# RP Demo

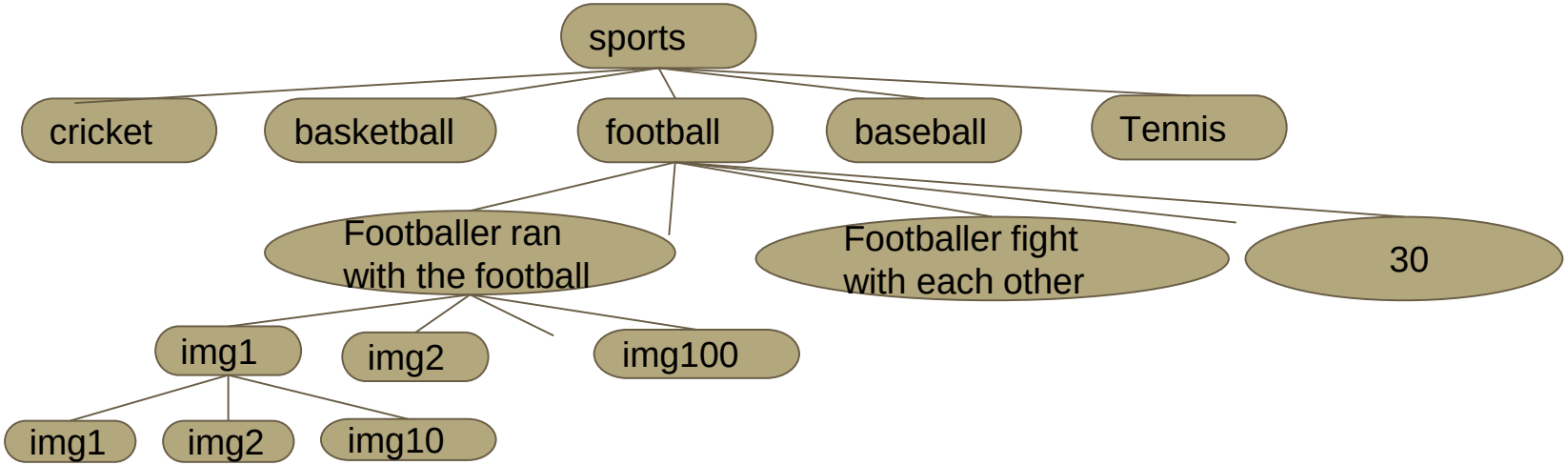
Under guidance of
Dr.Poonam Goyal

Presented By : Rashmi Gulhane(2015H112187P)
Date:10th May2016

# Creation of Dataset for Image Captioning

- Dataset needs to created on Sports with following type of sports Cricket,football,basketball,baseball,Tennis
- Dataset is created by crawling web.The script for crawling is written in Python using Selenium.
- For each sport on average 30 Queries are identified, each query consist of action and object.For each Query top 100 images,then for each of top 100 results its top 10 results , page Text and Url is stored
- So there are a total of around 1,65,000 pages which are crawled and its url for images stored.

# Directory Structure



# Total number of raw files

| 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 |

# PreProcessing

- 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.

# Statistics after preprocessing

| 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 |

# 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.

  Query_url : Consist of Url of top 100 images for each of the 30 Actions.

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

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

  Url_url : consist of mapping of each of top 100 url to its top 10 url.

- All the files in Football dataset are uniquely named so that there is convenience in understanding and using.

# Query_url.txt

```
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,
```

# url_data

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.)

# url_index

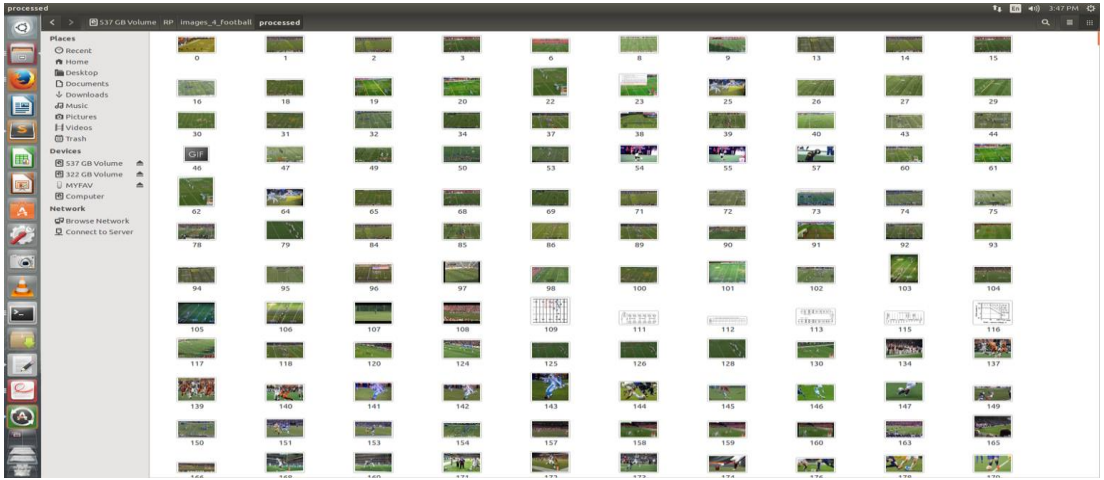
| | |
|---|---|
| 0 | https://upload.wikimedia.org/wikipedia/commons/thumb/5/51/Broncosonsidekick.jpg/330px-Broncosonsidekick.jpg |
| 1 | http://birdbreakdowndotcom1.files.wordpress.com/2013/09/kickoff.png |
| 2 | http://birdbreakdowndotcom1.files.wordpress.com/2013/09/onsides-return.png |
| 3 | http://www.pewterreport.com/wp-content/uploads/2015/10/onsides1.jpg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

