B0002: Messy Code in Dictionary

Bug ID: B0002
Title: Messy Code in Dictionary

Vender: ZHU Renjie	**Product:** News Clustering System	**Version**: 1.0
Severity: High	**Updated**: 2014-10-14	**OS**: Windows 7
Status: Closed	**Assigned to:** Fang Zhou	

Description:

In the dictionary generated for each category, there exists some confusing messy code.

Steps to Reproduce:

1.) Prepare the articles under raw_data directory

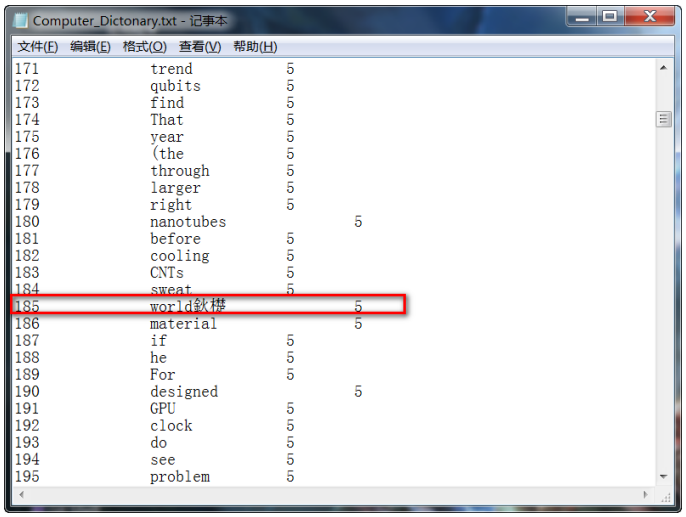
2.) Start the program

Expected result:

Only 26 English letters and numbers can appear in the generated dictionary. Other letters should be filtered.

Actual result:

Mess codes and punctuations like “?”, “)” appear in the generated dictionary.



Fang Zhou 2014-10-15 21:26 EDT

Previously, we just split by space, dots, commas, and Chinese punctuations. So some other characters and Chinese characters may appear in the document. To fix it, we just need to add a filter which will only allow 26 English letters, numbers and“-”, “_” to be read.

Before

```
while(scanner.hasNextLine())
{
	line=scanner.nextLine();
	s=line.split(",|\\.|\\s+|\t|\"|\'|????????? ?????????");
	copyArray(s);
}
```

After

```
String tokens[] = line.split("[^a-zA-Z-_0-9]");

for(String token: arr){
	if(!token.trim().isEmpty() && token.matches("[a-zA-Z-_]+"))
		this.words.add(token.toLowerCase());//all words in lower case
}
```

Zhu Renjie 2014-10-16 20:40 EDT

Test succeeds. No more mess codes and meaningless punctuations show in the output file.