**Contents**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **List of Figures**…………………………………………………………………………….. | | | | | | | Viii | |
| **List of Tables**…………………………………………………………………………… | | | | | | | Ix | |
| **Nomenclature**…………………………………………………………………………....... | | | | | | | X | |
| **1** | | | **Introduction**……………………………………………………………………... | | | | 1 | |
|  | | | 1.1 | | Problem Definition……………………………………………………….. | | 1 | |
|  | | | 1.2 | | Motivation ……………………………………………………………….. | | 4 | |
|  | | | 1.3 | | Scope of Project …………………………………………………………. | | 6 | |
|  | | | 1.4 | | Functional and Non-Functional Requirements ………………………….. | | 9 | |
|  | | | 1.5 | | Organization of the Report ……………………………………………….. | | 10 | |
| **2** | | | **Background Work**………………………………………………………………. | | | | 12 | |
| **3** | | **Implementation**  3.1 Technologies used …………………………………………………………  3.2 Algorithm / Methodology ………………………………………………  3.3 Implementation …………………………………………………………… | | | | |  | |
| **4** | | **Results and Discussion** | | | | |  | |
| **5** | | | **Conclusions and Learnings**……………………………………… | | | | | 55 | |
|  | | | 5.1 | | Conclusions………………………………………………………………. | | | 55 | |
|  | | | 5.2 | | Scope for Future Work……………………………………………………. | | | 56 | |
|  | | | 5.3 | | Learnings | | | 57 | |
| **References** ……………………………………………………………………………... | | | | | | | | 58 | |
| **Acknowledgements**………………………………………………………………………. | | | | | | | |  | |

**List of Figures**

|  |  |  |
| --- | --- | --- |
| 3.1 | Vector reprereeeeeeeeeedocument………………………………………………... | 15 |
| 3.2 | Vectoreeeeeeeeeeeeeeenal space for SVM………………………………………… | 16 |
| 3.3 | dddddddddddddddddvectors in two dimensional space…………………………... | 16 |
| 3.4 | ffffffffffffffffffffffffution in two dimensional space………………………………. | 17 |
| 3.5 | Plufffffffffffffffffffer- plane………………………………………………………… | 17 |
| 3.6 | Mgggggggggggg…………………………………………………………………… | 18 |
| 3.7 | Non- lggggggggggarable data points……………………………………………….. | 19 |
| 3.8 | Vectors in two dimensional space for KNN ……………………………………… | 25 |
| 3.9 | Label computation for K=1 and K=3……………………………………………… | 25 |
| 4.1 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 42 |
| 4.2 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 43 |
| 4.3 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 43 |
| 4.4 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 44 |
| 4.5 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 45 |
| 4.6 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 45 |
| 4.7 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 47 |
| 4.8 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 47 |
| 4.9 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 48 |
| 4.10 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 48 |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| 3.1 | Document Word Matrix……………………………………………………………... | 14 |
| 3.2 | Parameters for F1 computation……………………………………………………… | 27 |
| 4.1 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 37 |
| 4.2 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 38 |
| 4.3 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 38 |
| 4.4 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 39 |
| 4.5 | K evaluation for 20 newsgroup group 1…………………………………………….. | 40 |
| 4.6 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 40 |
| 4.7 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 41 |
| 4.8 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 41 |
| 4.9 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 51 |
| 4.10 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 51 |
| 4.11 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 52 |
| 4.12 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 52 |
| 4.13 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 53 |
| 4.14 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 53 |
| 4.16 | Comparisohhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhroup groups for SVM .. | 54 |

**Nomenclature**

|  |  |
| --- | --- |
| B | Bias for hyper-plane in SVM |
| C | Regularization parameter for SVM |
|  | Category to which documents belong (C1 or C0) |
| Di | ith Document |
|  | Decision function |
| FP | False positive |
| FN | False negative |
| F1 | Evaluation parameter combining precision and recall |
| IDFt | Inverse Document Frequency for the term t |
| K | Number of nearest neighbours(training documents) with which distance computation is to be performed |
| L | Words present in the document |
|  | Multinomial model feature vector for the ith document in training |
|  | Frequency of word Wt  in document Di |
|  | Multinomial model feature vector for the jth document in testing |
|  | Total number of words in Di |
|  | Number of documents |
|  | Number of documents of class Ck |
| NFt | Number of documents containing term t |
| TP | True positive |
| TN | True Negative |
| TFi | Frequency of ith word in a document |
|  | Dictionary |
| W | Word |
|  | Weight vector for defining hyper-plane is SVM |
| X | Input feature vector |
| Y | Output class of a classifier |
|  | Parameter to indicate Di has class Ck |
|  | Lagrange multipliers |
|  | Lagrange multipliers used for soft margin SVM |
|  | Distance parameter |
|  | Slack variable of ith training document violating margin |
|  |  |
|  |  |