Submitted in partial fulfillment of the requirements of

**BITS G540 Research Practice**

BY

RASHMI GULHANE

ID NO: 2015H112187P

Under the supervision of

**Dr. POONAM GOYAL**

Assistant Professor CSIS


**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI, PILANI CAMPUS**

(May 2016)

# Acknowledgements

First of all I liked to express my Sincere gratitude to my supervisor **Dr. Mrs. POONAM GOYAL** for giving me this opportunity to work under her expert guidance and helping me explore this field.

I would also like to thank **Mrs.chandramani choudhary** and **Mrs.dhanashree N P** for their valuable suggestion.

I am also thankful to CSIS Department, BITS Pilani for providing the opportunity to take this Course.

# Table of Contents:

# List of Figures



# List of Tables

# 1.Creation of Dataset for Image Captioning

### 1.1.Introduction

Nowadays sports like tennis,cricket,basketball,baseball and football are becoming very popular.We have many pages on the web which are dedicated to this sports.Extracting text and images relating to this sports will help in dataset creation which can be used by researcher who are working on projects related to image captioning.The process of dataset creation which we have followed is as follow.

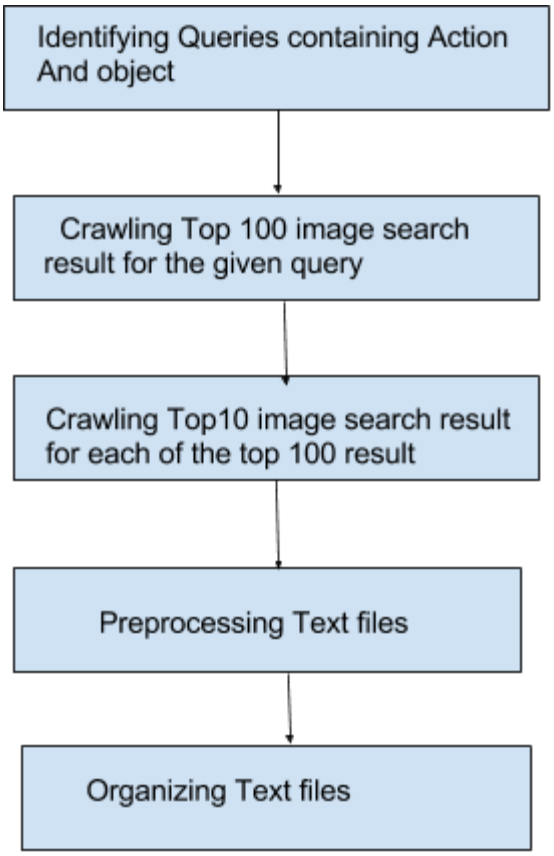


Figure 1

### 2.2.Technical Details

Crawler has been written in python language using selenium webdriver.The default web browser used for downloading is Firefox.Different libraries used for writing the driver are as follow

1)urlparse module is used to break the URL in different parts which are later used to extract text and download images.

2)urllib2 module is used for establishing the connection to the webpage.

3)Beautifulsoup python library is used for extracting text from webpages.

### 2.3.Directory Structure

The extracted images and text are stored in a directory Structure as given below where level 1 signifies the broad topic.Each of the Entity at this level is represented by 30 action.for each of this action top 100 url from google image search are extracted.Latter top 10 url of each of Top 100 Url are extracted.

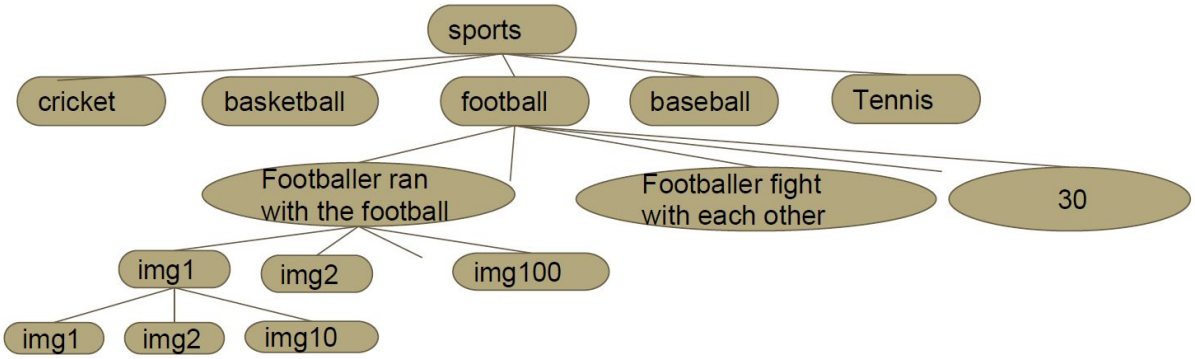


Figure 2:Directory Structure

**2.4.Statistics**

The count of web pages crawled and url downloaded for each of the category are as follow:

| CATEGORY | ACTION | TOP100 | TOP10 | NO OF TEXT FILES | NO. OF URL OF IMAGES |
|---|---|---|---|---|---|
| Cricket | 31 | 3100 | 31000 | 34100 | 34100 |
| Football | 29 | 2900 | 29000 | 31900 | 31900 |
| Baseball | 36 | 3600 | 36000 | 39600 | 39600 |
| Basketball | 30 | 3000 | 30000 | 33000 | 33000 |
| Tennis | 26 | 2600 | 26000 | 28600 | 28600 |

Table 1.Consist of count of files crawled

**2.5.PreProcessing of Data:**

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.After PreProcessing of files the following statistics are observed.

| CATEGORY | ACTION | NO OF TEXT FILES PROCESSED | NO OF TEXT FILES REMOVED | NO OF TEXT FILES REMAINING | NO. OF IMAGES TO BE DOWNLOADED |
|---|---|---|---|---|---|
| Cricket | 31 | 34100 | 17938 | 16162 | 16162 |
| Football | 29 | 31900 | 6648 | 25252 | 25252 |
| Baseball | 36 | 39600 | 14614 | 22286 | 22286 |
| Basketball | 30 | 33000 | 17951 | 15049 | 15049 |
| Tennis | 26 | 28600 | 12971 | 15629 | 15629 |

Table 2.Consist of count of files after preprocessing.

