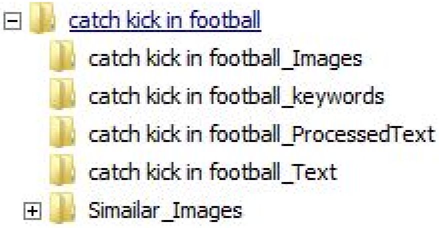
1)TopUrl_100:

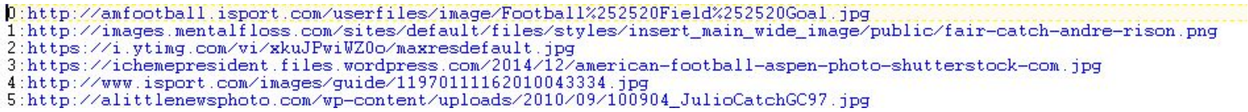
Language type :The script is written in python
FileName:topUrl_100.py
Objective:Download the top100 url and crawl text of pages for a given Query.
Input:Array of Query
Output:suppose the input query was catch kick in football.5 folders will get created



```
⊟ catch kick in football
    catch kick in football_Images
    catch kick in football_keywords
    catch kick in football_ProcessedText
    catch kick in football_Text
  ⊞ Simailar_Images
```

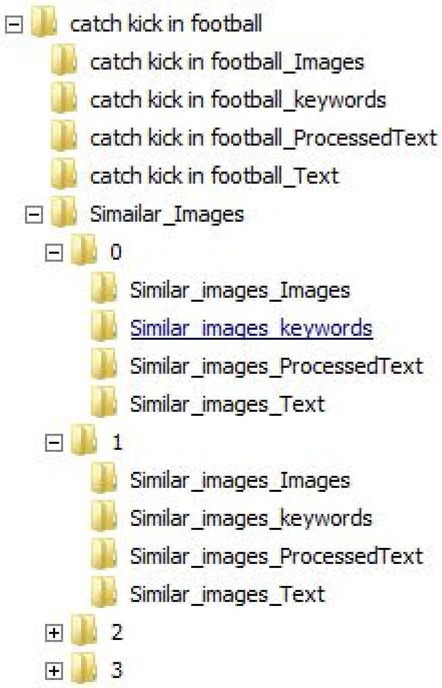
Inside catch kick in football_Text folder there will be 100 text file created(i.e id 0 to id 99) 0 correspond to the first result returned by the google image search for text and 99 the 99th result.

URL of images:the URL of corresponding 100 images will be downloaded to catch kick in football_urls.txt file which is a child directory of directory catch kick in football.For each line in this file the interpretation is as follow:
fileid:url_of_image



```
0:http://amfootball.isport.com/userfiles/image/Football%252520Field%252520Goal.jpg
1:http://images.mentalfloss.com/sites/default/files/styles/insert_main_wide_image/public/fair-catch-andre-rison.png
2:https://i.ytimg.com/vi/xkuJPwiWZ0o/maxresdefault.jpg
3:https://ichemepresident.files.wordpress.com/2014/12/american-football-aspen-photo-shutterstock-com.jpg
4:http://www.isport.com/images/guide/11970111162010043334.jpg
5:http://alittlenewsphoto.com/wp-content/uploads/2010/09/100904_JulioCatchGC97.jpg
```

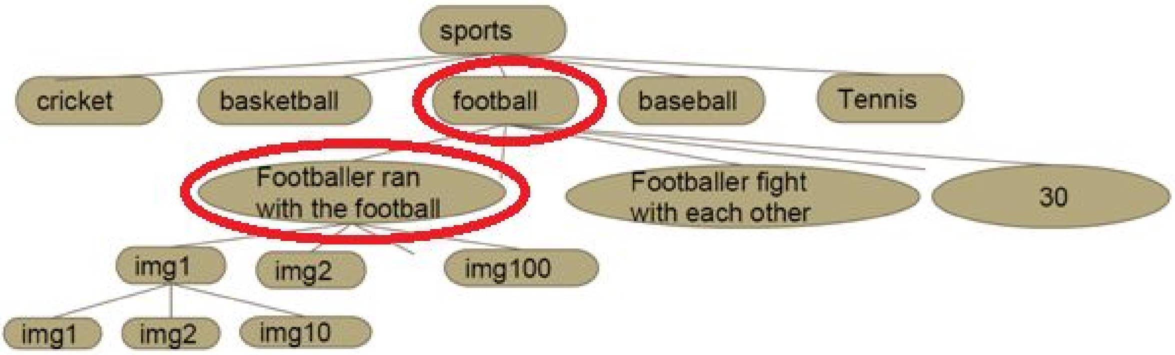
2)TopUrl_10:
Language type :The script is written in python
FileName:topUrl_10.py
Objective:Download the top10 url and crawl text of pages for a given Url in the catch kick in football_urls.text generated by downloading the top100.
Input:catch kick in football_urls.txt file.
Output:Inside the similar_Images folder each folder name corresponds to the text id of top 100 images and consist of top 10 crawled pages for that url .inside similar images folder folder name as text id of top 100 files you will find a filename as Similar_images_urls.txt which consist of top 10 downloaded images url.



