Module 4: Probabilistic Blocking, Part II (Extra Details)

Rebecca C. Steorts

Locality Sensitive Hashing (LSH) to the Rescue

We want to hash items several times such that similar items are more likely to be hashed into the same bucket.

1. Divide the **signature matrix** into b bands with r rows each so $m = b * r$ where m is the number of times that we drew a permutation of the characteristic matrix in the process of minhashing
2. Each band is hashed to a bucket by comparing the minhash for those permutations
   - If they match within the band, then they will be hashed to the same bucket
3. **If two documents are hashed to the same bucket they will be considered candidate pairs**

We only check *candidate pairs* for similarity

Candidate Pairs

In order to avoid looking at all-to-all comparisons, we instead will check *candidate pairs* of records that are hashed to the same bucket.

Goals:

1. dis-similar pairs will never hash to the same bucket (and never checked).
2. dis-similar pairs hashed to the same bucket are false positives.
3. hope that truly similar pairs hash to the same bucket. Those that do not are false negatives.

Banding Signature Matrix

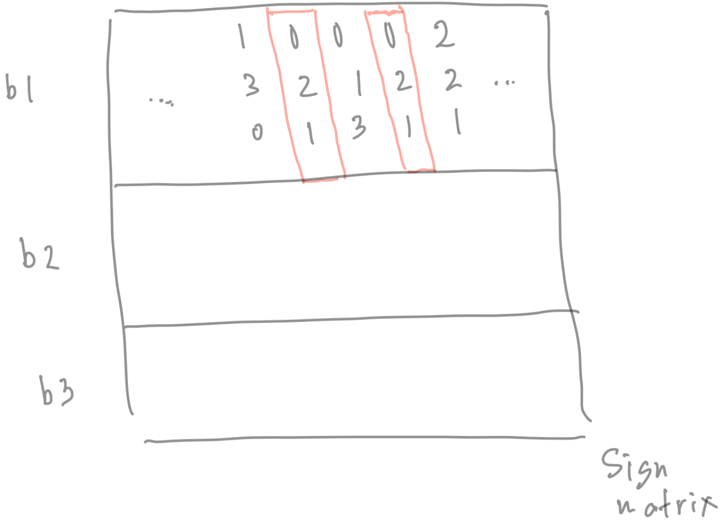

Figure 1: Signature matrix divided into 3 bands with 3 rows ber band.

Banding Technique

- Assume b bands with r rows per band. Each record pair had Jaccard similarity s.
- The minhash signature for a record pair in a row of the signature matrix is s.

Banding Technique

We can calculate the probability that record pairs (their signatures) become a candidate pairs as follows:

1. The probability the signatures agree in all rows of one particular band is s^r.
2. The probability that the signatures do not agree in at least one row of a particular band is $1 - s^r$.
3. The probability the signatures do not agree in all rows of any of the bands is $(1 - s^r)^b$.
4. The probability that the signatures agree in all row of at least one band, and thus, is a candidate pair is

$$1 - (1 - s^r)^b.$$

Note that $r = m/b \implies 1 - (1 - s^{m/b})^b$.

Putting it all together

1. Find the set of k-shingles.
2. Pick a length for the min-hash signatures m.
3. Find a threshold t that defines how similar record pairs should be to be considered as a candidate pair. Specifically, choose the number of rows and bands such that $br = m$. The threshold is approximately $(1/b)^{1/r}$.

- If avoidance of false negatives is important, select b and r to produce a lower threshold and than t.
- If avoidance of false positives is important, select b and r to produce a higher threshold than t.

5. Construct candidate pairs using LSH.
6. Filter out any candidate pairs that exceed your threshold. (This step is optional for speeding up computation in terms of filtering to avoid all-to-all comparisons).
7. If you have ground truth, you can compute precision, recall, and reduction ratio.