**2.6.Organizing data**

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.

**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,
```

Figure 3

**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.)
```

Figure 4

**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
```

Figure 5

**Url_ur**l : 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,
```

Figure 6

All the files in All dataset are uniquely named within a category so that there is convenience in understanding and using.

# 2.Omega Index Calculation

## 2.1.Literature Survey

Omega Index[1] 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. Rand Index is given as:

$$R.I = a+d/a+b+c+d$$

where:
a = number of pairs of objects assigned to same cluster by both clustering
b = number of pairs of objects assigned to same cluster by first clustering but to different cluster by the other
c = number of pairs of objects assigned to different cluster by first clustering but to same cluster by the other
d = number of pairs of objects assigned to different cluster by both clustering
However, in real clustering it is quite likely that a data point can be assigned to more than one cluster. Such clustering is called overlapping clustering or also known as non-disjoint clustering. Rand Index fails for such clustering.

Adjusted Rand Index is an improvement to Rand Index and gives better metric than R.I, however it has the same limitation of being applicable only to disjoint clustering.

Omega Index considers takes a pair of objects and counts the number of clusters in which that pair appears together. It then calculates an observed agreement in between the solutions which is given by:

Where J and K are the maximum count of the clusters in which any pair of objects appear together in the first and second solutions respectively. $A_j$ is the number of pairs that both solutions have agreed upon to assign to number of clusters j and N is total number of possible object pairs.
The expected agreement between two solutions is given by:

Where $N_{j1}$ is the total number of pairs assigned to number of clusters j in solution 1 and $N_{j2}$ is the total number of pairs assigned to number of clusters j in solution 2.

Omega Index is then finally calculated as:
A highest score of 1 denotes that both solutions agree in a perfect manner on clustering of objects.

## 2.2.Implementation details

Implementation is done in C.The code is scalable and can handle upto 10 lakhs data points with varied number of cluster.The processing of nC2 pairs of data points was done on fly so that memory is not a constraint while calculating omega index with large number of data points.Ground truth and actual Dataset with 0.1% and 5% error rate has been created programmatically.

### 2.3.Results

With 100 clusters and variable number of data points between 1000 to 100000 in the cluster having 0.1% error rate  the results for omega index are as follow:

| Same Number of Clusters for both Data Sets | | | | | | |
|---|---|---|---|---|---|---|
| *Number of Clusters* | *Number of points* | *Number of Varying Points* | *Rand Index* | *Adjusted Rand Index* | *Omega Index* | *Variation* |
| 100 | 1000 | 1 | 0.999922 | 0.995637 | 0.995637 | 0.10% |
| 100 | 10000 | 10 | 0.999956 | 0.997775 | 0.997775 | 0.10% |
| 100 | 50000 | 50 | 0.999961 | 0.99801 | 0.99801 | 0.10% |
| 100 | 100000 | 100 | 0.996201 | 0.992964 | 0.992964 | 0.10% |

Table 3

With 100 clusters and variable number of data points between 1000 to 100000 in the cluster having 5% error rate  the results for omega index are as follow:

| Same Number of Clusters for both Data Sets | | | | | | |
|---|---|---|---|---|---|---|
| *Number of Clusters* | *Number of points* | *Number of Varying Points* | *Rand Index* | *Adjusted Rand Index* | *Omega Index* | *Variation* |
| 100 | 1000 | 50 | 0.998999 | 0.946865 | 0.946865 | 5% |
| 100 | 10000 | 500 | 0.998301 | 0.975185 | 0.975185 | 5% |
| 100 | 50000 | 2500 | 0.997392 | 0.953821 | 0.953821 | 5% |
| 100 | 100000 | 5000 | 0.996358 | 0.991328 | 0.991328 | 5% |

Table 4

With variable clusters in ground truth and cluster to be evaluated, variable number of data points between 1000 to 100000 in the cluster having 0.1% error rate  the results for omega index are as follow:

| Varying Number of Clusters for both Data Sets | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Number of Clusters* | | *Number of points* | | *Number of Varying Points* | *Rand Index* | *Adjusted Rand Index* | *Omega Index* | *Variation* |
| *Dataset1,Dataset2* | | | | | | | | |
| 100 | 90 | 1000 | | 1 | 0.997981 | 0.897999 | 0.897999 | 0.10% |
| 100 | 90 | 10000 | | 10 | 0.997963 | 0.905539 | 0.905539 | 0.10% |
| 100 | 90 | 50000 | | 50 | 0.997958 | 0.905991 | 0.905991 | 0.10% |
| 100 | 90 | 100000 | | 100 | 0.997985 | 0.975681 | 0.975681 | 0.10% |

Table 5

# 3.Conclusion and Future Work

In my research I have created dataset containing images and Text relating to sports such as cricket,football,basketball,baseball and tennis.After crawling files relating to this category from web it is preprocessed and organised.Images from the corresponding url are downloaded.A program for Omega index has been written in C which will help in evaluating soft clusters.Future work is to apply w1 algorithm on this dataset to extract important keywords.

# 4.Reference

[1] G. Murray, G. Carenini, R. Ng, "Using the omega index for evaluating abstractive community detection”
[2]www.google.com