Assignement 1

Let's first assume we use b bands each of which containing r rows. Then, let's suppose that a certain pair of documents have a *Jaccard Similarity* with a value of s. The probability that the minhash signatures for these documents in each certain line of the signature matrix agree with each other equals s (From $Pr[h(C_1) = h(C_2)] = Sim(C_1, C_2)$). Thus, we have the following:

- The probability that the signatures in *all lines* from a certain band *agree* with each other equals s^r (thanks to the independence)
- The probability that the signatures in *at least one line* from a certain band does ***not*** *agree* with each other equals $1 - s^r$
- The probability that the signatures in *at least one line* from each band do ***not*** *agree* with each other equals $(1 - s^r)^b$
- The probability that the signatures in *at least one band and all lines of that band agree* with each other and consequently a *candidate pair* equals $1 - (1 - s^r)^b$

We therefore need to tune r and b in order to capture the pairs with the similarity we want and exclude the pairs we do not want[1].

Let's now consider the two thresholds θ_1 and θ_2, with $\theta_1 > \theta_2$. Two sets X and Y are considered "similar" (i.e. true positive) whenever $Jaccard(X, Y) \geq \theta_1$, and "not similar" (i.e. true negative) whenever $Jaccard(X, Y) < \theta_2$. From the introduction we can easily find the following:

- Given a *true positive* pair (X, Y), we want the probability of considering them as a negative pair (*false negative* probability):

  $$p_1 \geq (1 - t_1^r)^b = P(FN)$$

  with $t_1 \in [\theta_1, 1]$.

  Which means that *at least one line* from each band do ***not*** *agree*, even if in reality $Jaccard(X, Y) \geq \theta_1$.

- Given a *true negative* pair (X, Y), we want the probability of considering them as a positive pair (*false positive* probability):

  $$p_2 \geq 1 - (1 - t_2^r)^b = P(FP)$$

  with $t_2 \in [0, \theta_2)$.

  Which means that *at least one band and all lines of that band agree*, even if in reality $Jaccard(X, Y) < \theta_2$.

[1] We can always post-process to exclude false positives with low similarity (by calculating the exact similarity), but any false negatives cannot be recovered.

Assignement 2

Section 2.1

Given A square invertible, n-dimensional matrix with SVD

$$A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i u_i v_i^T$$

we have the inverse

$$A^{-1} = (U\Sigma V^T)^{-1} = (V^T)^{-1}\Sigma^{-1}U^{-1}$$

Now, since $U^{-1} = U^T$ and $V^{-1} = V^T$ (because they are orthonormal[2]), we can write

$$A^{-1} = V\Sigma^{-1}U^T = \sum_{i=1}^{n} \frac{1}{\sigma_i} v_i u_i^T = B$$

(note Σ is a diagonal matrix and the entries are all non-negative).
Indeed $A^{-1}A = (V\Sigma^{-1}U^T)(U\Sigma V^T) = V\Sigma^{-1}(U^T U)\Sigma V^T = V\Sigma^{-1}\Sigma V^T = VV^T = I$

Section 2.2

Let's first consider the following relationships:

$$A = U\Sigma V^T \iff VA = U\Sigma \quad \text{and} \quad B = V\Sigma^{-1}U^T \iff UB = V\Sigma^{-1}$$

In this case we are generalizing the concept and considering that A is square but not necessarily invertible, which means $rank(A) = r < n$ and there are $n - r$ components in the null space $N(A)$. Thus, we can write the following vector form:

$$\begin{cases} Av_i = \sigma u_i & \text{for } i = 1, ..., r \\ Bu_i = v_i \frac{1}{\sigma_i} & \text{for } i = 1, ..., r \end{cases} \qquad \begin{cases} Av_i = 0 & \text{for } i = r+1, ..., n \\ Bu_i = 0 & \text{for } i = r+1, ..., n \end{cases}$$

If any of the singular values $\sigma_i = 0$, the corresponding entry in the inverse would be $1/0$ and therefore Σ^{-1} cannot exist. We therefore need the *Moore–Penrose inverse* (pseudo-inverse) $A^+ = (A^*A)^{-1}A^* = V\Sigma^\dagger U^T = B$ that has a 0 entry in each of the $n - r$ components in the diagonal (* stands for the *Hermitian matrix*).

$$\Sigma^\dagger = \begin{bmatrix} \frac{1}{\sigma_1} & & \cdots & & 0 \\ & \ddots & & & \\ \vdots & & \frac{1}{\sigma_r} & & \vdots \\ & & & \ddots & \\ 0 & & \cdots & & 0 \end{bmatrix}$$

Thus, for every vector $x = \sum_{i=1}^{r} \alpha_i v_i$ that can be expressed as a linear combination of the right singular vectors of A, the following holds:

$$\begin{cases} BAx = I_r x = x & \text{for } i = 1, ..., r \\ BAx = B0 = 0 & \text{for } i = r+1, ..., n \end{cases}$$

With full-rank A we only have the first equation only and we are back to *Section 2.1.*

[2] $UU^T = U^T U = I$ and $VV^T = V^T V = I$