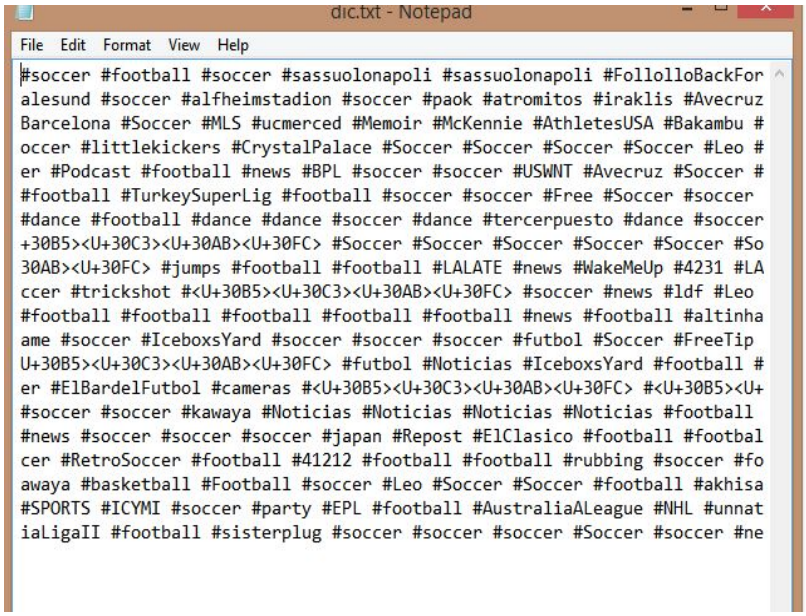
Problem 1 : Word Count on tweets

Input: Tweets for a given domain Output: Word-cloud for the Input Processing: MR on HDFS

1 First of all extracted tweets using the file **Problem1_WordCloud.iypnb** in R and then the extracted tweets were saved in the file dic.txt



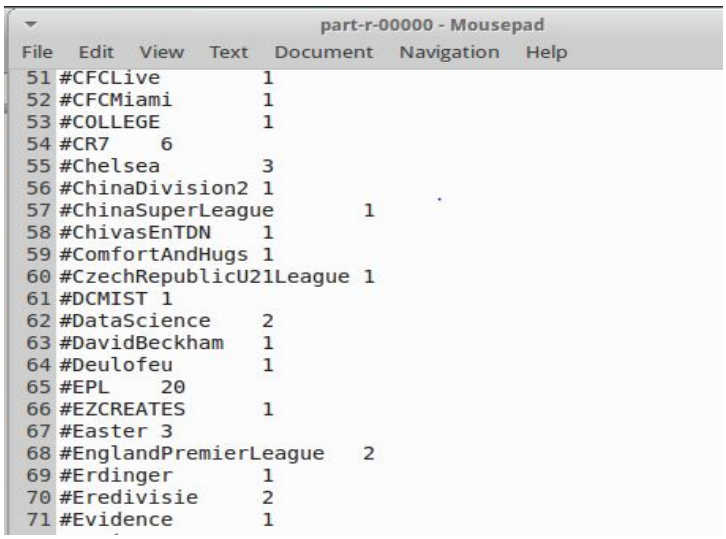
dic.txt - Notepad

File Edit Format View Help

```
#soccer #football #soccer #sassuolonapoli #sassuolonapoli #FollolloBackFor
alesund #soccer #alfheimstadion #soccer #paok #atromitos #iraklis #Avecruz
Barcelona #Soccer #MLS #ucmerced #Memoir #McKennie #AthletesUSA #Bakambu #
occer #littlekickers #CrystalPalace #Soccer #Soccer #Soccer #Soccer #Leo #
er #Podcast #football #news #BPL #soccer #soccer #USWNT #Avecruz #Soccer #
#football #TurkeySuperLig #football #soccer #soccer #Free #Soccer #soccer
#dance #football #dance #dance #soccer #dance #tercerpuesto #dance #soccer
+30B5><U+30C3><U+30AB><U+30FC> #Soccer #Soccer #Soccer #Soccer #Soccer #So
30AB><U+30FC> #jumps #football #football #LALATE #news #WakeMeUp #4231 #LA
ccer #trickshot #<U+30B5><U+30C3><U+30AB><U+30FC> #soccer #news #ldf #Leo
#football #football #football #football #football #news #football #altinha
ame #soccer #IceboxsYard #soccer #soccer #soccer #futbol #Soccer #FreeTip
U+30B5><U+30C3><U+30AB><U+30FC> #futbol #Noticias #IceboxsYard #football #
er #ElBardelFutbol #cameras #<U+30B5><U+30C3><U+30AB><U+30FC> #<U+30B5><U+
#soccer #soccer #kawaya #Noticias #Noticias #Noticias #Noticias #football
#news #soccer #soccer #soccer #japan #Repost #ElClasico #football #footbal
cer #RetroSoccer #football #41212 #football #football #rubbing #soccer #fo
awaya #basketball #Football #soccer #Leo #Soccer #Soccer #football #akhisa
#SPORTS #ICYMI #soccer #party #EPL #football #AustraliaALeague #NHL #unnat
iaLigaII #football #sisterplug #soccer #soccer #soccer #Soccer #soccer #ne
```

