

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



11. Klasse Übungsaufgaben Wahrscheinlichkeit, Unabhängigkeit

- Die folgenden drei Kolmogorov-Axiome sind für Wahrscheinlichkeiten fundamental:
 - $P(\emptyset) = 0$.
 - $P(E) \geq 0$ für alle Ereignisse E .
 - $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ für alle Ereignisse E_1, E_2 mit $E_1 \cap E_2 = \emptyset$.
 Folgen Sie aus (1)-(3) die Rechenregel $P(A \cap B) = P(A) + P(B) - P(A \cup B)$.
- Bei einer Verkehrskontrolle wird ein Fahrzeug zufällig herausgegriffen und auf Punktsicherheitskennzeichen von Vorder- bzw. Rücklicht untersucht. Die Wahrscheinlichkeit, dass zwar das Vorder-, aber nicht das Rücklicht funktioniert, betrage $0,07$. Die Wahrscheinlichkeit, ein Fahrzeug mit defektem Rücklicht herauszugreifen, sei $0,05$.
 - Die Wahrscheinlichkeit, dass mindestens eines der beiden Lichter defekt ist, sei $0,08$. Zeigen Sie, dass damit in Hinblick auf die Funktionsfähigkeit Vorder- und Rücklicht keine Unabhängigkeit vorliegt.
 - Berechnen Sie, wie groß die Wahrscheinlichkeit für mindestens eines der beiden Defekte sein müsste, damit sich Unabhängigkeit ergibt.
- Es werden die Erwartungswerte der Besucher einer Kantine berechnet, in der unter anderem Currywurst angeboten wird. Sei E_C „Spätkommender isst Currywurst“ und E_{G} „Es ist Currywurst“. Es sei $P(E_C) = 1 - 0,6$.

Formulieren Sie P_C und $P_{G|C}$ in Worten. Berechnen Sie die Wahrscheinlichkeit:

 - Für zwei Ereignisse A und B gelte $P(A) = 0,4$, $P(A \cap B) = \frac{1}{5}$ und $P(B) = \frac{1}{2}$. Berechnen Sie $P(A \cup B)$.
 - Stellen Sie die Wahrscheinlichkeiten in einem Diagramm der schonendsten Art dar, in dem die Wahrscheinlichkeiten durch entsprechend große Flächenanteile wiedergegeben sind. Woran erkennt man, ob Unabhängigkeit vorliegt? $P(B)$
- (Aus dem Abitur 1988)

„Zu jedem Ziffernschloss gehört eine „Geheimzahl“, mit der der Schloss geöffnet werden kann. Im Folgenden werden als Geheimzahlen vierstellige Zahlen verwendet, die aus den Ziffern 1 bis einschließlich 7 gebildet werden können. Dabei wird die Produktion so gesteuert, dass alle möglichen Geheimzahlen gleichwahrscheinlich sind. Betrachten wir die Ereignisse

 - „Die Geheimzahl enthält genau zwei gleiche Ziffern“ und
 - „Die Geheimzahl besteht nur aus ungeraden Ziffern“
 - Berechnen Sie $P(Z)$.
 - Sind die Ereignisse Z und U unabhängig? Begründen Sie Ihre Antwort.
 - Mit welcher Wahrscheinlichkeit erhält man ein Element aus U , wenn man nur aus den Elementen von Z zufällig auswählt?

0
1
0
0
0
0
1
0
0
0
0
0
0
1
0
:

6
1
0
2
4

{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 $(n \text{ possible characters})^k \gg \text{characters in average document}$

Min Hashing

<i>Element</i>	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:

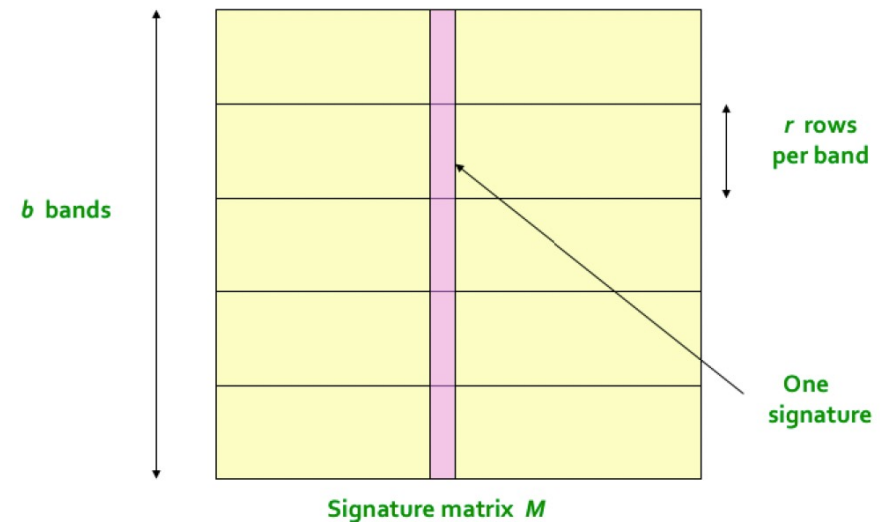
Identical in one of the bands = candidate pair

