**SPARK HOMEWORK**

**1.Write a map-reduce program that creates an inverted list of words and the files that contain them. That is, for each word, you will display a list of files that contain that word. A sample output is given below:**

**‘anger’ [‘histories’, ‘tragedies’]**

**‘laugh’ [‘comedies’, ‘poems’, ‘histories’]**

**…….**

import string

allStopWords={'about':1, 'above':1, 'after':1, 'again':1, 'against':1, 'all':1, 'am':1, 'an':1, 'and':1, 'any':1, 'are':1, 'arent':1, 'as':1, 'at':1, 'be':1, 'because':1, 'been':1, 'before':1, 'being':1, 'below':1, 'between':1, 'both':1, 'but':1, 'by':1, 'cant':1, 'cannot':1, 'could':1, 'couldnt':1, 'did':1, 'didnt':1, 'do':1, 'does':1, 'doesnt':1, 'doing':1, 'dont':1, 'down':1, 'during':1, 'each':1, 'few':1, 'for':1, 'from':1, 'further':1, 'had':1, 'hadnt':1, 'has':1, 'hasnt':1, 'have':1, 'havent':1, 'having':1, 'he':1, 'hed':1, 'hell':1, 'hes':1, 'her':1, 'here':1, 'heres':1, 'hers':1, 'herself':1, 'him':1, 'himself':1, 'his':1, 'how':1, 'hows':1, 'i':1, 'id':1, 'ill':1, 'im':1, 'ive':1, 'if':1, 'in':1, 'into':1, 'is':1, 'isnt':1, 'it':1, 'its':1, 'its':1, 'itself':1, 'lets':1, 'me':1, 'more':1, 'most':1, 'mustnt':1, 'my':1, 'myself':1, 'no':1, 'nor':1, 'not':1, 'of':1, 'off':1, 'on':1, 'once':1, 'only':1, 'or':1, 'other':1, 'ought':1, 'our':1, 'ours ':1, 'ourselves':1, 'out':1, 'over':1, 'own':1, 'same':1, 'shant':1, 'she':1, 'shed':1, 'shell':1, 'shes':1, 'should':1, 'shouldnt':1, 'so':1, 'some':1, 'such':1, 'than':1, 'that':1, 'thats':1, 'the':1, 'their':1, 'theirs':1, 'them':1, 'themselves':1, 'then':1, 'there':1, 'theres':1, 'these':1, 'they':1, 'theyd':1, 'theyll':1, 'theyre':1, 'theyve':1, 'this':1, 'those':1, 'through':1, 'to':1, 'too':1, 'under':1, 'until':1, 'up':1, 'very':1, 'was':1, 'wasnt':1, 'we':1, 'wed':1, 'well':1, 'were':1, 'weve':1, 'were':1, 'werent':1, 'what':1, 'whats':1, 'when':1, 'whens':1, 'where':1, 'wheres':1, 'which':1, 'while':1, 'who':1, 'whos':1, 'whom':1, 'why':1, 'whys':1, 'with':1, 'wont':1, 'would':1, 'wouldnt':1, 'you':1, 'youd':1, 'youll':1, 'youre':1, 'youve':1, 'your':1, 'yours':1, 'yourself':1, 'yourselves':1}

filename = sc.wholeTextFiles("shakespeare/shakespeare1")

filename1 = filename.map(lambda x: (x[0].encode('utf-8').split('/')[-1],x[1].encode('utf-8').translate(None,string.punctuation).lower().split()))

def removeallstopwords(lst):

list1=[]

for res in lst:

if res not in allStopWords and len(res)>1:

list1.append(res)

return list1

nostopwords = filename1.map(lambda x: (x[0],removeallstopwords(x[1])))

values = nostopwords.flatMapValues(lambda x: x).map(lambda x: (x[1],x[0])).distinct().groupByKey()

result = values.map(lambda x: (x[0],list(x[1])))

result.take(50)





**1b).Modify your map-reduce program (in 1) to display the number of times the word occurs in each file. A sample output is shown below:**

