### **a) Stop Words**

#### **Document 1:**

researchers focus computational phenotyping produce disease prediction models

machine learning statistical tools.

#### **Document 2:**

researchers develop tools use Bayesian statistical information generate causal models

large complex phenotyping datasets.

#### **Document3:**

researchers build computational information engine uses machine learning combine

gene function gene interaction information disparate genomic data sources.

### **N-gram Approach:**

##### **Document 1 (n=3):**

researchers focus computational

phenotyping produce disease

prediction models machine

learning statistical tools.

##### **Document 2 (n=3):**

researchers develop tools

use Bayesian statistical

information generates causal

models from large

complex phenotyping datasets.

##### **Document 3 (n=3):**

researchers build computational

information engine uses

machine learning combine

gene function gene

interaction information disparate

disparate genomic data sources.

**Parsing and dictionary creation:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **WORDS** | **DOC 1** | **DOC 2** | **DOC 3** | **Total Count** |
| **DOC1 WORDS** |  |  |  |  |
| Researches | 1 | 1 | 1 | 3 |
| Focus | 1 | 0 | 0 | 1 |
| Computational | 1 | 0 | 1 | 2 |
| Phenotyping | 1 | 1 | 0 | 2 |
| Produce | 1 | 0 | 0 | 1 |
| Disease | 1 | 0 | 0 | 1 |
| Prediction | 1 | 0 | 0 | 1 |
| Models | 1 | 1 | 1 | 3 |
| Machine | 1 | 0 | 1 | 2 |
| Learning | 1 | 0 | 1 | 2 |
| Statistical | 1 | 1 | 0 | 2 |
| Tools | 1 | 1 | 0 | 2 |
| Develop | 0 | 1 | 0 | 1 |
| Use | 0 | 1 | 0 | 1 |
| Bayesian | 0 | 1 | 0 | 1 |
| Information | 0 | 1 | 2 | 3 |
| **DOC2 WORDS** |  |  |  |  |
| Researchers | 1 | 1 | 1 | 3 |
| Tools | 1 | 1 | 0 | 2 |
| Statistical | 1 | 1 | 0 | 2 |
| Generate | 0 | 1 | 0 | 1 |
| Causal | 0 | 1 | 0 | 1 |
| Large | 0 | 1 | 0 | 1 |
| Complex | 0 | 1 | 0 | 1 |
| Datasets | 0 | 1 | 0 | 1 |
| Phenotyping | 1 | 1 | 0 | 2 |
| Models | 1 | 1 | 1 | 3 |
| **DOC3 WORDS** |  |  |  |  |
| Researchers | 1 | 1 | 1 | 3 |
| Build | 0 | 0 | 1 | 1 |
| Computational | 1 | 0 | 1 | 2 |
| Information | 0 | 1 | 2 | 3 |
| Machine | 1 | 0 | 1 | 2 |
| Learning | 1 | 0 | 1 | 2 |
| engine | 0 | 0 | 1 | 1 |
| Uses | 0 | 0 | 1 | 1 |
| Combine | 0 | 0 | 1 | 1 |
| Gene | 0 | 0 | 2 | 2 |
| Function | 0 | 0 | 1 | 1 |
| Interaction | 0 | 0 | 1 | 1 |
| Disparate | 0 | 0 | 1 | 1 |
| Genomic | 0 | 0 | 1 | 1 |
| Data | 0 | 0 | 1 | 1 |
| Sources | 0 | 0 | 1 | 1 |