# url_url

```
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,
```

# Image Download Code

All the url from the url_index.txt files are extracted and images are downloaded from web.

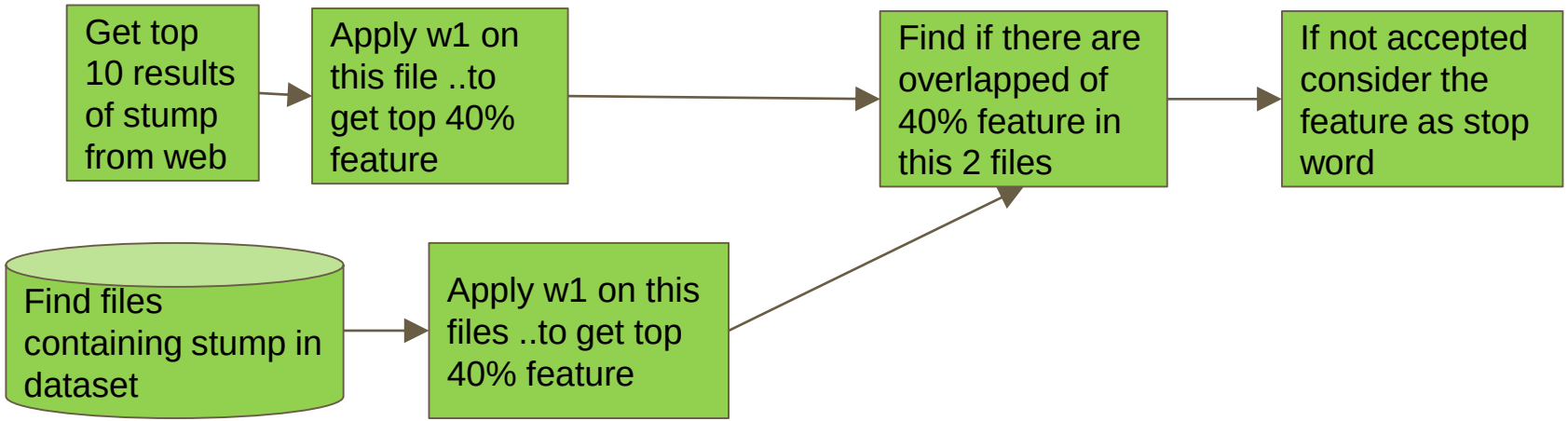


# W1 feature extraction method(In progress)

We have used this program to get the features and its frequency on sample dataset.

We have accepted the top 40% feature.for the rest of the feature we are applying the following approach to decide whether to accept it or reject it.

eg .let the word be stump.



# Implementation of Omega Index for Cluster Evaluation

- Implementation is done in C
- The code is scalable and can handle upto 10 lakhs data points.
- Creation of Dataset for Testing purpose(Ground Truth and a Dataset with 0.1% error in it)
- Program to create Ground Truth and actual Dataset with 0.1% of error has been Written

# Readings of Omega Index

| 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 | 0.992364 | 0.992364 | 0.10% |

| 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 | 0.997992 | 0.953821 | 0.953821 | 5% |
| 100 | 100000 | 5000 | 0.996358 | 0.991328 | 0.991328 | 5% |

| Varying Number of Clusters for both Data Sets | | | | | | |
|---|---|---|---|---|---|---|
| Number of Clusters Dataset1 | Dataset2 | Number of points | Number of Varying Points | Rand Index | Adjusted Rand Index | Omega Index | Variation |
| 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% |

# Thank You