
Analyzing NFL Dataset



CS 850 4 - Big Data
Summer 2015

Introduction

- **NFL-** National Football League is a professional American Football League.
 - Proliferation of Social networks has given **power to people** for **expressing** their views about any **topic** which they find interesting.
 - Twitter is a #1 medium where people **express their polarity** about their favourite teams in **real-time**.
-

Business Question - Why all this ?

- This project dealt with **uncovering insights** in a subset of tweets that dealt with NFL teams.
 - Our goal in this project was to **identify 10 clusters** in the provided NFL data set.
 - Also, we intend to **identify the top 5 words** in each cluster.
 - Additionally, we also plan to visualize each cluster with the word cloud for that cluster.
-

Bigger Picture

Clustering of Tweets and finding Top 5 words finds its applicability in following areas :

- User Segmentation based on Interests
 - Trending Topics Detection.
 - Finding Breaking News by Geo Location
 - Sentiment Analysis and Opinion Mining and many more....
-

Dataset Description

The dataset provided for analysis was in CSV format and belonged to tweets data relating to the NFL domain. The data had 2000 samples and 6 features namely:

content, id, tstamp, profile link, screenname, timezone

content	id	tstamp	profilelink	screenname	timezone
NFL flexes Dallas Cowboys-Washington Redskins game	cbbbcf9395705611c3e	2012-12-24T0	http://a0.twimg.com/profi	Fight4EveryY	Pacific Time (US & Canada)
@special_event32 redskins still suck	9b50b8be10460eab6c	2012-12-24T0	http://a0.twimg.com/profi	_jpappps	Quito
RG3 leads Redskins over Eagles 27-20 (The Associated	77e1a37031884642b8	2012-12-24T0	http://a0.twimg.com/profi	CowboysPage	Athens
Correct me if I'm wrong, but #Giants can still get into	0d4f533e658b47eefec	2012-12-24T0	http://a0.twimg.com/profi	jazadal	London
RG3 leads Redskins over Eagles 27-20 http://t.co/UZq	a4a58402d1c33f85f3f	2012-12-24T0	http://a0.twimg.com/profi	lbgood122	

Technologies Used

- R Programming Language (3.2.1)
- RStudio (0.99.447)
- Data Science Packages- caret, tm, word cloud

Scaling this Project :

- Re-implementing same logic in Apache Spark and deploying across several cluster machines, which can handle millions of tweets in real time.
-

Methodology - Details

- Reading CSV data into R workspace as a data frame
 - Converting the raw machine log CSV data that was read into a data frame into a text corpus for the ease of text mining in R.
 - This text corpus was later subjected to **data cleaning**
-

Methodology - Details Contd...

- Cleaned text corpus was later used to create a **document term matrix**.
 - Later the vectors in the document term matrix was **normalized**
 - Finally this normalized matrix was used for **K-means clustering**.
-

Methodology - Details Contd...

What is TF-IDF

- $TF(t) = (\text{Number of times term } t \text{ appears in a document}) / (\text{Total number of terms in the document})$.
 - $IDF(t) = \log_e(\text{Total number of documents} / \text{Number of documents with term } t \text{ in it})$.
-

Results

Cluster 1

seahawks
nfl face
redskins
playoffs
likely
ers
make

Cluster 4

primetime
cowboys
redskins
sunday
night
game
flexed
time
football
winner
moved
shifts
going
get
espn
nfl
packers
vikings
breaking
dec
week
watch
dallas
httr
espn
next
got
prime
nbc
nfl

Cluster 5

happy
playoff
nfl
picture
cowboys
dallas
holidays
sound
merry
wilson
trending
music
christmas
russell
google
new
santa
trend
bruno
tracker

Results Contd..

Top 5 words per cluster

<i>Cluster 1</i>	<i>playoffs, redskins, likely, make, ers</i>
<i>Cluster 2</i>	<i>beat, redskins, cowboys, next, week</i>
<i>Cluster 3</i>	<i>rgiii, redskins, nfl, luck, wilson</i>
<i>Cluster 4</i>	<i>sunday, night, redskins, cowboys, game</i>
<i>Cluster 5</i>	<i>nfl, playoff, picture, happy, holidays</i>
<i>Cluster 6</i>	<i>redskins, cowboys, nfl, game, dallas</i>
<i>Cluster 7</i>	<i>redskins, east, nfc, espnfirsttake, cowboys</i>
<i>Cluster 8</i>	<i>redskins, dallascowboys, sportscenter, leads, eagles</i>
<i>Cluster 9</i>	<i>cowboys, redskins, dreams, east, imagine</i>
<i>Cluster 10</i>	<i>fan, cowboys, redskins, nfl, saints</i>

Key Challenges

- Identifying proper R Packages (that are not outdated) for carrying out our methodologies
 - Identifying necessary R packages compatible with R 3.2.1
 - Unable to install package due to unavailability of package in selected CRAN mirror.
-

Key Findings / Observations

- We observed that some words were repeating in different clusters, suggesting the presence of **overlapping clusters** in the dataset.
 - Also we observed that data cleaning step of removing numbers **removed all instances of 49ers** , which probably would have been a top word in a cluster.
 - Clusters varied with different runs of R program, as K-means clustering provide different clusters each time due to **randomized centroid selection**.
-

Conclusion

- We would like to conclude that Data Science technique of **Clustering** can be useful to uncover **actionable insights** from data.
 - These clusters can be **visualized** for easier **communication of results** to business team.
 - Clustering results can be used to **segment customers** and **classify people**.
-

Future Work

Based on our first iteration of work , we learnt many things that we would like to improve in our future work :

1. **Named Entity Recognition** based Data Cleaning (Which will identify entities like 49ers in data)
2. **Removing Swear Words** based on a pre-defined dictionary.
3. **Scaling** the R project to **production grade with Apache Spark** (MLLIB Library) and **Scala** (ScalaNLP), which can work with **bigger and Real-Time** data.

Q & A

Thank You
