

## BiGram

```
1 '''
2 Created on Feb 4, 2013
3
4 @author: Masum
5 '''
6 import os
7
8 from BiGramHTMLParser import BiGramHTMLParser
9 from collections import defaultdict
10
11 class BiGram():
12     def __init__(self, config):
13         self.config = config
14         self.ht = defaultdict(int)
15         self.htmlparser = BiGramHTMLParser(self.config, self.ht)
16         self.start_batch_processing()
17         self.write_file_map()
18
19     def start_batch_processing(self):
20         file_id=0
21
22         for in_file in os.listdir(self.config['str_src_dir']):
23             #if in_file not in ['medium.html','simple.html']: continue #for testing
24             with open(self.config['str_src_dir']+ in_file, 'r') as f:
25                 self.htmlparser.feed(f.read(),file_id)
26                 file_id += 1
27
28     def write_file_map(self):
29         #writing bigram file to a file named under document id
30         with open(self.config['str_dst_dir']+ self.config['str_doc_id_file_name'],'wb+') as f:
31             for words, count in sorted(self.ht.iteritems(), key=lambda (k,v): (v,k), reverse=
32 True):
33                 f.write(words[0]+" "+words[1]+" "+str(count)+"\n")
```