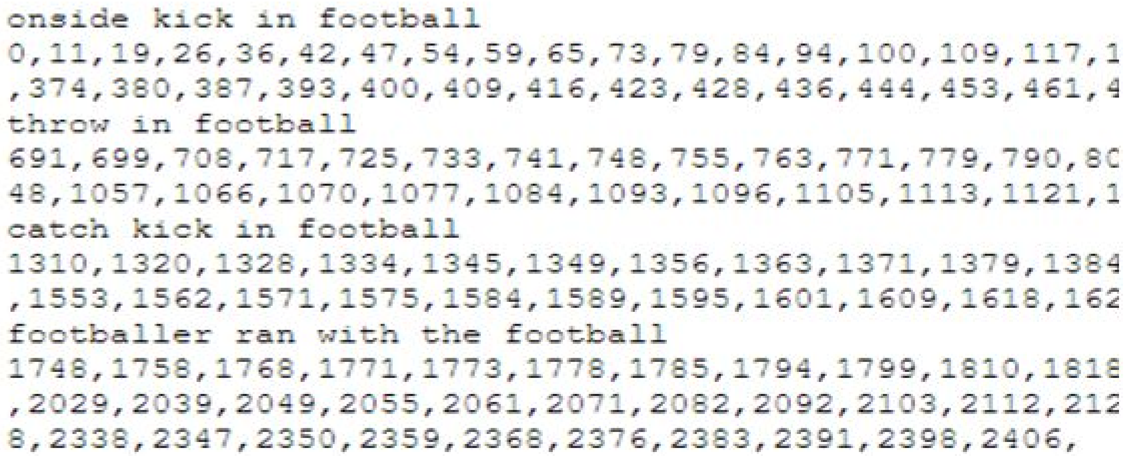
```
⊟ catch kick in football
    catch kick in football_Images
    catch kick in football_keywords
    catch kick in football_ProcessedText
    catch kick in football_Text
  ⊟ Simailar_Images
    ⊟ 0
        Similar_images_Images
        Similar_images_keywords
        Similar_images_ProcessedText
        Similar_images_Text
    ⊟ 1
        Similar_images_Images
        Similar_images_keywords
        Similar_images_ProcessedText
        Similar_images_Text
    ⊞ 2
    ⊞ 3
```

3)Preprocess.py
FileName:
After doing a Domain Analysis we found that many files which had 4 or less number of lines had a repeated or meaningless data,We have deleted such files and all its reference in dataset for top 10 images.It was observed that 1 and (2+3) line of file where exact same so we have deleted such duplicate lines.There were many blank spaces in the files we have deleted all the blank lines.Files which had no contain and where empty are deleted.
Input:Path of any folder whose files you want to preprocess.You could provide path where all the top100 results are stored.Or a specific top100 query path.

4)Organizing file in for image captioning.
FileName:FILEREADER.py
Purpose:Data is organized in the form of standardised dataset,thereby eliminating redundant information.For Each subtype for Sports eg football ,we have represented all the data related to football in 4 files.the description of each of the file is as follow.
Input:Parent folder where top100 results are stored.or you want this for a specific query you can give path of specific top100 query.You can give path of any of the folder circled in red.