2 Input this file to Hadoop system and run all the command given in Hadoop VM Guide

3 Get the output from the Hadoop and saved it as OutputWordCloud.txt



part-r-00000 - Mousepad

File Edit View Text Document Navigation Help

```
#CFCLive        1
#CFCMiami       1
#COLLEGE        1
#CR7    6
#Chelsea        3
#ChinaDivision2 1
#ChinaSuperLeague       1
#ChivasEnTDN    1
#ComfortAndHugs 1
#CzechRepublicU21League 1
#DCMIST 1
#DataScience    2
#DavidBeckham   1
#Deulofeu       1
#EPL    20
#EZCREATES      1
#Easter 3
#EnglandPremierLeague   2
#Erdinger       1
#Eredivisie     2
#Evidence       1
```

4 Used this file again on R file mentioned above and created the Word Cloud using the library WordCloud

Problem 2 : Word co-occurrence on tweets

1 For cleaning and extracting the tweets I used R and did in the file WordCoOccurence.iypnb and created the file dic_lab1.txt

2 dic_lab1.txt is the file that I will input to my hadoop for which I will get the corresponding output shown below are the types of tweet in the txt file.

```
clean_tweet
'unnatural nipples model fayetteville selfie boyfriend soccer ex'   'BrazilUCup Criciuma U v Palmeiras USoccer'
'Team jerseys have arrived Come toth and thNYFEST nyc soccer drink charity'
'Soccer Sport New Nike Magista Opus FG Mens Soccer Cleats ACCLime Green'
'SoccerSeattle Sounders drop in WeekMLS Power Rankings after loss to Vancouver WhitecapsSportsRoadhouse'
'SoccerFor Sounders every decision starts with dataSportsRoadhouse'
'SoccerSan Jose The scout who found Neymar is coming to MLSSportsRoadhouse'
'SoccerImpact forward JacksonHamel looks for playing time after scoring first MLS goalSportsRoadhouse'
'Soccer Sport Nike Hypervenom Phantom FG Soccer Cleats Mens'
'SoccerLA Galaxy slip in latest MLS Power RankingsINSIDERSportsRoadhouse'
'SoccerLA Galaxy sign Jack McInerney after former Timbers striker clears waiversSportsRoadhouse'
'Soccer Sport NEW AUTHENTIC ADIDAS MessiFG Mens Soccer CleatsBlueSilver AQ'
'Leo Messi NIKE LionelMessi FC BARCELONA HOME JerseyFINAL BERLINDETAIL ...'
'Ronaldo CristianoRonaldo REAL MadridSoccerSTARZ MINI Soccer FIGURERealMadrid'
'Watch Cristiano Ronaldos hat trick to become the first player to scoreChampions League goals The.....'
'BT Sport highlight Leicester CityAtleticosoccer ESPN'
'So excited to coach my daughters soccer team for the first time Its going to be a great seasonGo Lady Mustangs...'
'WATCH Goals aside Ronaldo was awfulsoccer ESPN'   'WATCH FC crew rips RealBavern refereesoccer ESPN'
```

