(A, B) = 2\frac{|A \cap B|}{|A| + |B|}$. This can be used to get a fairly accurate approximation using a similar algorithm to the min hashing algorithm. The two differences are that (i) the Jacard Similarity and the S-Dice Similarity we must address the multiplication of 2, and (ii) the larger denominator in the S-Dice Similiarity requires $|A \cap B| \leq |A| + |B|$ as elements may exist in one set but not the other. The inequality is larger the more similar the sets are, and the adapted algorithm for S-Dice account for the duplicate values given a large enough number of iterations.

In the min hashing algorithm, there are three types of rows (i) x rows with 1 in both columns, (ii) there are y rows with a 1 in one column and a 0 in the other, and (iii) there are zeros in both. The Jacard Similarity is $J_s(A, B) = x/(x + y)$ while the S-Dice similarity is $S_D(A, B) = 2x/(2x + y)$.

1. Build a matrix as described in the min hash algorithm where the i^{th} and j^{th} columsn is 1 if the element $s_{i,j}$ is in a set S_j and 0 otherwise. *In addition*, when building the matrix, if both $s_{i,j} = s_{i,m} \in S_j$ and S_m, then add other duplicate row for the element $s_{i,j}$. We can then randomly permute the columns of this matrix.

2. For each column, find the value fo the map function $M(S)$ which is equal to the value of the element which appears earliest in the rows, *i.e.* $M(S_j) = s_{i,j}$ such that $\min_i(x_{i,j} \neq 0)$ for $s_{i,j} \in S_j$ and $x_{i,j}$ equal the values at index (i, j).

3. Estimate the S-Dice similarity as $S_D(S_j, S_m) = \begin{cases} 1 & M(S_j) = M(S_M) \\ 0 & \text{else} \end{cases}$

The above works because there are a total of $2x + y + z$ and therefore row r is of the type where both columsn are 1 with probability $2x/(2x + y)$ since we can ignore the large z number of rows. With k random permutations, we get a set $\{m_1, m_2, \ldots, m_k\}$ and k random variables $\{X_1, X_2, \ldots, X_k\}$ and we can estimate $S_D(S_j, S_m)$ as $S_D(S_j, S_m) = \frac{1}{k}\sum_{\ell=1}^{k} X_\ell$