```
import urllib
import re
from collections import Counter
from collections import OrderedDict
import json
from HTMLParser import HTMLParser
import htmlentitydefs

#HTMLParser functions
class HTMLTextExtractor(HTMLParser):...

def html_to_text(html):
    s = HTMLTextExtractor()
    s.feed(html)
    return s.get_text()

#Get the URL and store into a string and clean it
def f1URLRead(urlStr):
    file = urllib.urlopen(urlStr)
    myfile = file.read()
    myfile = re.sub("<[^<]+?>", "", myfile)
    myfile = re.sub("[^a-zA-Z]+", " ", myfile)
    return myfile

#Get the string and delimit it to a list
def f2Tokenize(inputStr):
    inputStr = inputStr.lower()
    delimit = inputStr.split()
    return delimit

#Get the ignoreList, compare it to the list I have, and get the non-overlapping ones into a list
def f3CompareList(inputList, inputURL):
    ignoreURL = urllib.urlopen(inputURL)
    holdStr = ignoreURL.read()
    holdStr = holdStr.lower()
    ignoreList = holdStr.split()
    cleanList = [i for i in inputList if i not in ignoreList]
    return cleanList
```

```
40      return myfile
41
42  #Get the string and delimit it to a list
43  def f2Tokenize(inputStr):
44      inputStr = inputStr.lower()
45      delimit = inputStr.split()
46      return delimit
47
48  #Get the ignoreList, compare it to the list I have, and get the non-overlapping ones into a list
49  def f3CompareList(inputList, inputURL):
50      ignoreURL = urllib.urlopen(inputURL)
51      holdStr = ignoreURL.read()
52      holdStr = holdStr.lower()
53      ignoreList = holdStr.split()
54      cleanList = [i for i in inputList if i not in ignoreList]
55      return cleanList
56
57  #takes in list, count and return a dict
58  def f4DictFromList(inputList, inputNum):
59      tempDict = Counter(inputList)
60      return OrderedDict(tempDict.most_common(inputNum))
61
62  #takes in map and num input, and turn the string into a json
63  def f5JsonString(inputMap):
64      jsonStr = json.dumps(inputMap)
65      print jsonStr
66
67  #Calls functions
68  firstStr = f1URLRead("https://lyle.smu.edu/~seungkil/3342/index.html")
69  wordList = f2Tokenize(firstStr)
70  listForMap = f3CompareList(wordList, "http://lyle.smu.edu/~coyle/3342/hw3.IgnoreWords.txt")
71  dictForJson = f4DictFromList(listForMap, 5)
72  f5JsonString(dictForJson)
```

```
C:\Python27\python.exe C:/Users/lg/PycharmProjects/URLtoString/URLtoString.py
{"was": 11, "training": 4, "language": 4, "programming": 4, "is": 4}

Process finished with exit code 0
```