1 . no stop words term frequency ratio used

2. removed stopwords using stopwords\_1.txt, term frequency ratio used

3. removed stopwords using stopwords\_2.txt, term frequency ratio used

4. removed stopwords using stopwords\_3.txt, term frequency ratio used

5. removed stopwords using stopwords\_3.txt + randomly select 1000 features, term frequency ratio used

6 . TF-IDF ratio used

7. removed stopwords using stopwords\_1.txt, TF-IDF ratio used

8. removed stopwords using stopwords\_2.txt, TF-IDF ratio used

9. removed stopwords using stopwords\_3.txt, TF-IDF ratio used

10. removed stopwords using stopwords\_3.txt + randomly select 1000 features, TF-IDF ratio used

For the first 5 datasets each number n=Ti/D where

Ti is the number of times a given word T appears in a given article

D is the total number of words In that articles

|  |  |  |
| --- | --- | --- |
| File (dataset) name | Number of feature | Preprocessing used |
| 1 | 7612 | no stop words term frequency ratio used |
| 2 | 7248 | removed stopwords using stopwords\_1.txt, term frequency ratio used |
| 3 | 6704 | removed stopwords using stopwords\_2.txt, term frequency ratio used |
| 4 | 5546 | removed stopwords using stopwords\_3.txt, term frequency ratio used |
| 5 | 1000 | removed stopwords using stopwords\_3.txt + randomly select 1000 features, term frequency ratio used |
| 6 | 7612 | TF-IDF ratio used |
| 7 | 7248 | removed stopwords using stopwords\_1.txt, TF-IDF ratio used |
| 8 | 6704 | removed stopwords using stopwords\_2.txt, TF-IDF ratio used |
| 9 | 5546 | removed stopwords using stopwords\_3.txt, TF-IDF ratio used |
| 10 | 1000 | removed stopwords using stopwords\_3.txt + randomly select 1000 features, TF-IDF ratio used |

For the first last 5 datasets each number n=Ti/D\* log(N\_docs/Ni) (this is tf-idf) where

Ti is the number of times a given word T appears in a given article

D is the total number of words In that articles

N\_docs is the number of articles in our case 40

Ni the number of articles word i appears in

Each row is a new article, each column is a different word