Refactoring Report 005

Refactoring ID: 005
Title: Enhance the logic structure of calculating information gain
Date: 2014/11/20
Vender: Acumen **Product**: News Classification & Recommendation
Platform: All platform it supports
PIC: FANG Zhou

Situation (Code Smell):

- In rare cases, some problems (NaN) occur when calculating information gain for one word

Refactoring Plan:

- Handle rare cases by implementing certain programming tricks

Diff:

Original

```
if(tot_file_num-tot_num_file == 0){
	pcw_b = 0;
} else {
	pcw_b = (num_file_in_one_cat[i]-num_file)*1.0/(tot_file_num-tot_num_file);
}


if(pcwf.size() != pcw_bf.size()){
	System.out.println("ERROR");
} else {
	double pw = all_cat_count.get(key)*1.0/tot_file_num;
	double IG = this.entropy + pw * getLogSum(pcwf) + (1-pw)*getLogSum(pcw_bf);

	if(info_gain.get(key) != null){
		System.out.println("ERROR2");
	} else {
		info_gain.put(key, IG);
	}
}
```

Updated

```
pw = 1.0 * wFileCount / this.totFileNum;
for(int i=0; i<this.categoryNum; i++){
	pcw[i] = 1.0*wFileCountCat[i]/wFileCount;
	pcw_b[i] = 1.0*(num_file_in_one_cat.get(i)-wFileCountCat[i])/(this.totFileNum - wFileCount);//Note that pcw_b may be 0
}
double entropy_w = 0, entropy_wb =0;

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
}
		double ig = entropy + pw * entropy_w + (1 - pw) * entropy_wb;
```