Introduction to locality sensitive hashing

Andee Kaplan

Duke University
Department of Statistical Science
andrea.kaplan@duke.edu

February 8, 2018

Slides available at http://bit.ly/cimat-lsh

Goal and outline

Goal: Introduce locality sensitive hashing, a fast method of blocking for record linkage, and get some experience doing LSH in R.

1. Defining similarity

2. Representing data as sets (shingling)

3. Hashing

4. Hashing with compression (minhashing)

5. Too many pairs to compare! (LSH)

6. Evaluation

Finding similar items

- We want to find similar items

    - Maybe we are looking for near duplicate documents (plagiarism)

    - More likely, we are trying to block our data which we can later pass to a record linkage process

- How do we define *similar*?

Jaccard similarity

There are many ways to define similarity, we will use *Jaccard similarity* for this task.
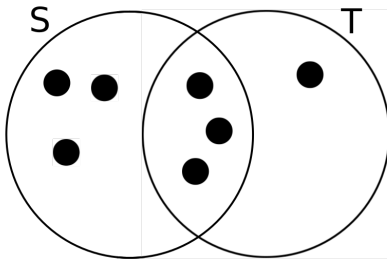
$$Jac(S, T) = \frac{|S \cap T|}{|S \cup T|}$$

Figure: Two sets S and T with Jaccard similarity 3/7. The two sets share 3 elements in common, and there are 7 elements in total.

How to represent data as sets

We want to talk about similarity of data $\Rightarrow$ we need sets to compare!

- One way is to construct from the data the set of **short strings** that appear within it

- Similar documents/datasets will have many common elements, i.e. many commong short strings

- We can do construct these short strings using *shingling*

k-shingling (how-to)

1. Think of a document or record as a string of characters

2. A *k*-shingle (k-gram) is any sub-string (word) of length *k* found within the document or record

3. Associate with each document or record the set of *k*-shingles that appear one or more times within it

Let's try

Suppose our document is the string "Hello world" and $k = 2$, then

- the set of 2-shingles is {he, el, ll, lo, ow, wo, or, rl, ld}

- the set of 3-shingles is {hel, ell, llo, low, owo, wor, orl, rld}

Your turn

We have the following two records:

	First name	Last name
129	MICHAEL	VOGEL
130	MICHAEL	MEYER

1. Compute the 2-shingles for each record

2. Using Jaccard similarity, how similar are they?

Your turn solution

1. The 2-shingles for the first record are
   {mi, ic, ch, ha, ae, el, lv, vo, og, ge, el} and for the second are
   {mi, ic, ch, ha, ae, el, lm, me, ey, ye, er}.

2. There are 6 items in common {mi, ic, ch, ha, ae, el} and 16
   items total
   {mi, ic, ch, ha, ae, el, lv, vo, og, ge, el, lm, me, ey, ye, er}, so
   the Jaccard similarity is $\frac{6}{16} = \frac{3}{8} = 0.375$

Useful packages/functions in R

Example data

Your turn

shinging and Jaccard similarity with R

Hashing

Why do it?

Similarity preserving summaries of sets

Characteristic matrix

Minhashing

LSH (avoid pairwise comparisons)

Banding and buckets

banding in R

Putting it all together

"Easy" LSH in R

Evaluation

Your turn

perform LSH and evaluate how we did