Asmt 2: Document Similarity and Hashing

Turn in through Canvas by 5pm, then come to class:
Wednesday, February 19
20 points

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 (8 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. Unfortunately, the documents have a few extra characters (e.g. 2,1,0,8, newline), please ignore them. Sorry.

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

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

[G3] Construct 3-grams based on words, for all documents.

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

A: (4 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.

B: (4 points) Compute the Jaccard distance between all pairs of documents for each type of k-gram. You should report $3 \times 6 = 18$ different numbers.

2 Min Hashing (6 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 (I suggest over $m \geq 10{,}000$).

A: (5 points) Using grams G2, build a min-hash signature for document `D1` and `D2` using $t = \{10, 50, 100, 250, 500\}$ hash functions. For each value of t report the Hamming similarity between the pair of documents `D1` and `D2`, estimating the Jaccard similairty:

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

You should report 5 numbers.

B: (1 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.

3 LSH (6 points)

Consider computing an LSH using $m = 100$ hash functions. We want to find all documents which have Jaccard similarity above $\tau = .3$.

A: (4 points) Use the trick mentioned in class and the notes to estimate the best values of rows r in each of b blocks to provide the S-curve

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

with good separation at τ. Report these values.

B: (2 points) Using your choice of r and b and $S(\cdot)$, what is the probability ~~that you will need to check the exact Jaccard similarity~~ of each pair of the four documents using G2 for being estimated for having similarity greater that τ? Report 6 numbers. *(Show your work.)*

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.