Module 10: Project

Project Help Guide

Version 1.0

Module 10– Project

Project Help Guide

Table of Contents

Introduction

This document will tell you how to analyse the NFL dataset and generate the optimised output for the same.

Analysing the Dataset

The number of steps applied for analysing the dataset are mentioned below:

1. You have to point towards the folder containing the input data.

The input file containing the dataset **"NFL_SocialMedia_sample_data1.csv"** is in the following format:

content	id	tstamp	profilelink	screenname	timezone
NFL flexes Dallas	cbbbcf9395705611c3eeeffaa610a602	2012-12-24T09:51:43Z	http://a0.twi	Fight4EveryYard	Pacific Time (U
@special_event3	9b50b8be10460eab6c0f6f3590067bd7	2012-12-24T09:52:19Z	http://a0.twi	_jpappps	Quito
RG3 leads Redskir	77e1a37031884642b8d1bccad99516c6	2012-12-24T09:52:30Z	http://a0.twi	CowboysPage	Athens
Correct me if I'm	0d4f533e658b47eefecadee60b61278e	2012-12-24T09:52:37Z	http://a0.twi	jazadal	London
RG3 leads Redskir	a4a58402d1c33f85f3f38c0978255c7c	2012-12-24T09:53:16Z	http://a0.twi	lbgood122	
RT @_2KnoMeIz2	83e6a52e1c43d23659d9916e1302daad	2012-12-24T09:53:48Z	http://a0.twi	Tone301	Central Time (
"@LBSports: Reds	b0a3d2ff3c4f0d86a98a9fecc52a3676	2012-12-24T09:54:29Z	http://a0.twi	MBran23	Arizona
RT @PGPackersN	f84dcd20886cd289b90525099ab7b96a	2012-12-24T09:54:41Z	http://a0.twi	DarrenLohr	Central Time (
(Random) the Re	5a01eefaacd6e21bffac014452d1f74f	2012-12-24T09:55:39Z	http://a0.twi	_DEADLYnovem	Eastern Time (
YOUR Monday Mc	59f3064cb1d918377a71a2f715b52948	2012-12-24T09:56:02Z	http://a0.twi	NBA_Raptors_	Arizona
Old time rivalry- (	e4423e8a88933c744af2e94737494d33	2012-12-24T09:56:08Z	http://a0.twi	mzcaz8941	
RT @Mani_Ova_N	0435b355e2aeecdda515666bd6d09ade	2012-12-24T09:57:34Z	http://a0.twi	_MarvinJr	Eastern Time (
RG3 leads Redskir	84f97a411805092b26f65a6c2d181162	2012-12-24T09:58:22Z	http://a0.twi	ravenschatroom	Quito

The data is divided into the following columns:

→ Content
→ Id
→ tstamp
→ profilelink
→ screenmane
→ timezone

NFL_SocialMedia_sample_data1.csv file is present in the LMS.

2. After the input data has been fed, read machine log and separate out 'log' and 'time' columns.
3. Convert the machine log into text corpus.
4. Convert to Lower Case.
5. Remove the Stopwords.
6. Remove Punctuations.
7. Remove Numbers.
8. Eliminate the white spaces.

9. Create a dtm (Document Term Matrix)
10. Determine the Term Frequency and tfxidf
11. Use the K-means package to do document clustering.
12. Normalize the Vectors so that Euclidean makes sense.
13. Cluster the data into 10 clusters.

14. Point towards the folder containing the Interim Data i.e. **"Cluster_Out.csv"**

The **Cluster_Out.csv** will look like this:

	Log	Cluster
1	NFL flexes Dallas Cowboys-Washington Redskins game http://t.co/Yim1aAzy	2
2	@special_event32 redskins still suck	2
3	RG3 leads Redskins over Eagles 27-20 (The Associated Press) PHILADELPHIA (AP) -- With one http:/	1
4	Correct me if I'm wrong, but #Giants can still get into playoffs if #Packers def #Vikings + #Redskins	2
5	RG3 leads Redskins over Eagles 27-20 http://t.co/UZqXBQoV	1
6	RT @_2KnoMeIz2LuvMe: BREAKING NEWS - NFL - Cowboys-Redskins flexed to Sunday night game	2
7	"@LBSports: Redskins fan Kevin Durant talks trash about Cowboys http://t.co/PKcXJ0bA" another r	2
8	RT @PGPackersNews: NFL just announced next Sunday's Packers-Vikings game has been moved to	2
9	(Random) the Redskins are a legitimate team now we can beat you with or without a healthy RG3	2
10	YOUR Monday Morning Blog talks NFL and Toronto Raptors YwqLuyK5 CalvinJohnson #BigBen #RGII	2
11	Old time rivalry- Cowboys vs Redskins... Heading for showdown on Sunday night, 12-30-12.	2
12	RT @Mani_Ova_Matter: I'm really mad as shit, I knew we couldn't beat the redskins tho	2
13	RG3 leads Redskins over Eagles 27-20 (The Associated Press): PHILADELPHIA (AP) -- With one more	1

15. Find the top 5 words discussed in each of the 10 clusters. Write these cluster wise top words into the **"TopWords.csv"** file and generate the word cloud for each of the cluster.

The output file containing the cluster-wise topwords will look like as follows:

	Log Group	Log Count	Top Words	Word Count	Counter
redskins	1	119	redskins	120	1
eagles	1	119	eagles	99	2
iii	1	119	iii	48	3
griffin	1	119	griffin	30	4
returns	1	119	returns	28	5
redskins1	2	1813	redskins	1428	1
cowboys	2	1813	cowboys	650	2
nfl	2	1813	nfl	409	3
game	2	1813	game	265	4
dallas	2	1813	dallas	233	5
boyzzredsk	3	1	boyzzredski	1	1

The output file **"Topwords.csv"** is present in the LMS for reference.