3 Now we need to provide the code for running this *operation for pairs* I have used java code mentioned in WordCount1.java and got the corresponding output as shown below

part-r-00000 - Mousepad

File Edit View Text Document Navigation Help

```
RMAFCBucl White 1
RMAFCBucl Who   1
RMAFCBucl Wikki 1
RMAFCBucl Womens        1
RMAFCBucl World 2
RMAFCBucl WorldFootball 1
RMAFCBucl YCIS  1
RMAFCBucl Yanks 1
RMAFCBucl a     2
RMAFCBucl account       1
RMAFCBucl an    1
RMAFCBucl and   3
RMAFCBucl any   1
RMAFCBucl approach      1
RMAFCBucl are   2
RMAFCBucl at    1
RMAFCBucl auto  1
RMAFCBucl autograph     1
RMAFCBucl baseball      1
RMAFCBucl basketball    1
RMAFCBucl bbwla 1
RMAFCBucl beats 1
RMAFCBucl becomes       1
RMAFCBucl bets  1
```

4 Now we need to provide the code for running this *operation for Stripes* I have used java code mentioned in WordCount2.java and got the corresponding output as shown below

```
JankovicSportsRoadhouse { = 240}{yoga = 236}{Kassai = 256}{la = 243}{another = 184}{bboy =
Jardim  { = 241}{yoga = 237}{Kassai = 257}{la = 244}{another = 185}{bboy = 237}{insists =
Ja…     { = 242}{yoga = 238}{Kassai = 258}{la = 245}{another = 186}{bboy = 238}{insists =
JerseyFINAL     { = 243}{yoga = 239}{Kassai = 259}{la = 246}{another = 187}{bboy = 239}{ir
Jerseys { = 244}{yoga = 240}{Kassai = 260}{la = 247}{another = 188}{bboy = 240}{insists =
JobraniHow      { = 245}{yoga = 241}{Kassai = 261}{la = 248}{another = 189}{bboy = 241}{ir
John    { = 246}{yoga = 242}{Kassai = 262}{la = 249}{another = 190}{bboy = 242}{insists =
Jose    { = 247}{yoga = 243}{Kassai = 263}{la = 250}{another = 191}{bboy = 243}{insists =
JuveSportsRoadhouse     { = 248}{yoga = 244}{Kassai = 264}{la = 251}{another = 192}{bboy =
JuventusSportsRoadhouse { = 249}{yoga = 245}{Kassai = 265}{la = 252}{another = 193}{bboy =
Kassai  { = 250}{yoga = 246}{Kassai = 266}{la = 253}{another = 194}{bboy = 246}{insists =
Keeping { = 251}{yoga = 247}{Kassai = 267}{la = 254}{another = 195}{bboy = 247}{insists =
LA      { = 252}{yoga = 248}{Kassai = 268}{la = 255}{another = 196}{bboy = 248}{insists =
LCFCAtleti      { = 253}{yoga = 249}{Kassai = 269}{la = 256}{another = 197}{bboy = 249}{ir
LaLiga  { = 254}{yoga = 250}{Kassai = 270}{la = 257}{another = 198}{bboy = 250}{insists =
Lady    { = 255}{yoga = 251}{Kassai = 271}{la = 258}{another = 199}{bboy = 251}{insists =
Latest  { = 256}{yoga = 252}{Kassai = 272}{la = 259}{another = 200}{bboy = 252}{insists =
Leader  { = 257}{yoga = 253}{Kassai = 273}{la = 260}{another = 201}{bboy = 253}{insists =
League  { = 258}{yoga = 254}{Kassai = 274}{la = 261}{another = 202}{bboy = 254}{insists =
Leeds   { = 259}{yoga = 255}{Kassai = 275}{la = 262}{another = 203}{bboy = 255}{insists =
Left    { = 260}{yoga = 256}{Kassai = 276}{la = 263}{another = 204}{bboy = 256}{insists =
Legend  { = 261}{yoga = 257}{Kassai = 277}{la = 264}{another = 205}{bboy = 257}{insists =
Leicester       { = 262}{yoga = 258}{Kassai = 278}{la = 265}{another = 206}{bboy = 258}{ir
LeicesterAtleti { = 263}{yoga = 259}{Kassai = 279}{la = 266}{another = 207}{bboy = 259}{ir
```

Problem 3 : Featured Activity 1: Wordcount on Classical Latin text

1 For this first we input two files which are given as test files lucan.bellum_civile.part.1.tess and vergil.aeneid.tess

2 Now we need to run the basic algorithm as provided in the featured activity 1 and created a java file WordCountLatin.java which contains the code which will provide me my desired output .

