TCSS 554 Information Retrieval and Search

Spring 2019

Assignment 1

Zac Lu [tclu82@uw.edu](mailto:tclu82@uw.edu)

1. The number of word tokens in the database (before: 241,263 words, after text processing: 116,826 words.)
2. The number of unique words in the database: 11,998
3. The number of words that occur only once in the database 6,454
4. The average number of word tokens per document: 595.7111111111111
5. For 30 most frequent words in the database:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Term | Tf | Tf(weight) | df | IDF | tf\*idf | p(term) |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |