Refactoring Report 003

Refactoring ID: 003
Title: Enhance the strategy of selecting common words
Date: 2014/11/1
Vender: Acumen **Product**: News Classification & Recommendation
Platform: All platform it supports
PIC: FANG Zhou

Situation (Code Smell):

- Common words are defined manually, making it exhausting to calculate information gain

Refactoring Plan:

- Add a method to automatically add word to common word dictionary if it appears in all documents

Diff:

Original

```
public static void buildDict(String commonDictPath, String totalDictPath,
ArrayList<Hashtable<String, Integer>> wordCountFileList){
//read the words in common dictionary
ReadContent reader = new ReadContent(commonDictPath);
HashSet<String> comWordSet = new HashSet<String>();//to enhance the searching process
for(String word : reader.processDocument())
    comWordSet.add(word);

    HashMap<String, Integer> totalWords = new HashMap<String, Integer>();
    int fileNum = wordCountFileList.size();

    for(Hashtable<String, Integer> table : wordCountFileList){
        for(String word: table.keySet()){
            if(comWordSet.contains(word))
                continue;
            if(totalWords.containsKey(word)){
                int count = totalWords.get(word) + 1;
                totalWords.put(word, count);
            }else
                totalWords.put(word, 1);
        }
    }


    writeToFile(commonDictPath, totalDictPath, comWordSet, totalWords);
}
```

Updated

we add one more condition to buildDict():

```
//filter the common word and build the word dictionary
for(String word : totalWords.keySet()){
    if(totalWords.get(word) == fileNum)//every file have this word
        comWordSet.add(word);
}
```