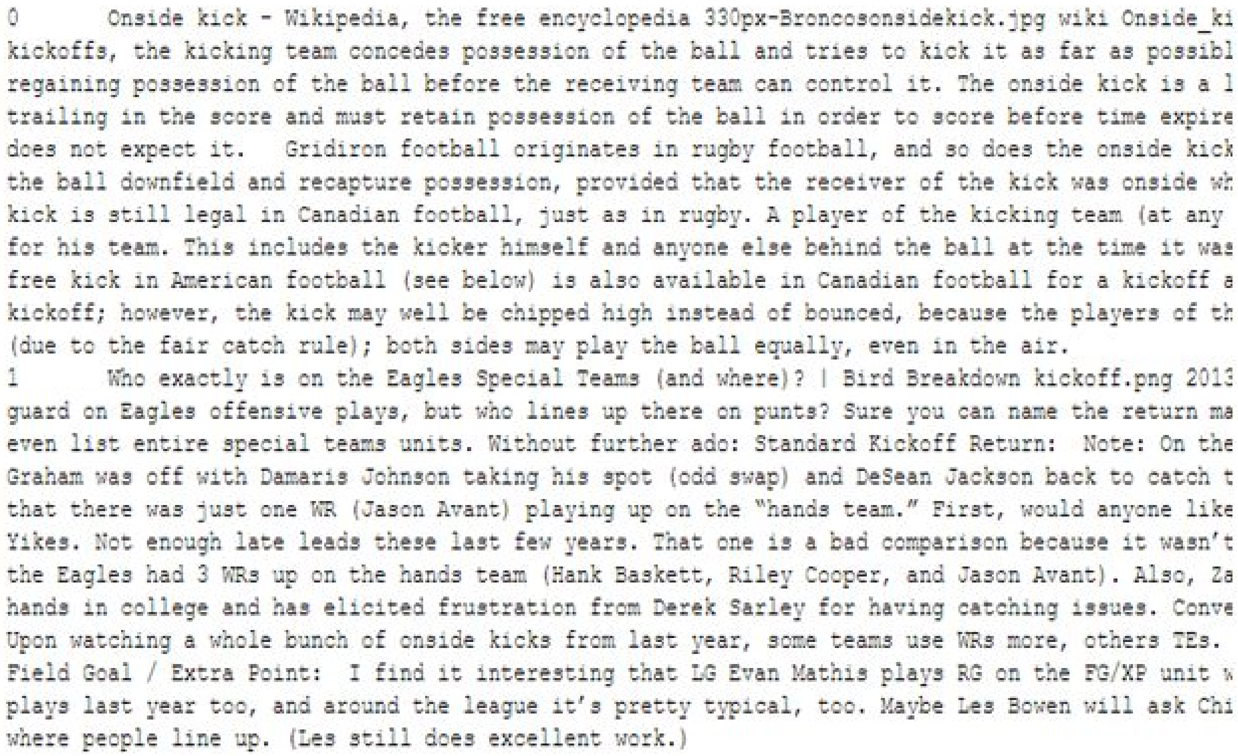
Output:the data of whole folder gets organised in 4 files.the description of files are as follow:
1)**Query_url** : Consist of id of top 100 images for each of the 30 Actions in a specific sport category eg.football.



```
onside kick in football
0,11,19,26,36,42,47,54,59,65,73,79,84,94,100,109,117,1
,374,380,387,393,400,409,416,423,428,436,444,453,461,4
throw in football
691,699,708,717,725,733,741,748,755,763,771,779,790,80
48,1057,1066,1070,1077,1084,1093,1096,1105,1113,1121,1
catch kick in football
1310,1320,1328,1334,1345,1349,1356,1363,1371,1379,1384
,1553,1562,1571,1575,1584,1589,1595,1601,1609,1618,162
footballer ran with the football
1748,1758,1768,1771,1773,1778,1785,1794,1799,1810,1818
,2029,2039,2049,2055,2061,2071,2082,2092,2103,2112,212
8,2338,2347,2350,2359,2368,2376,2383,2391,2398,2406,
```

Here the first line is the **query**.2nd line is the **top 100 image id** for that query(this id signifies the url id,image id and text file id for the top100 result).

