1 Apr 2 Schedule

- Test: 7:00 - 8:10 (closed book)
- No core-sets
- Chap 1, 4, 3, Linear programming.

2 Locality Sensitive Hashing

Use min hash to compress documents into small signatures, and preserve expected similarity for pairs of document. How do we find pairs with large similarity?

- $n = 10^6$ documents, signature of length 250 (4 bytes each), so 1000 bytes per document, 1KB $\times 10^6 = 1$GB
- #pairs is too large $O(n^2)$: $\frac{1}{2} \times 10^{12}$ pairs

2.1 High level idea of LSH

Use multiple hash function $h_1, h_2, \ldots, h_k$

- $d_i \rightarrow S_i < x_1^1, x_1^2, \cdots >$
- $\sigma_1, \sigma_2, \sigma_3, \ldots, \sigma_{250}$
- $d_j \rightarrow S_j < x_j^1, x_j^2, \cdots >$
- If the first value hashed from $d_i$ and $d_j$ are the same, then Jaccard similarity can be high.

2.2 Recall Jaccard similarity

$$J(d_i, d_j) = \frac{|d_i \cap d_j|}{|d_i \cup d_j|}$$
2.3 Bands

Define $d_i^j$ as entries of $d_i$ from $(k-1)r$. Each hash function is applied to a subset of the entries of a column. And we call such a subset a band. Two bands of document $d_i$ and $d_j$ are considered a match if all entries match.

<table>
<thead>
<tr>
<th>$h_1$ band 1 $r$</th>
<th>$d_1 = \begin{pmatrix} 1 \ 17 \ 4 \end{pmatrix}$</th>
<th>$d_2 = \begin{pmatrix} 2 \ 17 \ 4 \end{pmatrix}$</th>
<th>$\leftarrow$ not match $\rightarrow$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_2$ band 2</td>
<td>$d_1^2 = \begin{pmatrix} 3 \ 17 \ 1 \end{pmatrix}$</td>
<td>$d_2^2 = \begin{pmatrix} 3 \ 17 \ 1 \end{pmatrix}$</td>
<td>$\leftarrow$ matches $\rightarrow$</td>
</tr>
<tr>
<td>$h_3$ band 3</td>
<td>$d_1^3 = \begin{pmatrix} 2 \ 5 \ 6 \end{pmatrix}$</td>
<td>$d_2^3 = \begin{pmatrix} 2 \ 5 \ 7 \end{pmatrix}$</td>
<td>$\leftarrow$ not match $\rightarrow$</td>
</tr>
<tr>
<td>$h_b$ band b</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>

2.4 Analysis

$b$ bands, $r$ rows in each band. Suppose $(x, y)$ two columns have similarity $J(x, y) = s \Rightarrow$ probability that minhash agree on any entry is $s$

- Suppose $J(d_i, d_j) = s$, $0 < s < 1$
- Fix a band $k$, $Pr(d_i^k = d_j^k) = s^r = (0.8)^5 = 0.32768$ is the probability of two bands match in all entries.
- Probability of mismatch of two bands: $1 - s^r$
- Probability that all bands mismatch? $(1 - s^r)^b$
- Probability that at least one match $\geq 1 - (1 - s^r)^b$

2.4.1 Example

- If $b = 16$, $r = 4$, 64 min hash, $(\frac{1}{5})^{1/4} = s = \frac{1}{2}$
- If $b = 20$, $r = 5$,

<table>
<thead>
<tr>
<th>$s$</th>
<th>$1 - (1 - s^5)^{20}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.006</td>
</tr>
<tr>
<td>0.3</td>
<td>0.047</td>
</tr>
<tr>
<td>0.4</td>
<td>0.186</td>
</tr>
<tr>
<td>0.5</td>
<td>0.47</td>
</tr>
<tr>
<td>0.6</td>
<td>0.802</td>
</tr>
<tr>
<td>0.7</td>
<td>0.975</td>
</tr>
<tr>
<td>0.8</td>
<td>0.9996</td>
</tr>
</tbody>
</table>
2.5 Distance Metric?

Function $d(x, y)$ that satisfies:

1. $d(x, x) = 0$
2. $d(x, y) = d(y, x)$
3. $d(x, y) \leq d(x, z) + d(z, y)$

2.5.1 Example

- Jaccard distance $J_d = 1 - J_s$
- Euclidean distance
- Hamming distance
- Edit distance
- LP norm: $L_p(\vec{x}, \vec{y}) = \left(\sum_{i=1}^{d} (x_i - y_i)^p\right)^{1/p}$

2.6 Theory of LSH

- $d(x, y) \leq d_1 \Rightarrow$ prob of a match is high, $\geq p_1$
- $d(x, y) \geq d_2 \Rightarrow$ prob of a match is low, $\leq p_2$

3 PageRank

Please read chapter 5.