```
'''
Created on Mar 11, 2012

@author: Masum
'''
import sys
from HTMLParser import HTMLParser
from shared import parse


class BiGramHTMLParser(HTMLParser):
    text, N, newline ='', 0,{'br':'\n','BR':'\n'}

    def __init__(self, config, ht):
        HTMLParser.__init__(self)
        self.config = config
        self.ht = ht


    def handle_data(self, raw):
        self.text = self.text+ raw+ self.newline.get(self.lasttag,'')

    #format is (token =>value, here val)
    def feed(self,data, did):
        #extracting html text
        try:
            HTMLParser.feed(self,data)
        except:
            sys.exc_clear()

        #tokenizing extracted data
        for line in self.text.splitlines():
            tokens= parse(line, self.config)
            if not tokens: continue
            for i in xrange(len(tokens)):
                if i > 0 :
                    self.ht[(tokens[i], tokens[i-1])] +=1;
                    self.N +=1

        #cleaning for next feed
        self.text  = ''; self.N=0
        HTMLParser.reset(self)


```