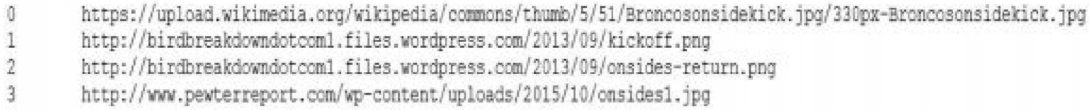
Url_data : Consist of all the text data of all the pages crawled from web related to sport category football.



```
0       Onside kick - Wikipedia, the free encyclopedia 330px-Broncosonsidekick.jpg wiki Onside_ki
kickoffs, the kicking team concedes possession of the ball and tries to kick it as far as possibl
regaining possession of the ball before the receiving team can control it. The onside kick is a l
trailing in the score and must retain possession of the ball in order to score before time expire
does not expect it.   Gridiron football originates in rugby football, and so does the onside kick
the ball downfield and recapture possession, provided that the receiver of the kick was onside wh
kick is still legal in Canadian football, just as in rugby. A player of the kicking team (at any
for his team. This includes the kicker himself and anyone else behind the ball at the time it was
free kick in American football (see below) is also available in Canadian football for a kickoff a
kickoff; however, the kick may well be chipped high instead of bounced, because the players of th
(due to the fair catch rule); both sides may play the ball equally, even in the air.
1       Who exactly is on the Eagles Special Teams (and where)? | Bird Breakdown kickoff.png 2013
guard on Eagles offensive plays, but who lines up there on punts? Sure you can name the return ma
even list entire special teams units. Without further ado: Standard Kickoff Return:  Note: On the
Graham was off with Damaris Johnson taking his spot (odd swap) and DeSean Jackson back to catch t
that there was just one WR (Jason Avant) playing up on the "hands team." First, would anyone like
Yikes. Not enough late leads these last few years. That one is a bad comparison because it wasn't
the Eagles had 3 WRs up on the hands team (Hank Baskett, Riley Cooper, and Jason Avant). Also, Za
hands in college and has elicited frustration from Derek Sarley for having catching issues. Conve
Upon watching a whole bunch of onside kicks from last year, some teams use WRs more, others TEs.
Field Goal / Extra Point:  I find it interesting that LG Evan Mathis plays RG on the FG/XP unit w
plays last year too, and around the league it's pretty typical, too. Maybe Les Bowen will ask Chi
where people line up. (Les still does excellent work.)
```

Each line has the following format
<text id of the file>\t<contain of the file>

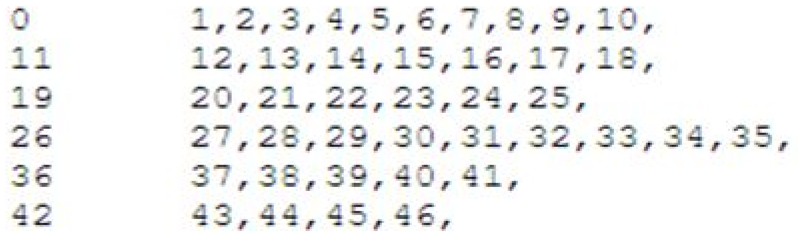
Url_index : Consist of mapping between id and its corresponding image url.



```
0       https://upload.wikimedia.org/wikipedia/commons/thumb/5/51/Broncosonsidekick.jpg/330px-Broncosonsidekick.jpg
1       http://birdbreakdowndotcom1.files.wordpress.com/2013/09/kickoff.png
2       http://birdbreakdowndotcom1.files.wordpress.com/2013/09/onsides-return.png
3       http://www.pewterreport.com/wp-content/uploads/2015/10/onsides1.jpg
```

Each line has the following format
<text id of the file>\t<The corresponding image id of the file>

Url_url : consist of mapping of id of each of top 100 url to its corresponding id of top 10 searched results.



```
0       1,2,3,4,5,6,7,8,9,10,
11      12,13,14,15,16,17,18,
19      20,21,22,23,24,25,
26      27,28,29,30,31,32,33,34,35,
36      37,38,39,40,41,
42      43,44,45,46,
```

Each line has the following format
<text id of the file of top 100 result>\t<Text id of the corresponding top 10 of the top 100>
All the files in All dataset are uniquely named within a category so that there is convenience in understanding and using.

5)Downloading images:

Objective:Downloading the images given the url.
FileName:image_download.py
Input : path of **Url_index**.txt file.
Output : Image downloaded on physical machine in processed folder.images with the size of less than 255 bytes are downloaded in the unprocessed folder. The name of the image follows the following format:
<Text id of the file as in url_index>_.<image format>

6)Image Rename:
Objective:Removing unwanted characters from the image name in the processed folder.
FileName:imagerename.py
Input : absolute path of processed folder.
Output : image name without unwanted character like '\n'

Sequence of execution is step no 1,2,3,4,5,6

Part 2:

Omega index:
FileName:
Omega Index is an index for comparing non-disjoint clustering that is a clustering in which a data point can be assigned to more than one cluster. Omega Index overcomes the drawback of Rand Index which is basically a metric to compare disjoint clustering that is a clustering in which each data point is assigned to only one cluster.

The input file format is as follow:

Line 1:no of cluster
Line2:no of points
Next line onwards 1 cluster on each line

The ground truth file has to be named file.txt.the cluster whose file has to be tested its file should be named file2.txt