**‘anger’ {‘histories’: 3, ‘tragedies’: 8}**

**‘laugh’ {‘comedies’: 7, ‘poems’: 2, ‘histories’: 15}**

**………**

import string

allStopWords={'about':1, 'above':1, 'after':1, 'again':1, 'against':1, 'all':1, 'am':1, 'an':1, 'and':1, 'any':1, 'are':1, 'arent':1, 'as':1, 'at':1, 'be':1, 'because':1, 'been':1, 'before':1, 'being':1, 'below':1, 'between':1, 'both':1, 'but':1, 'by':1, 'cant':1, 'cannot':1, 'could':1, 'couldnt':1, 'did':1, 'didnt':1, 'do':1, 'does':1, 'doesnt':1, 'doing':1, 'dont':1, 'down':1, 'during':1, 'each':1, 'few':1, 'for':1, 'from':1, 'further':1, 'had':1, 'hadnt':1, 'has':1, 'hasnt':1, 'have':1, 'havent':1, 'having':1, 'he':1, 'hed':1, 'hell':1, 'hes':1, 'her':1, 'here':1, 'heres':1, 'hers':1, 'herself':1, 'him':1, 'himself':1, 'his':1, 'how':1, 'hows':1, 'i':1, 'id':1, 'ill':1, 'im':1, 'ive':1, 'if':1, 'in':1, 'into':1, 'is':1, 'isnt':1, 'it':1, 'its':1, 'its':1, 'itself':1, 'lets':1, 'me':1, 'more':1, 'most':1, 'mustnt':1, 'my':1, 'myself':1, 'no':1, 'nor':1, 'not':1, 'of':1, 'off':1, 'on':1, 'once':1, 'only':1, 'or':1, 'other':1, 'ought':1, 'our':1, 'ours ':1, 'ourselves':1, 'out':1, 'over':1, 'own':1, 'same':1, 'shant':1, 'she':1, 'shed':1, 'shell':1, 'shes':1, 'should':1, 'shouldnt':1, 'so':1, 'some':1, 'such':1, 'than':1, 'that':1, 'thats':1, 'the':1, 'their':1, 'theirs':1, 'them':1, 'themselves':1, 'then':1, 'there':1, 'theres':1, 'these':1, 'they':1, 'theyd':1, 'theyll':1, 'theyre':1, 'theyve':1, 'this':1, 'those':1, 'through':1, 'to':1, 'too':1, 'under':1, 'until':1, 'up':1, 'very':1, 'was':1, 'wasnt':1, 'we':1, 'wed':1, 'well':1, 'were':1, 'weve':1, 'were':1, 'werent':1, 'what':1, 'whats':1, 'when':1, 'whens':1, 'where':1, 'wheres':1, 'which':1, 'while':1, 'who':1, 'whos':1, 'whom':1, 'why':1, 'whys':1, 'with':1, 'wont':1, 'would':1, 'wouldnt':1, 'you':1, 'youd':1, 'youll':1, 'youre':1, 'youve':1, 'your':1, 'yours':1, 'yourself':1, 'yourselves':1}

filename = sc.wholeTextFiles("shakespeare/shakespeare1")

filename1 = filename.map(lambda x: (x[0].encode('utf-8').split('/')[-1],x[1].encode('utf-8').translate(None,string.punctuation).lower().split()))

def removeallstopwords(lst):

list1=[]

for res in lst:

if res not in allStopWords and len(res)>1:

list1.append(res)

return list1

nostopwords = filename1.map(lambda x: (x[0],removeallstopwords(x[1])))

values = nostopwords.flatMapValues(lambda x: x).map(lambda x: (x[1]+ '%'+ x[0],1))

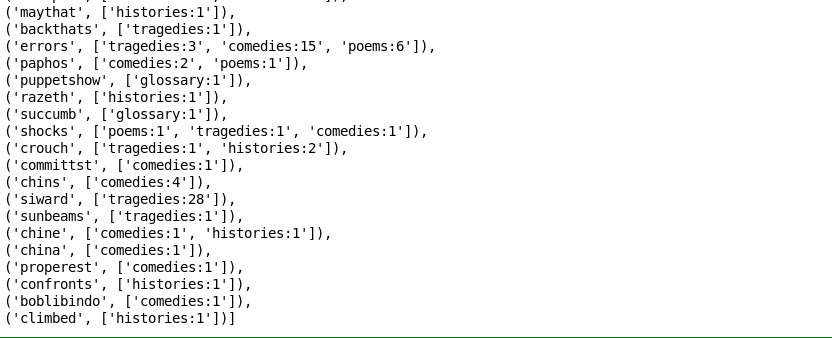
reducer = values.reduceByKey(lambda x,y : x+y)

result = reducer.map(lambda x: (x[0].split('%')[0],(x[0].split('%')[1]+':'+str(x[1])))).groupByKey()

finalresult = result.map(lambda x: (x[0],list(x[1])))

finalresult.take(50)





