**R52 and R8 of Reuters 21578**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Reuters 21578** | | | | |
| **# Topics** | **# train docs** | **# test docs** | **# other** | **Total # docs** |
| 0 | 1828 | 280 | 8103 | 10211 |
| 1 | 6552 | 2581 | 361 | 9494 |
| 2 | 890 | 309 | 135 | 1334 |
| 3 | 191 | 64 | 55 | 310 |
| 4 | 62 | 32 | 10 | 104 |
| 5 | 39 | 14 | 8 | 61 |
| 6 | 21 | 6 | 3 | 30 |
| 7 | 7 | 4 | 0 | 11 |
| 8 | 4 | 2 | 0 | 6 |
| 9 | 4 | 2 | 0 | 6 |
| 10 | 3 | 1 | 0 | 4 |
| 11 | 0 | 1 | 1 | 2 |
| 12 | 1 | 1 | 0 | 2 |
| 13 | 0 | 0 | 0 | 0 |
| 14 | 0 | 2 | 0 | 2 |
| 15 | 0 | 0 | 0 | 0 |
| 16 | 1 | 0 | 0 | 1 |

The distribution of documents per class is the following for **R8** and **R52**:

|  |  |  |  |
| --- | --- | --- | --- |
| **R8** | | | |
| **Class** | **# train docs** | **# test docs** | **Total # docs** |
| acq | 1596 | 696 | 2292 |
| crude | 253 | 121 | 374 |
| earn | 2840 | 1083 | 3923 |
| grain | 41 | 10 | 51 |
| interest | 190 | 81 | 271 |
| money-fx | 206 | 87 | 293 |
| ship | 108 | 36 | 144 |
| trade | 251 | 75 | 326 |
| **Total** | **5485** | **2189** | **7674** |

|  |  |  |  |
| --- | --- | --- | --- |
| **R52** | | | |
| **Class** | **# train docs** | **# test docs** | **Total # docs** |
| acq | 1596 | 696 | 2292 |
| alum | 31 | 19 | 50 |
| bop | 22 | 9 | 31 |
| carcass | 6 | 5 | 11 |
| cocoa | 46 | 15 | 61 |
| coffee | 90 | 22 | 112 |
| copper | 31 | 13 | 44 |
| cotton | 15 | 9 | 24 |
| cpi | 54 | 17 | 71 |
| cpu | 3 | 1 | 4 |
| crude | 253 | 121 | 374 |
| dlr | 3 | 3 | 6 |
| earn | 2840 | 1083 | 3923 |
| fuel | 4 | 7 | 11 |
| gas | 10 | 8 | 18 |
| gnp | 58 | 15 | 73 |
| gold | 70 | 20 | 90 |
| grain | 41 | 10 | 51 |
| heat | 6 | 4 | 10 |
| housing | 15 | 2 | 17 |
| income | 7 | 4 | 11 |
| instal-debt | 5 | 1 | 6 |
| interest | 190 | 81 | 271 |
| ipi | 33 | 11 | 44 |
| iron-steel | 26 | 12 | 38 |
| jet | 2 | 1 | 3 |
| jobs | 37 | 12 | 49 |
| lead | 4 | 4 | 8 |
| lei | 11 | 3 | 14 |
| livestock | 13 | 5 | 18 |
| lumber | 7 | 4 | 11 |
| meal-feed | 6 | 1 | 7 |
| money-fx | 206 | 87 | 293 |
| money-supply | 123 | 28 | 151 |
| nat-gas | 24 | 12 | 36 |
| nickel | 3 | 1 | 4 |
| orange | 13 | 9 | 22 |
| pet-chem | 13 | 6 | 19 |
| platinum | 1 | 2 | 3 |
| potato | 2 | 3 | 5 |
| reserves | 37 | 12 | 49 |
| retail | 19 | 1 | 20 |
| rubber | 31 | 9 | 40 |
| ship | 108 | 36 | 144 |
| strategic-metal | 9 | 6 | 15 |
| sugar | 97 | 25 | 122 |
| tea | 2 | 3 | 5 |
| tin | 17 | 10 | 27 |
| trade | 251 | 75 | 326 |
| veg-oil | 19 | 11 | 30 |
| wpi | 14 | 9 | 23 |
| zinc | 8 | 5 | 13 |
| **Total** | **6532** | **2568** | **9100** |

File description

All of these are text files containing one document per line.

Each document is composed by its class and its terms.

Each document is represented by a "word" representing the document's class, a TAB character and then a sequence of "words" delimited by spaces, representing the terms contained in the document.

Pre-processing

From the original datasets, in order to obtain the present files, Ana applied the following pre-processing:

all-terms Obtained from the original datasets by applying the following transformations:

Substitute TAB, NEWLINE and RETURN characters by SPACE.

Keep only letters (that is, turn punctuation, numbers, etc. into SPACES).

Turn all letters to lowercase.

Substitute multiple SPACES by a single SPACE.

The title/subject of each document is simply added in the beginning of the document's text.

no-short Obtained from the previous file, by removing words that are less than 3 characters long. For example, removing "he" but keeping "him".

no-stop Obtained from the previous file, by removing the 524 SMART stopwords. Some of them had already been removed, because they were shorter than 3 characters.

stemmed Obtained from the previous file, by applying Porter's Stemmer to the remaining words. Information about stemming can be found here.

Last updated April 2007.