b) **DCOUMENT VECTORS:**

**TERM WEIGHT calculations:**

For researchers,

TF = 1/12 = 0.083

IDF = log(3/3) = 0

Weight = TF \* IDF

= (1/12)\*0

=0

Similarly, for all words, calculations are done for all terms and recorded in the table below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **WORDS** | **DOC 1** | **DOC 2** | **DOC 3** | **TF=(no. of appearances/**  **Total no. of**  **Words in doc)** | **IDF=**  **Log(no. of docs/**  **Total no. of docs in which the term appears)** | **WEIGHT=**  **(TF\*IDF)** |
| **DOC1 WORDS** |  |  |  |  |  |  |
| Researches | 1 | 1 | 1 | 1/12=0.083 | 0 | 0 |
| Focus | 1 | 0 | 0 | 0.083 | 0.477 | 0.039 |
| Computational | 1 | 0 | 1 | 0.083 | 0.176 | 0.0146 |
| Phenotyping | 1 | 1 | 0 | 0.083 | 0.176 | 0.0146 |
| Produce | 1 | 0 | 0 | 0.083 | 0.477 | 0.039 |
| Disease | 1 | 0 | 0 | 0.083 | 0.477 | 0.039 |
| Prediction | 1 | 0 | 0 | 0.083 | 0.477 | 0.0.39 |
| Models | 1 | 1 | 1 | 0.083 | 0.176 | 0.0146 |
| Machine | 1 | 0 | 1 | 0.083 | 0.176 | 0.0146 |
| Learning | 1 | 0 | 1 | 0.083 | 0.176 | 0.0146 |
| Statistical | 1 | 1 | 0 | 0.083 | 0.176 | 0.0146 |
| Tools | 1 | 1 | 0 | 0.083 | 0.176 | 0.0146 |
| Develop | 0 | 1 | 0 | 0.0769 | 0.477 | 0.036 |
| Use | 0 | 1 | 0 | 0.0769 | 0.477 | 0.036 |
| Bayesian | 0 | 1 | 0 | 0.0769 | 0.477 | 0.036 |
| Information | 0 | 1 | 2 | 0.0769 | 0.176 | 0.0135 |
| **DOC2 WORDS** |  |  |  |  |  |  |
| Researchers | 1 | 1 | 1 | 0.0769 | 0 | 0 |
| Tools | 1 | 1 | 0 | 0.0769 | 0.176 | 0.0135 |
| Statistical | 1 | 1 | 0 | 0.0769 | 0.176 | 0.0135 |
| Generate | 0 | 1 | 0 | 0.0769 | 0.477 | 0.036 |
| Causal | 0 | 1 | 0 | 0.0769 | 0.477 | 0.036 |
| Large | 0 | 1 | 0 | 0.0769 | 0.477 | 0.036 |
| Complex | 0 | 1 | 0 | 0.0769 | 0.477 | 0.036 |
| Datasets | 0 | 1 | 0 | 0.0769 | 0.477 | 0.036 |
| Phenotyping | 1 | 1 | 0 | 0.0769 | 0.176 | 0.0135 |
| Models | 1 | 1 | 1 | 0.0769 | 0.176 | 0.0135 |
| **DOC3 WORDS** |  |  |  |  |  |  |
| Researchers | 1 | 1 | 1 | 0.0555 | 0 | 0 |
| Build | 0 | 0 | 1 | 0.0555 | 0.477 | 0.026 |
| Computational | 1 | 0 | 1 | 0.0555 | 0.176 | 0.978 |
| Information | 0 | 1 | 2 | 0.1111 | 0.176 | 0.0195 |
| Machine | 1 | 0 | 1 | 0.0555 | 0.477 | 0.026 |
| Learning | 1 | 0 | 1 | 0.0555 | 0.477 | 0.978 |
| engine | 0 | 0 | 1 | 0.0555 | 0.477 | 0.026 |
| Uses | 0 | 0 | 1 | 0.0555 | 0.477 | 0.026 |
| Combine | 0 | 0 | 1 | 0.0555 | 0.477 | 0.026 |
| Gene | 0 | 0 | 2 | 0.1111 | 0.477 | 0.0529 |
| Function | 0 | 0 | 1 | 0.0555 | 0.477 | 0.026 |
| Interaction | 0 | 0 | 1 | 0.0555 | 0.477 | 0.026 |
| Disparate | 0 | 0 | 1 | 0.0555 | 0.477 | 0.026 |
| Genomic | 0 | 0 | 1 | 0.0555 | 0.477 | 0.026 |
| Data | 0 | 0 | 1 | 0.0555 | 0.477 | 0.026 |
| Sources | 0 | 0 | 1 | 0.0555 | 0.477 | 0.026 |