**2) PROBLEM 2:**

import string

allStopWords={'about':1, 'above':1, 'after':1, 'again':1, 'against':1, 'all':1, 'am':1, 'an':1, 'and':1, 'any':1, 'are':1, 'arent':1, 'as':1, 'at':1, 'be':1, 'because':1, 'been':1, 'before':1, 'being':1, 'below':1, 'between':1, 'both':1, 'but':1, 'by':1, 'cant':1, 'cannot':1, 'could':1, 'couldnt':1, 'did':1, 'didnt':1, 'do':1, 'does':1, 'doesnt':1, 'doing':1, 'dont':1, 'down':1, 'during':1, 'each':1, 'few':1, 'for':1, 'from':1, 'further':1, 'had':1, 'hadnt':1, 'has':1, 'hasnt':1, 'have':1, 'havent':1, 'having':1, 'he':1, 'hed':1, 'hell':1, 'hes':1, 'her':1, 'here':1, 'heres':1, 'hers':1, 'herself':1, 'him':1, 'himself':1, 'his':1, 'how':1, 'hows':1, 'i':1, 'id':1, 'ill':1, 'im':1, 'ive':1, 'if':1, 'in':1, 'into':1, 'is':1, 'isnt':1, 'it':1, 'its':1, 'its':1, 'itself':1, 'lets':1, 'me':1, 'more':1, 'most':1, 'mustnt':1, 'my':1, 'myself':1, 'no':1, 'nor':1, 'not':1, 'of':1, 'off':1, 'on':1, 'once':1, 'only':1, 'or':1, 'other':1, 'ought':1, 'our':1, 'ours ':1, 'ourselves':1, 'out':1, 'over':1, 'own':1, 'same':1, 'shant':1, 'she':1, 'shed':1, 'shell':1, 'shes':1, 'should':1, 'shouldnt':1, 'so':1, 'some':1, 'such':1, 'than':1, 'that':1, 'thats':1, 'the':1, 'their':1, 'theirs':1, 'them':1, 'themselves':1, 'then':1, 'there':1, 'theres':1, 'these':1, 'they':1, 'theyd':1, 'theyll':1, 'theyre':1, 'theyve':1, 'this':1, 'those':1, 'through':1, 'to':1, 'too':1, 'under':1, 'until':1, 'up':1, 'very':1, 'was':1, 'wasnt':1, 'we':1, 'wed':1, 'well':1, 'were':1, 'weve':1, 'were':1, 'werent':1, 'what':1, 'whats':1, 'when':1, 'whens':1, 'where':1, 'wheres':1, 'which':1, 'while':1, 'who':1, 'whos':1, 'whom':1, 'why':1, 'whys':1, 'with':1, 'wont':1, 'would':1, 'wouldnt':1, 'you':1, 'youd':1, 'youll':1, 'youre':1, 'youve':1, 'your':1, 'yours':1, 'yourself':1, 'yourselves':1}

def removeallstopwords(lst):

list1=[]

for res in lst:

if res not in allStopWords and len(res)>1:

list1.append(res)

return list1

filename = sc.wholeTextFiles("MapReduceData1/MapReduceData/hw3data")

filename1 = filename.flatMap(lambda x: (x[1].strip().split('\n'))).map(lambda x: x.encode('ascii','ignore').split(':::')).map(lambda x: (x[1],x[2].lower().split()))

nostopwords = filename1.map(lambda x: (x[0],removeallstopwords(x[1]))).flatMapValues(lambda x : x).map(lambda x: (x[1],x[0].split("::")))

values = nostopwords.flatMapValues(lambda x: x).map(lambda x: (x[1]+ '%'+ x[0],1))

reducer = values.reduceByKey(lambda x,y : x+y)

result = reducer.map(lambda x: (x[0].split('%')[0],(x[0].split('%')[1]+':'+str(x[1])))).groupByKey()

finalresult = result.map(lambda x: (x[0],list(x[1])))

finalresult.saveAsTextFile("hwk3.txt")

