[BRY17] https://medium.com/engineering-brainly/locality-sensitive-hashing-explained-304eb39291e4 , Hubert Brylkowski, 2017 (Letzter Abruf:11.06.2018)

1. **Locality sensitive hashing**

**Jaccard Index:** Für die Jaccard Index Bildung ist die Definition, ob eine Frage das Duplikat von einer anderen Frage ist, wichtig. [BRY17] behauptet, dass die Jaccard Indexierung geeignet ist für diesen Fall. Der Jaccard Index entspricht einen Schnittpunkt über eine Union. Es wird erstmal die Berechnung der Anzahl an gemeinsamer Elemente aus zwei Mengen durchgeführt und danach wird die Teilung dieser berechneten Anzahl an gemeinsamer Elemente aus zwei Mengen durch die Anzahl der zur ersten Menge oder zu zweiten Mengen, oder noch zu beiden Menge gehörende Elemente. Ein Beispiel dafür wäre: „wer hat mich gestern angerufen“ & „wer hat mich heute angerufen“

wer   
hat  
mich   
angerufen

heute

angerufen

Abbildung 1: Beispiel von Wörtermengen

Der Schnitt dieser beiden Mengen ist 4 und die Vereinigung ist 4 + 1 + 1 = 6. Das Jaccard Index wird durch die Operation 4/6 den Wert 0,67 haben. [BRY17] behauptet, dass je mehr die Wörter üblich, desto größer ist das Jaccard Index und die Wahrscheinlich, dass zwei Fragen Duplikate sind.

* 1. Min part of MinHash (Minimale Index)

Die Erstellung eines einfachen numerischen Fingerabdruckes mit fester Größe für jeden Satz stellt sich als Lösung zum Problem der Aufteilung von jeder Frage in die Wörter anschließend der Vergleich von jeder Frage. Die folgenden Sätze werden benutzt:

„wer hat mich gestern angerufen“

„wer hat mich heute angerufen“

„wer hat dich gleich abgeholt“

Diese haben folgende Jaccard Indexe:

J („wer hat mich gestern angerufen“, „wer hat mich heute angerufen“) = 0,67

J („wer hat mich heute angerufen“, „wer hat dich gleich abgeholt“) = 0,4

J („wer hat mich gestern angerufen“, „wer hat dich gleich abgeholt“) = 0,4

Für die Berechnung des MinHashs müssen ein Wörterbuch (Wörtermenge) aus allen in diesem Fall in Frage kommende Frage und eine zufällige Permutation, erstellt werden.  
Dafür kommt die folgende Menge in Frage („wer“, „hat“, „mich“, „gestern“, „angerufen“, „heute“, „dich“, „gleich“, „abgeholt“)

Es wird eine Iteration geben auf die Zeile hinaus und Das Index jeder Zelle wird geschrieben, falls das gesuchte Wort im Satz zu finden ist [BRY17]. Die Tabelle 1 zeigt ein Beispiel davon.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | Word | wer hat mich gestern angerufen | wer hat mich heute angerufen | wer hat dich gleich abgeholt |
| 1 | Wer | 1 | 1 | 1 |
| 2 | Hat | 2 | 2 | 2 |
| 3 | Mich | 3 | 3 |  |
| 4 | Gestern | 4 |  |  |
| 5 | Angerufen | 5 | 5 |  |
| 6 | Heute |  | 6 |  |
| 7 | Dich |  |  | 7 |
| 8 | Gleich |  |  | 8 |
| 9 | abgeholt |  |  | 9 |

Tabelle 1: Tabelle von indexierte Wörter mit Verweis auf deren vorkommen in Sätze

Nur das Vorkommen des ersten Wortes ist wichtig, bzw. der minimale Index wird geliefert. Hiermit ist ein Mindestwert für alle Fragen und ein erstes Teil des Fingerabdrucks erhalten. Für den Erhalt des zweiten Teils des Fingerabdrucks muss eine weitere zufällige Permutation erfolgen. Abhängig von wie groß der Fingerabdruck sein soll können weiteren Permutationen stattfinden. Z. B es können noch 5 folgende Permutationen erstellt werden:

(„dich“, „gestern“, „wer“, „abgeholt“, „mich“, „gleich“, „hat“, „angerufen“, „heute“)

(„angerufen“, „dich“, „hat“, „heute“, „wer“, „gleich“, „mich“, „abgeholt“, „gestern“)

(„mich“, „gestern“, „gleich“, „abgeholt“, „dich“, „wer“, „angerufen“, „heute“, „hat“)

(„heute“, „abgeholt“, „gestern“, „gleich“, „wer“, „mich“, „dich“, „hat“, „angerufen“)

(„abgeholt“, „mich“, „wer“, „gestern“, „dich“, „hat“, „heute“, „gleich“, „angerufen“)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | Word | wer hat mich gestern angerufen | wer hat mich heute angerufen | wer hat dich gleich abgeholt |
| 1 | Dich |  |  | 1 |
| 2 | Gestern | 2 |  |  |
| 3 | Wer | 3 | 3 | 3 |
| 4 | abgeholt |  |  |  |
| 5 | Mich |  |  |  |
| 6 | Gleich |  |  |  |
| 7 | Hat |  |  |  |
| 8 | Angerufen |  |  |  |
| 9 | heute |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | Word | wer hat mich gestern angerufen | wer hat mich heute angerufen | wer hat dich gleich abgeholt |
| 1 | Angerufen | 1 | 1 |  |
| 2 | Dich |  |  | 2 |
| 3 | hat |  |  |  |
| 4 | Heute |  |  |  |
| 5 | wer |  |  |  |
| 6 | Gleich |  |  |  |
| 7 | mich |  |  |  |
| 8 | Abgeholt |  |  |  |
| 9 | Gestern |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | Word | wer hat mich gestern angerufen | wer hat mich heute angerufen | wer hat dich gleich abgeholt |
| 1 | Mich | 1 | 1 |  |
| 2 | Gestern | 2 |  |  |
| 3 | Gleich |  |  | 3 |
| 4 | abgeholt |  |  |  |
| 5 | Dich |  |  |  |
| 6 | wer |  |  |  |
| 7 | Angerufen |  |  |  |
| 8 | heute |  |  |  |
| 9 | hat |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | Word | wer hat mich gestern angerufen | wer hat mich heute angerufen | wer hat dich gleich abgeholt |
| 1 | Heute |  | 1 |  |
| 2 | abgeholt |  |  | 2 |
| 3 | Gestern | 3 |  |  |
| 4 | Gleich |  |  |  |
| 5 | wer |  |  |  |
| 6 | Mich |  |  |  |
| 7 | Dich |  |  |  |
| 8 | Hat |  |  |  |
| 9 | Angerufen |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | Word | wer hat mich gestern angerufen | wer hat mich heute angerufen | wer hat dich gleich abgeholt |
| 1 | abgeholt |  |  | 1 |
| 2 | Mich | 2 | 2 |  |
| 3 | wer |  |  |  |
| 4 | Gestern |  |  |  |
| 5 | dich |  |  |  |
| 6 | hat |  |  |  |
| 7 | Heute |  |  |  |
| 8 | Gleich |  |  |  |
| 9 | Angerufen |  |  |  |

MinHash („wer hat mich gestern angerufen“) = [1,2,1, 1,3, 2]

MinHash („wer hat mich heute angerufen“) = [1,3,1,1,1,2]

MinHash („wer hat dich gleich abgeholt“) = [1,1,2,3,2,1]

MinHashÄhnlichkeit („wer hat mich gestern angerufen“, „wer hat mich heute angerufen“) = 3/5

MinHashÄhnlichkeit („wer hat mich gestern angerufen“, „wer hat dich gleich abgeholt“) = 2/5

MinHashÄhnlichkeit („wer hat mich heute angerufen“, „wer hat dich gleich abgeholt“) =2/5

MinHash („Who was the first King of Poland“) =[2, 1,1, 2,1,1]

MinHash („Who was the first ruler of Poland“) =

MinHash („Who was the last pharaoh of Egypt“)