March Madness

Twitter Data Analysis with Big Data Technologies

Meet the team!

Yipei Li

Julian Krauth

Karim El Chamaa

Edward Melgar

Ann-Charlotte Verstreken

Pieter Van Poecke

Alexandre Collot

What is March Madness?

The NCAA Division I men's **basketball tournament** is a single-elimination tournament of 68 college teams that compete for the national championship during March.

Big Data Pipeline: Source

Source	Ingestion	Storage	Processing	Serving

Sentiment140

Bot Repository
by Indiana University

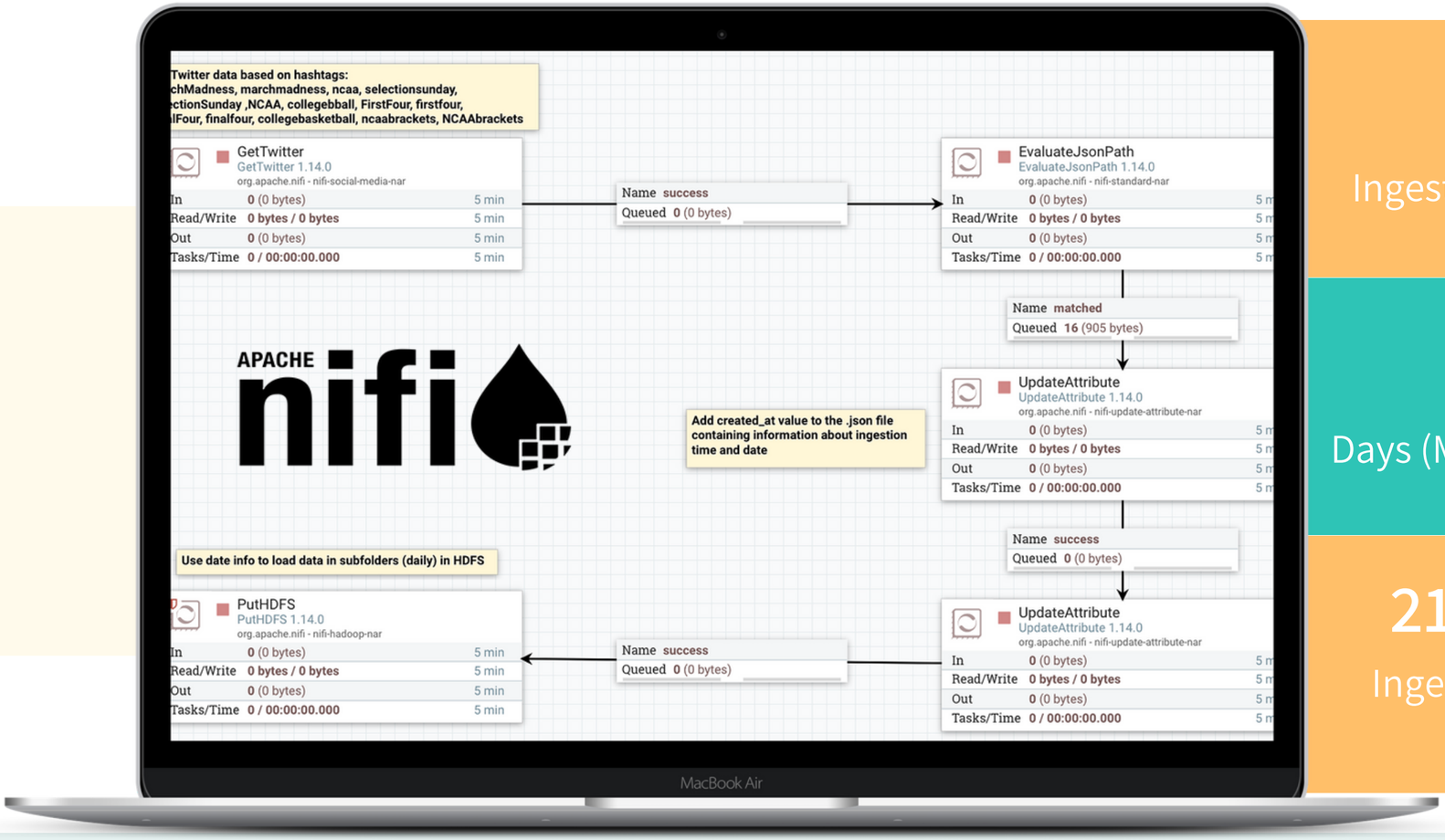
Data Sources

	#MarchMadness, #NCAA, #SelectionSunday, #collegebball, #FirstFour, #FinalFour, #NCAABrackets	JSON file (Semi-structured)
Sentiment140	Pre-annotated Twitter data with positive and negative sentiment category	CSV (Structured)
Bot Repository by Indiana University	Pre-annotated Twitter data with target feature (bot or not bot) in a CSV	JSON file (Semi-structured) CSV (Structured)
	Twitter accounts informaion on 51 college basketball teams	CSV (Structured)