s = true similarity between 2 documents

b = number of bands

r = number of rows in each band

$$P(\text{candidate pair}) = 1 - (1 - s^r)^b$$



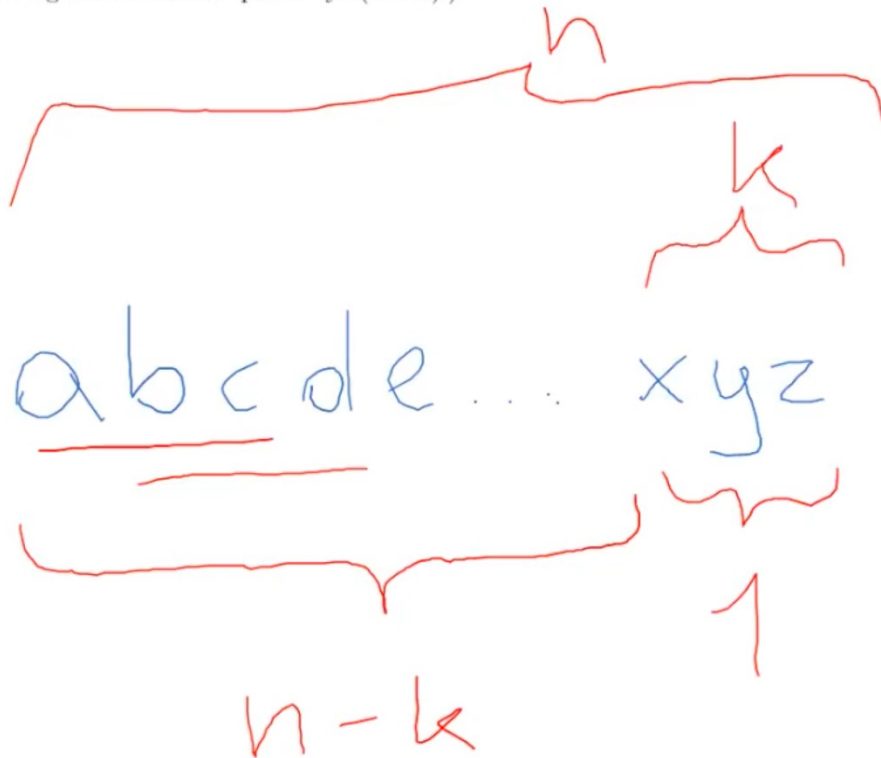
length of signature = $b * r$

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).)



$$n = \# \text{ char}$$

$$k = 3$$

$$\underline{n - k + 1}$$

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

<i>Element</i>	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	<u>$2x+1 \bmod 6$</u>	$3x+2 \bmod 6$	$5x + 2 \bmod 6$
<u>0</u>	1	2	2
<u>1</u>	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: $h_1(x) = 2x + 1 \bmod 6$; $h_2(x) = 3x + 2 \bmod 6$; $h_3(x) = 5x + 2 \bmod 6$.

<i>Element</i>	S_1	S_2	S_3	S_4	(h1) $2x+1 \bmod 6$	(h2) $3x+2 \bmod 6$	(h3) $5x+2 \bmod 6$
0	0	<u>1</u>	0	1	1	2	2
1	0	<u>1</u>	0	0	3	5	1
2	<u>1</u>	0	0	1	<u>5</u>	<u>2</u>	<u>0</u>
3	0	0	1	0	1	5	5
4	0	0	1	1	3	2	4
5	<u>1</u>	0	0	0	<u>5</u>	<u>5</u>	<u>3</u>

	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)	(h2)	(h3)
$2x+1 \bmod 6$	$3x+2 \bmod 6$	$5x+2 \bmod 6$
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?

$\text{Sim}(S_1, S_2) = 0/4 = 0$, estimated $1/3$
 $\text{Sim}(S_1, S_3) = 0/4 = 0$, estimated $1/3$
 $\text{Sim}(S_1, S_4) = 1/4$, estimated $2/3$
 $\text{Sim}(S_2, S_3) = 0/4 = 0$, estimated $2/3$
 $\text{Sim}(S_2, S_4) = 1/4$, estimated $2/3$
 $\text{Sim}(S_3, S_4) = 1/4$, estimated $2/3$

Characteristic matrix

<i>Element</i>	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:

$\text{Sim}(S_1, S_4) = \text{SIM}(\{2,5\}, \{0,2,4\})$

For estimated similarity (from signatures) we use the formula:

$\text{Sim}(S_1, S_4) = \frac{1}{K} \sum_{k=1}^K I(h_k(S_1) = h_k(S_2))$

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 basically sum up the rows where signatures 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)