Project 1.

Sentence 1: The Big Data platform for students is Blackboard

Sentence 2: Questions on MinHash project by NTNU is on Piazza

Sentence 3: NTNU Big Data platform are Blackboard and Piazza

Sentence 4: The project data for students are on Blackboard not Piazza

* 1. Extracting K-shingles

1: [‘big’, ‘data’, ‘platform’, ‘students’, ‘blackboard’]

2: [‘questions’, ‘minhash’, ‘project’, ‘ntnu’, ‘students’, ‘piazza’]

3: [‘ntnu’, ‘big’, ‘data’, ‘platform’, ‘blackboard’, ‘piazza’]

4: [‘project, ‘data’, ‘students’, ‘blackboard’, ‘piazza’]

* 1. Creating Unique Words Dictionary

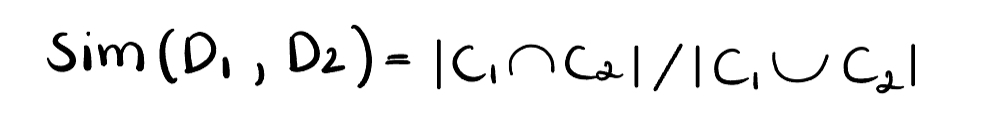
Unique Words Dictionary: {‘big’, ‘data’, ‘platform’, ‘students’, ‘blackboard’, ‘questions’, ‘minhash’, ‘project’, ‘ntnu’, ‘piazza’}

1.3 Input Matrix of sentences

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | big | data | platform | students | blackboard | questions | minhash | project | ntnu | piazza |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| 3 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| 4 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |

Jaccard Similarity of every possible sentence combination:

Formula:



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 |
| 1 | 1 | 1/5 | 4/6 | 3/6 |
| 2 | 1/5 | 1 | 4/8 | 3/7 |
| 3 | 4/6 | 4/8 | 1 | 4/8 |
| 4 | 3/6 | 3/7 | 4/8 | 1 |

Sentence pair 1 and 3 has the highest Jaccard similarity and are therefore the most similar.

* 1. Calculating Min-Hash Signature Matrix

|  |  |  |
| --- | --- | --- |
| **H1** | **H2** | **H3** |
| 5 | 9 | 10 |
| 6 | 6 | 7 |
| 7 | 3 | 4 |
| 8 | 10 | 1 |
| 9 | 7 | 8 |
| 10 | 4 | 5 |
| 1 | 1 | 2 |
| 2 | 8 | 9 |
| 3 | 5 | 6 |
| 4 | 2 | 3 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** |
| big | 1 | 0 | 1 | 0 |
| data | 1 | 0 | 1 | 1 |
| platform | 1 | 0 | 1 | 0 |
| students | 1 | 0 | 0 | 1 |
| blackboard | 1 | 0 | 1 | 1 |
| questions | 0 | 1 | 0 | 0 |
| minhash | 0 | 1 | 0 | 0 |
| project | 0 | 1 | 0 | 1 |
| ntnu | 0 | 1 | 1 | 0 |
| piazza | 0 | 1 | 1 | 1 |

Signature matrix:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| H1 | 5 | 1 | 3 | 2 |
| H2 | 3 | 1 | 2 | 2 |
| H3 | 1 | 2 | 3 | 1 |