Explanation of the code:
In my code in MAPPER I input the file split on the basis of ”>” and then for each word I normalise the word and then emit the word along with the location.

Then in REDUCER I input all the key value pairs and then check for the existing lemma from the hashmap that Icreated for la.lexicon.csv file and then if the lemma exist i change the word to corresponding lemma and then emit it otherwise I write the word and its location the output.

3 After running the code using all the commands mentioned in the Hadoop VM guide I get my output in the format as shown below:

```
File  Edit  View  Text  Document  Navigation  Help
976 aspera  <luc. 1.42>
977 astra   <luc. 1.232>
978 astra,  <luc. 1.641>
979 asylum. <luc. 1.97>
980 at      <luc. 1.417><luc. 1.352>
981 atque   <luc. 1.217><luc. 1.116><luc. 1.282><luc. 1.689>
982 atra    <luc. 1.541><luc. 1.579>
983 attonitam       <luc. 1.676>
984 attonitus       <luc. 1.616>
985 auctor  <luc. 1.30>
986 auctore <luc. 1.485>
987 auctoribus,     <luc. 1.454>
988 audaces <luc. 1.474>
989 audax   <luc. 1.382>
990 audito  <luc. 1.519>
991 auersi  <luc. 1.54>
992 auertere        <luc. 1.65>
993 aufert. <luc. 1.411>
994 augur   <luc. 1.601>
995 augusti <luc. 1.98>
996 auia    <luc. 1.569>
997 auidumque       <luc. 1.181>
998 auris   <luc. 1.132>
999 auro    <luc. 1.163>
```

Word <loc1><loc2><loc3>
atque <luc. 1.217><luc. 1.116><luc. 1.282><luc. 1.689>

Problem 4 : Featured Activity 2 : Word co-occurrence among multiple documents

1 For this first we will input tess files as above and then input various other tess files provided in hadoop

2 Now we need to run the basic algorithm as provided in the featured activity 1 along with the condition of word Co-ocuurence of two words and then on 3 grams and created a java file WordCoLatin.java which contains the code which will provide me my desired output

Explanation of the code:
In my code in MAPPER I input the file split on the basis of ”>” and then for each word I normalise the word and then see the nwighbour of the word using one more loop and then emit the word and its neioghbour as the key along with the location as the value.

Then in REDUCER I input all the key value pairs split on the basis of “,” and then check for the existing lemma of both the words and applied if-else condiotions if lemma of one word or both words exist from the hashmap that I created for la.lexicon.csv file and then if the lemma exist i change the word to corresponding lemma and then emit it otherwise I write the word and the neighbour and its location as the output.

3 After running the code using all the commands mentioned in the Hadoop VM guide I get my output in the format as shown below:

part-r-00000 - Mousepad

File Edit View Text Document Navigation Help

```
<<Mox,cauda>>   <luc. 1.208>
<<Mox,civis>>   <luc. 1.592>
<<Mox,et>>      <luc. 1.592>
<<Mox,iubeo>>   <luc. 1.592>
<<Mox,magnus>>  <luc. 1.195>
<<Mox,moenia>>  <luc. 1.195>
<<Mox,pavidus>> <luc. 1.592>
<<Mox,prospicus>>       <luc. 1.195>
<<Mox,qui2>>    <luc. 1.195>
<<Mox,saevus>>  <luc. 1.208>
<<Mox,sui>>     <luc. 1.208>
<<Mox,stimulo>> <luc. 1.208>
<<Mox,totus2>>  <luc. 1.592>
<<Mox,ubi>>     <luc. 1.208>
<<Mox,verber>>  <luc. 1.208>
<<Mox,urbs>>    <luc. 1.592><luc. 1.195>
<<Mulciber,Aetnae;>>    <luc. 1.545>
<<multa,aura>>  <luc. 1.132>
<<multa,do>>    <luc. 1.132>
<<multa,in>>    <luc. 1.132>
<<multa,popularis>>     <luc. 1.132>
<<multa,totus>> <luc. 1.132>
<<multa,uulgus;>>       <luc. 1.132>
<<multus,Roma>> <luc. 1.44>
<<multus,arma>> <luc. 1.44>
```

