|  |  |  |
| --- | --- | --- |
| **Word** | **Term frequency** | **Doc freq** |
| the | 255023 | 1000 |
| of | 159039 | 999 |
| and | 130032 | 1000 |
| in | 103676 | 1000 |
| to | 86516 | 1000 |
| a | 83365 | 999 |
| is | 46814 | 991 |
| for | 41480 | 978 |
| energy | 36346 | 890 |
| as | 31295 | 960 |
| by | 29310 | 992 |
| from | 29047 | 1000 |
| on | 28557 | 976 |
| with | 28017 | 1000 |
| retrieved | 25123 | 1000 |
| s | 23743 | 925 |
| that | 23695 | 940 |
| are | 22454 | 911 |
| power | 20577 | 720 |
| or | 18885 | 908 |
| be | 18688 | 921 |
| edit | 15840 | 937 |
| at | 15827 | 920 |
| was | 14965 | 839 |
| an | 14799 | 938 |
| it | 14446 | 944 |
| this | 13639 | 932 |
| which | 11809 | 899 |
| has | 11477 | 913 |
| new | 11191 | 856 |
| also | 11073 | 962 |
| can | 10782 | 851 |
| solar | 10447 | 507 |
| have | 9991 | 829 |
| 1 | 9599 | 985 |
| may | 9579 | 849 |
| world | 9382 | 741 |
| fuel | 9339 | 544 |
| water | 9216 | 700 |
| gas | 8747 | 592 |
| united | 8417 | 779 |
| used | 8354 | 758 |
| other | 8258 | 865 |
| 2 | 8229 | 962 |
| heat | 8174 | 477 |
| its | 8166 | 854 |
| not | 8139 | 837 |
| more | 8138 | 844 |
| use | 7977 | 831 |
| b | 7911 | 755 |
| carbon | 7893 | 576 |
| such | 7751 | 812 |

|  |
| --- |
| **List of Stop Words** |
| **the** |
| **of** |
| **and** |
| **in** |
| **to** |
| **a** |
| **is** |
| **for** |
| **as** |
| **by** |
| **from** |
| **on** |
| **with** |
| **retrieved** |
| **s** |
| **that** |
| **are** |
| **or** |
| **be** |
| **edit** |
| **at** |
| **an** |
| **it** |
| **this** |
| **which** |
| **has** |
| **also** |
| **1** |
| **2** |

For stop list, top 50 frequent terms are considered and threshold value of 890 for document frequency is taken into consideration.

The best terms are those with a higher tf-idf because those are the most common terms in the document while being rare in collection. The stop words are the very common terms in collection and they produce very low tf-idf.