The data set is made up of paper downloaded from the arXiv website. It is collected by using arXiv sanity preserver program. (<https://github.com/karpathy/arxiv-sanity-preserver/>) Note that all downloaded papers are in pdf format. We used the pdf to txt program provided by the arXiv sanity preserver program to convert all papers into txt format. The dataset includes 11 different classes, the table below illustrates the details of the data set.

|  |  |  |
| --- | --- | --- |
| Class name | Number of documents | Average words |
| cs.AI ( Artificial Intelligence)  cs.CE (Computational Engineering)  cs.CV (Computer Vision)  cs.DS (Data Structures)  cs.IT (Information Theory )  cs.NE (Neural and Evolutionary)  cs.PL (Programming Languages)  cs.SY (Systems and Control)  math.AC (Commutative Algebra )  math.GR (Group Theory)  math.ST (Statistics Theory) | 2995  2505  2525  4136  3233  3012  2901  3106  2885  3065  6025 | 6212  5777  5630  7439  5938  5856  7012  5948  5984  6642  6983 |

Data statistics: average number of words for each document in this data set is 6311, maximum and minimum words are 10000 and 1000