<< Word1,Word2 >> <loc1><loc2>
<<languor,tenuit>> <luc. 1.194>
<<languor,vestigium>> <luc. 1.194>

4 Repeating the same process with taking 3 grams by increasing one more loop in the mapper and java code is given in the file WordCoLatin1.java

The output that I get is as described below

part-r-00000 - Mousepad

File Edit View Text Document Navigation Help

```
<<linquo,tum,iubeo>>    <verg. aen. 3.289>
<<linquo,tum,portus>>   <verg. aen. 3.289>
<<linquo,tum,transtris:>>       <verg. aen. 3.289>
<<linquo,Ortygius,pelagus>>     <verg. aen. 3.124>
<<linquo,Ortygius,portus>>      <verg. aen. 3.124>
<<linquo,Ortygius,volo2>>       <verg. aen. 3.124>
<<linquo,pelagus,volo2>>        <verg. aen. 3.124>
<<linquo,portus,>>      <verg. aen. 3.124><verg. aen. 3.124>
<<Liparen,,arduus>>     <verg. aen. 8.417>
<<Liparen,,fumo>>       <verg. aen. 8.417><verg. aen. 8.417>
<<Liris,Pagasumque,alto>>       <verg. aen. 11.670>
<<Liris,Pagasumque,quis>>       <verg. aen. 11.670>
<<Liris,altum,habenas>> <verg. aen. 11.670>
<<Liris,qui,alto>>      <verg. aen. 11.670>
<<Liris,qui,habenas>>   <verg. aen. 11.670>
<<Liris,super;,alto>>   <verg. aen. 11.670>
<<Liris,super;,quis>>   <verg. aen. 11.670>
<<litus3,contrarius,>>  <verg. aen. 4.628><verg. aen. 4.628>
<<litus3,fluctus,unda>> <verg. aen. 4.628>
<<litus3,litus3,contrarius>>    <verg. aen. 4.628>
<<litus3,litus3,fluctus>>       <verg. aen. 4.628>
<<litus3,litus3,unda>>  <verg. aen. 4.628>
<<locus,ardea,quondam>> <verg. aen. 7.411>
<<longus,aequor,aro>>   <verg. aen. 2.780>
```

<<word1,word2,word3>> <loc1><loc2><loc3>

<<Liris,Pagasumque,alto>><verg. aen. 11.670>
<<Liris,Pagasumque,quis>> <verg. aen. 11.670>
<<Liris,altum,habenas>> <verg. aen. 11.670>

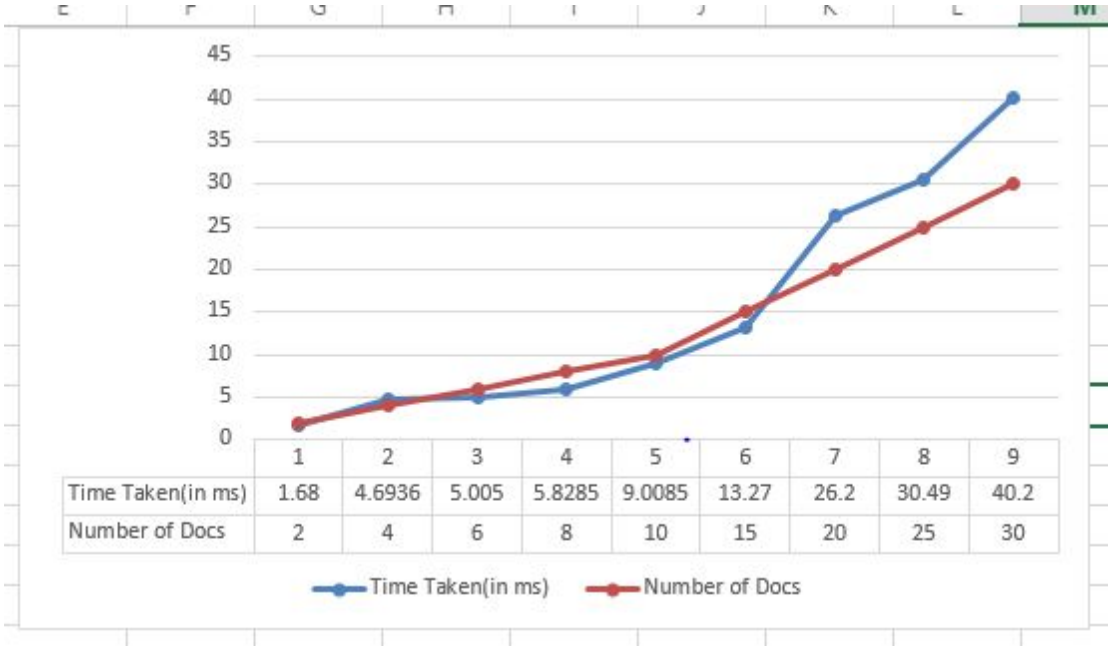
Now we analyse the codes mentioned above for 2 grams and 3 grams on increasing the number of documents to Discuss the results and plot the performance and scalability.

