BDA – Practical Sessions

Session 2

Introduction – Shingling, Min-Hashing, LSH

Jen Alchimowicz

Agenda for today

1. Quick recap

2. Exercises from the book (From Session 3 - exercises)

3. LSH exercise (From session 3 – practice)

4. Word counting in PySpark (From Session 2 - exercises)
(Not in class, I will provide solutions)

Agenda for today

1. Quick recap

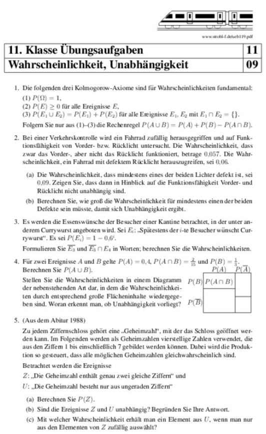
2. Exercises from the book (From Session 3 - exercises)

3. LSH exercise (From session 3 – practice)

4. Word counting in PySpark (From Session 2 - exercises)

Recap

Docu mnet $\xrightarrow{\text{Shingling}}$ Vector representation $\xrightarrow{\text{Min Hashing}}$ Signature $\xrightarrow{\text{LSH}}$ Candidate pairs



11. Klasse Übungsaufgaben 11
Wahrscheinlichkeit, Unabhängigkeit 09

$$\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix}$$

$$\begin{bmatrix} 6 \\ 1 \\ 0 \\ 2 \\ 4 \end{bmatrix}$$

{Document1, Document 5}
{Document3, Document 7}
{Document8, Document 2}

Jaccard Similarity

Jaccard Similarity of Sets:

- *Consider only unique elements*
- *E.g. SIM({1,2,3,4}, {2,3,5,7}) = 2/6*

Jaccard Similarity of Bags:

- *Include repeating items*
- *E.g. SIM({1,1,1,2}, {1,1,2,2,3}) = 3/9 = 1/3*
- *Maximum Jaccard similarity of bags is 1/2 (when the bags are equivalent)*

Shingling

- By character
  - E.g. ‘Hello! How are you?’ with k=3:
    [‘hel’, ‘ell’, ‘llo’, ‘lo!’, ‘o! ‘, ‘! h’, ‘ ho’, ‘how’, …]
- By word
  - E.g. ‘Hello! How are you?’ with k=2:
    [‘hello! how’, ‘how are’, ‘are you’, …]

- How to chose k?
- k should be picked large enough that the probability of any given shingle appearing in any given document is low
- Rule of thumb: k=5 for small documents, k=9 for large.
- You need to ensure that (n possible characters)^k >> characters in average document

Min Hashing

Element	S_1	S_2	S_3	S_4
0	0	1	0	1
1	0	1	0	0
2	1	0	0	1
3	0	0	1	0
4	0	0	1	1
5	1	0	0	0

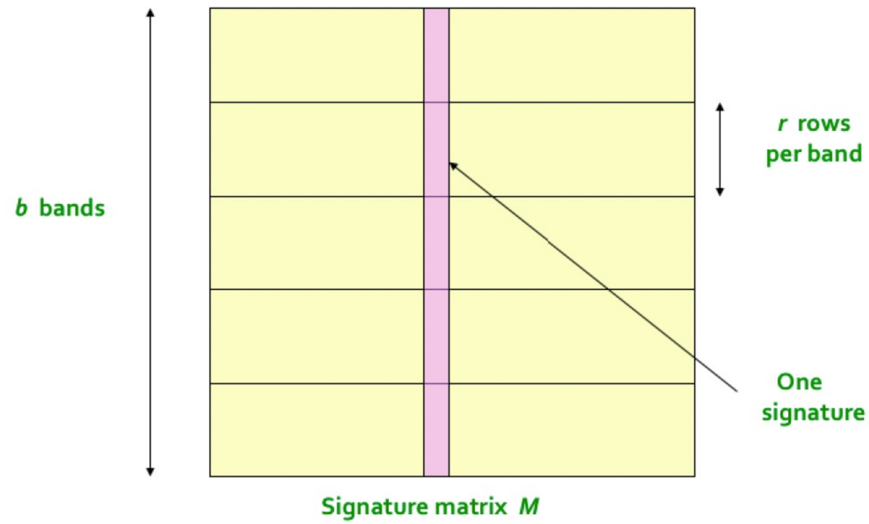
Few ways to do it:

- Clustering
  - Use random hash function to reduce size of the characteristic matrix
- Using permutations *(Min Hashing)*
  - *Permute rows and grab first '1' from top*
  - *Slow in practice, permuting rows on large datasets is computationally expensive*
- Using hash functions *(Fast Min Hashing)*
  - *Apply hash functions to imitate permutations*
  - *Faster than permutations*
  - *We will use this method in class*

Locality Sensitive Hashing

1. Apply hash function to each band
2. If two columns hash to the same bucket make them a candidate pair

Two columns will hash to the same bucket if they are identical in one of the bands, therefore:
Identical in one of the bands = candidate pair



length of signature = b*r

s = true similarity between 2 documents

b = number of bands

r = number of rows in each band

P(candidate pair) = $1 - (1 - s^r)^b$

Agenda for today

1. Quick recap

2. Exercises from the book (From Session 3 - exercises)

3. LSH exercise (From session 3 – practice)

4. Word counting in PySpark (From Session 2 - exercises)

HW3

1. (**Exercise 3.2.3 MMDS book**) What is the largest number of k-shingles a document of n bytes can have? You may assume that the size of the alphabet is large enough that the number of possible strings of length k is at least as n. (In UTF-8 encoding each letter occupies 1 byte(8 bits).)

