Eclipse JDT/Core

• Plot of the distribution of all the terms in the software lexicon. You can use Excel or anything like it.

• List of the 10 most frequent terms in the corpus (report their tf).

• List of the 10 terms with the highest document frequency (report their df).

• List of the 3 documents with the highest number of unique terms.

• List of the 3 documents with the lowest number of unique terms.

Weka

• Plot of the distribution of all the terms in the software lexicon. You can use Excel or anything like it.

Figure 1: Term frequency (tf) distribution

Figure 2: Document frequency (df) distribution

• List of the 10 most frequent terms in the corpus (report their tf).

|  |  |  |
| --- | --- | --- |
| **Rank** | **Term** | **tf** |
| 1 | String | 20975,00 |
| 2 | valu | 13479,00 |
| 3 | param | 12954,00 |
| 4 | set | 12123,00 |
| 5 | Instanc | 11925,00 |
| 6 | Overrid | 11127,00 |
| 7 | null | 8983,00 |
| 8 | instanc | 8388,00 |
| 9 | option | 7733,00 |
| 10 | pre | 7568,00 |

• List of the 10 terms with the highest document frequency (report their df).

|  |  |  |
| --- | --- | --- |
| **Rank** | **Term** | **df** |
| 1 | Gener | 1456,00 |
| 2 | version | 1240,00 |
| 3 | Copyright | 1237,00 |
| 4 | softwar | 1237,00 |
| 5 | impli | 1237,00 |
| 6 | distribut | 1233,00 |
| 7 | program | 1233,00 |
| 8 | copi | 1232,00 |
| 9 | term | 1231,00 |
| 10 | warranti | 1231,00 |

• List of the 3 documents with the highest number of unique terms.

|  |  |  |
| --- | --- | --- |
| **Rank** | **Document** | **Unique Terms** |
| 1 | KnowledgeFlowApp.java | 5080 |
| 2 | GUI.java | 2428 |
| 3 | HierarchicalBCEngine.java | 2272 |

• List of the 3 documents with the lowest number of unique terms.

|  |  |  |
| --- | --- | --- |
| **Rank** | **Document** | **Unique Terms** |
| 1547 | RandomizableDensityBasedClusterer.java | 1 |
| 1546 | LearningNode.java | 1 |
| 1545 | SupervisedFilter.java | 1 |