Asmt 2: Document Similarity and Hashing

Turn in a hard copy at the start of class:
Wednesday, February 20
20 points

Overview

In this assignment you will explore the use of shingling, 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/cs5955/A2/D1.txt`
- `http://www.cs.utah.edu/~jeffp/teaching/cs5955/A2/D2.txt`
- `http://www.cs.utah.edu/~jeffp/teaching/cs5955/A2/D3.txt`
- `http://www.cs.utah.edu/~jeffp/teaching/cs5955/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 Shingling (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.

[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 shingle once, duplicates are ignored.

A: How many distinct k-grams are there for each document with each type of k-gram?

B: 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 $m = 40127$. You are free to choose any hash function you want. If you want to implement your own, use a multiplicative hash (as described in notes).

A: Using grams G2, build a min-hash signature for document `D1` and `D2` using $t = \{10, 50, 100, 300, 600\}$ hash functions. For each value of t report the normalized L_0 distance between the pair of documents `D1` and `D2`, estimating the Jaccard distance:

$$\text{normalized-}L_0(a,b) = 1 - \frac{1}{t}\sum_{i=1}^{t} \begin{cases} 1 & \text{if } a_i = b_i \\ 0 & \text{if } a_i \neq b_i \end{cases}$$

You should report 5 numbers. Had given estimate for Jaccard Similarity above before. Either estimate will be accepted.

B: 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 $\tau = .25$.

A: 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: 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 having similarity greater that τ? Report 6 numbers. *(Show your work.)*

4 Bonus (3 points)

Show that the Jaccard distance $\mathtt{d_J}(A, B) = 1 - \mathsf{JS}(A, B)$ is a metric, or provide a counter example that it is not a metric.