

# **Assignment-1**

## **Team-5**

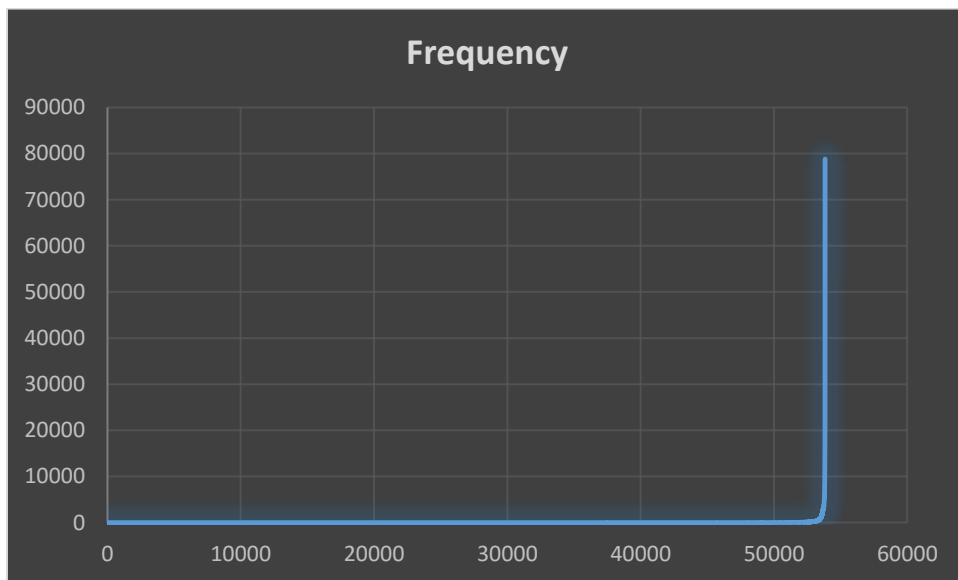
**Sanket Prabhu : srp140430**

**Chirag Chudasama : cbc140130**

### **For Oscar,**

- List of the 10 most frequent terms in the corpus (report their *tf*).
  1. String(66219)
  2. Set(40462)
  3. Date(21398)
  4. Value(19343)
  5. Name(18730)
  6. Program(15558)
  7. License(14869)
  8. Software(14697)
  9. List(14571)
  10. Null(13769)
- List of the 10 terms with the highest document frequency (report their *df*).
  1. General(3044)
  2. Version(3041)
  3. Copyright(3041)
  4. Software(3038)
  5. License(3037)
  6. Inc(3037)
  7. Terms(3036)
  8. Details(3036)
  9. Gnu(3036)
  10. Write(2923)

- List of the 3 documents with the highest number of unique terms.
  1. OceanDataProcessor.java (1601)
  2. ImportDemographicDataAction4.java(1886)
  3. OntarioMDSpec4DataTest.java(2860)
- List of the 3 documents with the lowest number of unique terms.
  1. S21.java(28)
  2. S22.java(26)
  3. ServiceCodeValidator.java(11)
- Plot of the distribution of all the terms in the software lexicon.



### **For OpenCMS,**

- List of the 10 most frequent terms in the corpus (report their *tf*).
  1. CMS(156349)
  2. String(72341)
  3. Resource(52050)
  4. Param(34811)
  5. Name(33170)
  6. Set(31894)
  7. List(31545)
  8. Value(21263)
  9. Open(21076)
  - 10.Type(19038)
  
- List of the 10 terms with the highest document frequency (report their *df*).
  1. Org(3453)
  2. Copy(3449)
  3. Version(3449)
  4. http(3449)
  5. www(3448)
  6. license(3447)
  7. Distributed(3447)
  8. Copyright(3447)
  9. Software(3447)
  - 10.Cms(3447)
  
- List of the 3 documents with the highest number of unique terms.
  1. CmsDriverManager.java(2134)
  2. CmsDefaultXmlContentHandler.java(1360)
  3. CmsSecurityManager.java(1330)

- List of the 3 documents with the lowest number of unique terms.
  1. CmsRpsServiceGenerator.java(82)
  2. SynchronisedRpcRequest.java(65) (This file appeared twice)
  3. CmsRelationValidatorInfoEntry.java(34)
- Plot of the distribution of all the terms in the software lexicon.

