B0003: IG Number NaN Error

Bug ID: B0003
Title: IG Number NaN Error

Vender: ZHU Renjie	**Product:** News Clustering System	**Version**: 3.0
Severity: High	**Updated**: 2014-11-14	**OS**: Windows 7
Status: Closed	**Assigned to:** Fang Zhou	

Description:

When running program, in the output file of IG number, the IG values for some words are NaN. The cause of this problem should be ln(0), when calculating the pcw_b, we should require that numorator and dominator could be 0 at the same time.

Steps to Reproduce:

Steps to Reproduce:

1.) Prepare the articles for each category under raw_data directory

2.) Make sure that there is not information gain file existed.

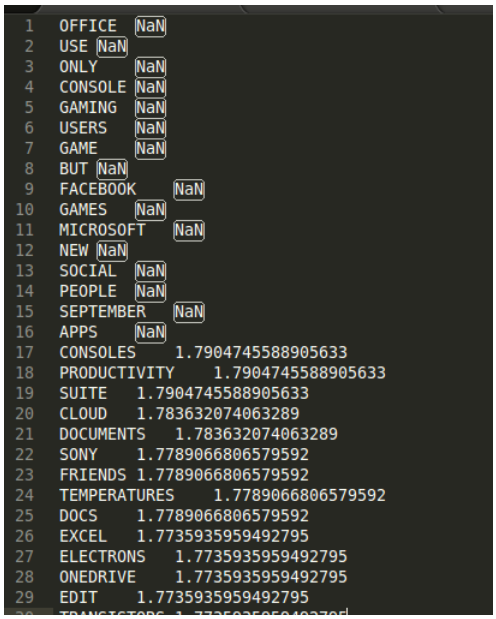
3.) Start the program

Expected result:

All the valid words should have an IG value of double type.

Actual result:

The IG values for some words are NaN.



Fang Zhou 2014-11-15 14:12 EDT

The reason for this problem is that when calculating IG, one attribute called pcw_b[i] is obtained by: pcw_b[i] = 1.0*(num_file_in_one_cat.get(i)-wFileCountCat[i])/(this.totFileNum - wFileCount); The denominator may be 0 when one word appears in all the documents. This could throw Divide by zero exception. To solve it, we filter out these words in BuildDictionary.java. If all documents contain this word, it will be putted into common word list which will not participate in the calculation of pcw_b[i] later.

Before

```
To obtain entropy, we use the following formular:
for(int i=0; i<this.categoryNum; i++){
        entropy_w += pcw[i] * Math.log(pcw[i]);
        entropy_wb += pcw_b[i] * Math.log(pcw_b[i]);
}
```

After

```
for(int i=0; i<this.categoryNum; i++){
entropy_w += pcw[i] * Math.log(pcw[i] + Double.MIN_VALUE);//to avoid pcw[i] = 0 case
entropy_wb += pcw_b[i] * Math.log(pcw_b[i] + Double.MIN_VALUE);//to avoid pcw_b[i] = 0
case
}
```

Zhu Renjie 2014-11-16 17:40 EDT

Test succeeds. No more NaN value for information gain values exists.