First we see the time taken in order to see the performance of the program by increasing the documents from 2 to 4 to 6 and so on till 30
Then we plot the number of documents and the time taken against each other.

We realize that when there are Word Co Occurence of two words the time taken initally is less and we get good performance for 2 docs but as we increase the number of docs the time taken increses which affects the performance of the system Which leads us to conclude that the performance of the system is inversely proportional to the number of documents which means by increasing the scalability.

The other thing which I realised was the computation was faster when we took only one word as the complexity of program increased for 2 the performance scale reduces and time taken to execute the time increases.

Here is the plot as shown below:

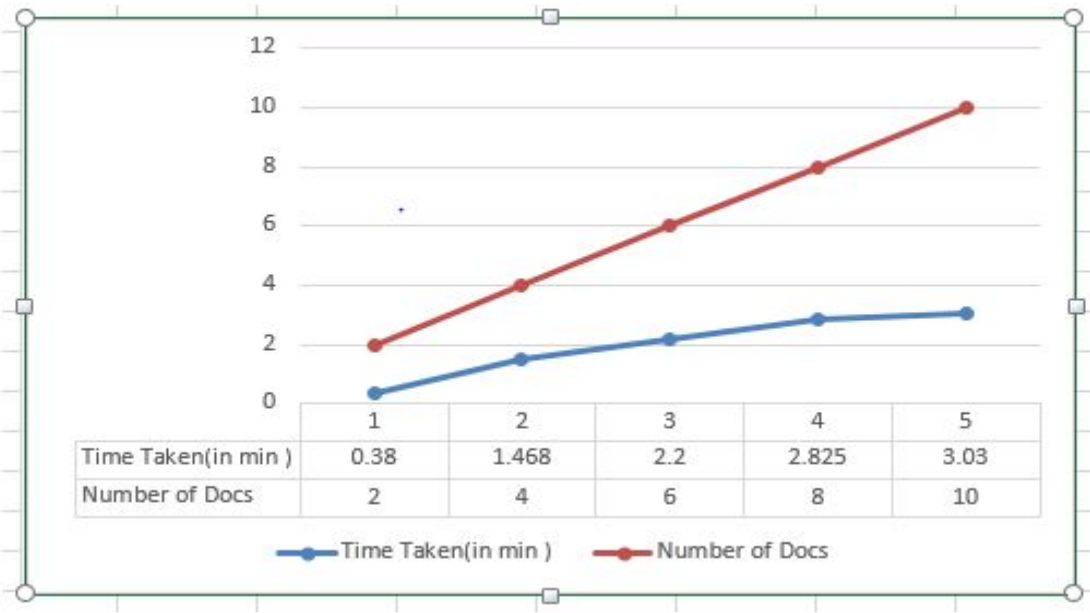


	1	2	3	4	5	6	7	8	9
Time Taken(in ms)	1.68	4.6936	5.005	5.8285	9.0085	13.27	26.2	30.49	40.2
Number of Docs	2	4	6	8	10	15	20	25	30

Now we increase the word Co Occurence to 3 grams and we see that the change is time taken by the `program or the performance is significantly affected.

The time taken increases greatly by small scale increase in number of documents as the complexity of cod increases for this program'

Here is the plot as shown below:



	1	2	3	4	5
Time Taken(in min)	0.38	1.468	2.2	2.825	3.03
Number of Docs	2	4	6	8	10