Big Data Pipeline: Ingestion

Source	Ingestion	Storage	Processing	Serving
Sentiment140 **Bot Repository** by Indiana University	APACHE nifi			

NiFi Setup & Configurations



14
Ingested hashtags

4
Days (March 11-14th)

216,673
Ingested tweets

Big Data Pipeline: Storage

Source	Ingestion	Storage	Processing	Serving
Sentiment140 **Bot Repository** by Indiana University	APACHE nifi	hadoop		

HDFS Data Storage

```
/datalake/raw/marchmadness/2022/03/07
/datalake/raw/marchmadness/2022/03/08
/datalake/raw/marchmadness/2022/03/09
/datalake/raw/marchmadness/2022/03/10
/datalake/raw/marchmadness/2022/03/11
/datalake/raw/marchmadness/2022/03/12
/datalake/raw/marchmadness/2022/03/13
/datalake/raw/marchmadness/2022/03/14
```

Date folders

Big Data Pipeline: Processing

Source

Sentiment140

Bot Repository
by Indiana University

Ingestion

Storage

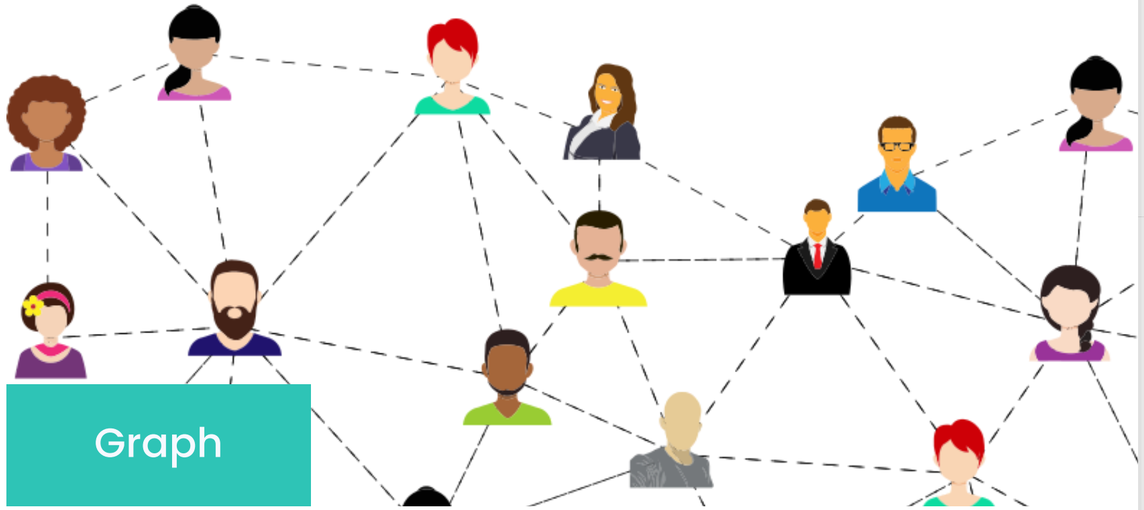
Processing

Serving

Analysis 1: GraphFrame with Hashtags

10 most common hashtag pairs

	src	dst	occurrences
0	MARCHMADNESS	SELECTIONSUNDAY	16573
1	SELECTIONSUNDAY	MARCHMADNESS	16573
2	NCAATOURNAMENT	MARCHMADNESS	1160
3	MARCHMADNESS	NCAATOURNAMENT	1160
4	GOBLUE	FORCOMPETITORSONLY	879
5	FORCOMPETITORSONLY	GOBLUE	879
6	NCAATOURNAMENT	SELECTIONSUNDAY	825
7	SELECTIONSUNDAY	NCAATOURNAMENT	825
8	WAREAGLE	MARCHMADNESS	732
9	MARCHMADNESS	WAREAGLE	732



GraphX

Analysis 1: GraphFrame with Hashtags

261 communities (Strongly Connected Components)

	component	TwitterHashtags
0	0	[UTES, BBQ, BANNERNUMBER6, TERPS, TWITTERMADNESS, 8CLAP, WEGOINSANE, HAWAIISB, UCLAMBB, NJCAA, EBOOK, TNWS, CATAMOUN...
1	12	[LT, 3G]
2	8589934602	[FAKEACTOR, DAFLOP, THEFLOP]
3	8589934604	[REDBIRDBSSKETBALL, GOBIRDS]
4	8589934605	[GOBUFFS, ELEVATEYOURGAME]
...	...	...
256	1374389534751	[TOTALWAR, THESIMS]
257	1374389534752	[DESNEWS, UTAH]
258	1425929142291	[SONICTHEHEDGEHOG, HARRYPOTTER]
259	1434519076873	[DIRECTV, MYTHTV]
260	1529008357391	[MINUSONEMILLIONODDS, TIMETOMAKESOMEDOLLARZ]