$n = \#\text{char}$

$k = 3$

$$\overbrace{\underbrace{abcde\ \ldots}_{n-k}\ \underbrace{\overbrace{xyz}^{k}}_{1}}^{n}$$

$n - k + 1$

HW3

3. (**Exercise 3.3.3 MMDS book**) In Fig. 3.5 is a matrix with six rows.

Element	S_1	S_2	S_3	S_4
0	0	1	0	1
1	0	1	0	0
2	1	0	0	1
3	0	0	1	0
4	0	0	1	1
5	1	0	0	0

Figure 3.5: Matrix for Exercise 3.3.3

- Compute the minhash signature for each column if we use the following three hash functions: $h_1(x) = 2x + 1 \bmod 6$; $h_2(x) = 3x + 2 \bmod 6$; $h_3(x) = 5x + 2 \bmod 6$.
- Which of these hash functions are true permutations?
- How close are the estimated Jaccard similarities for the six pairs of columns to the true Jaccard similarities?

HW3

3. (**Exercise 3.3.3 MMDS book**) In Fig. 3.5 is a matrix with six rows.

- Compute the minhash signature for each column if we use the following three hash functions: $h_1(x) = 2x + 1 \bmod 6$; $h_2(x) = 3x + 2 \bmod 6$; $h_3(x) = 5x + 2 \bmod 6$.

Rows	2x+1 mod 6	3x+2 mod 6	5x + 2 mod 6
0	1	2	2
1	3	5	1
2	5	2	0
3	1	5	5
4	3	2	4
5	5	5	3

HW3

3. (**Exercise 3.3.3 MMDS book**) In Fig. 3.5 is a matrix with six rows.

- Compute the minhash signature for each column if we use the following three hash functions: $h_1(x) = 2x + 1 \bmod 6$; $h_2(x) = 3x + 2 \bmod 6$; $h_3(x) = 5x + 2 \bmod 6$.

Element	S_1	S_2	S_3	S_4
0	0	1	0	1
1	0	1	0	0
2	1	0	0	1
3	0	0	1	0
4	0	0	1	1
5	1	0	0	0

(h1)	(h2)	(h3)
2x+1 mod 6	**3x+2 mod 6**	**5x + 2 mod 6**
1	2	2
3	5	1
5	2	0
1	5	5
3	2	4
5	5	3

$\Longrightarrow$

	S_1	S_2	S_3	S_4
h1	5	1	1	1
h2	2	2	2	2
h3	0	1	4	0

min(5,5) min(2,5) min(0,3)

HW3

3. (**Exercise 3.3.3 MMDS book**) In Fig. 3.5 is a matrix with six rows.

- Which of these hash functions are true permutations?

(h1) 2x+1 mod 6	(h2) 3x+2 mod 6	(h3) 5x + 2 mod 6
1	2	2
3	5	1
5	2	0
1	5	5
3	2	4
5	5	3

h1 – not a true permutation
h2 – not a true permutation
h3 – true permutation

HW3

3. (**Exercise 3.3.3 MMDS book**) In Fig. 3.5 is a matrix with six rows.

- How close are the estimated Jaccard similarities for the six pairs of columns to the true Jaccard similarities?

Sim(S_1, S_2) = 0/4 = 0, estimated 1/3
Sim(S_1, S_3) = 0/4 = 0, estimated 1/3
Sim(S_1, S_4) = 1/4, estimated 2/3
Sim(S_2, S_3) = 0/4 = 0, estimated 2/3
Sim(S_2, S_4) = 1/4, estimated 2/3
Sim(S_3, S_4) = 1/4, estimated 2/3

Characteristic matrix

Element	S_1	S_2	S_3	S_4
0	0	1	0	1
1	0	1	0	0
2	1	0	0	1
3	0	0	1	0
4	0	0	1	1
5	1	0	0	0

Signature

	S_1	S_2	S_3	S_4
h1	5	1	1	1
h2	2	2	2	2
h3	0	1	4	0

Notes:
For similarity between the shingle representations (from characteristic matrix) we use the Jaccard Similarity of sets. E.g:
Sim(S_1, S_4) = SIM({2,5}, {0,2,4})

For estimated similarity (from signatures) we use the formula:
Sim(S_1, S_4) = $\frac{1}{K}\sum_{k=1}^{K} I(h_k(S_1) = h_k(S_2))$
where K=length of signature and I is identity function that is 1 if $h_k(S_1) = h_k(S_2)$ else 0. We baiscally sum up the rows where singatures have equal elements and divide by length of signature.

Agenda for today

1. Quick recap

2. Exercises from the book (From Session 3 - exercises)

3. LSH exercise (From session 3 – practice)

4. Word counting in PySpark (From Session 2 - exercises)