Analysis 1: GraphFrame with Hashtags

Most relevant hashtags from pagerank

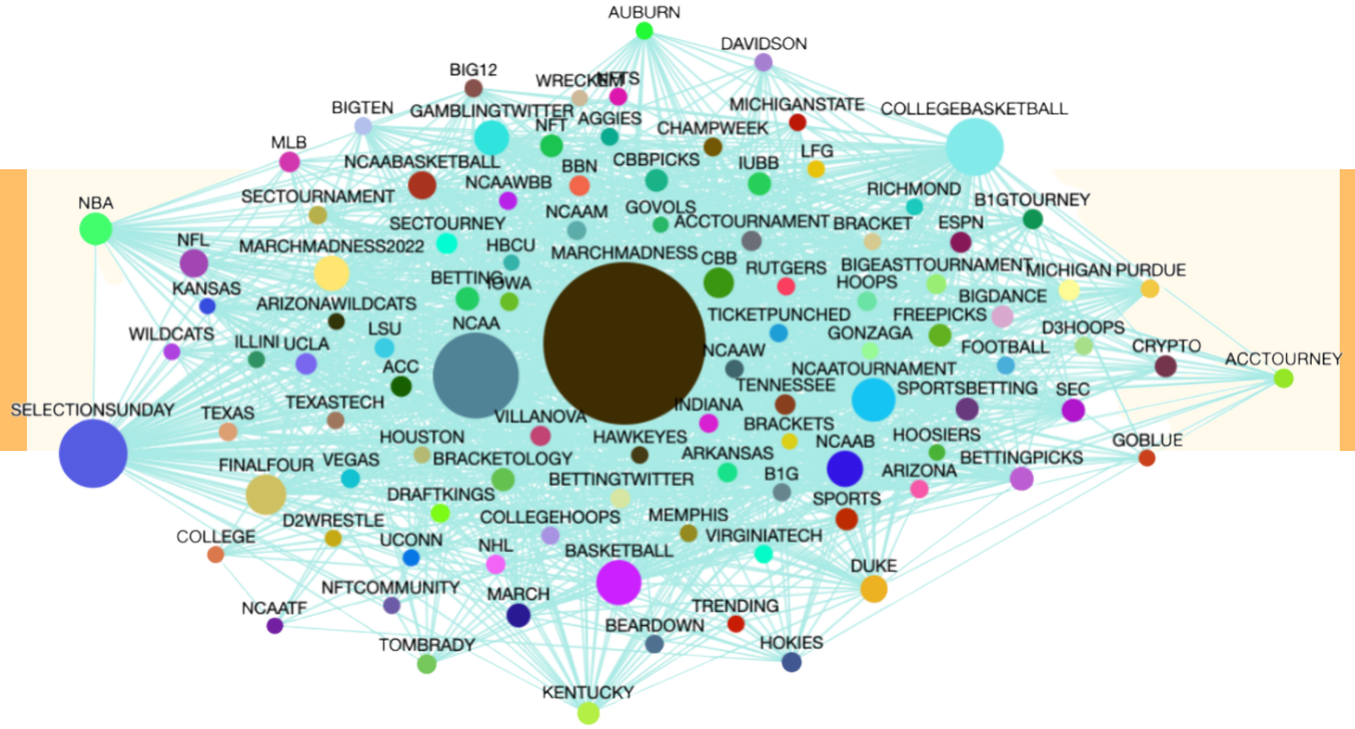
	id	pagerank
0	MARCHMADNESS	384.615978
1	NCAA	107.634358
2	SELECTIONSUNDAY	69.533240
3	COLLEGEBASKETBALL	48.919332
4	BASKETBALL	29.561056

Most important edges

	src	dst	occurrences	weight
0	NCAAQUARTERFINALS	GOMULES	2	1.0
1	YALEBULLDGODS	YALE	1	1.0
2	STIRFRIDAY	MARCHMADNESS	1	1.0
3	TRUTV	MARCHMADNESS	2	1.0
4	HOKIE	MARCHMADNESS	1	1.0

Analysis 1: GraphFrame with Hashtags

Connections between hashtags



5,175
Vertices

33,422
Edges

Analysis 2: Influencer Analysis

Reach score = Number of Followers - Number of Following

Joseph Duarte
@Joseph_Duarte

University of Houston beat writer @HoustonChron » Email: joseph.duarte@chron.com

Journalist Houston, Texas houstonchronicle.com
Joined August 2010

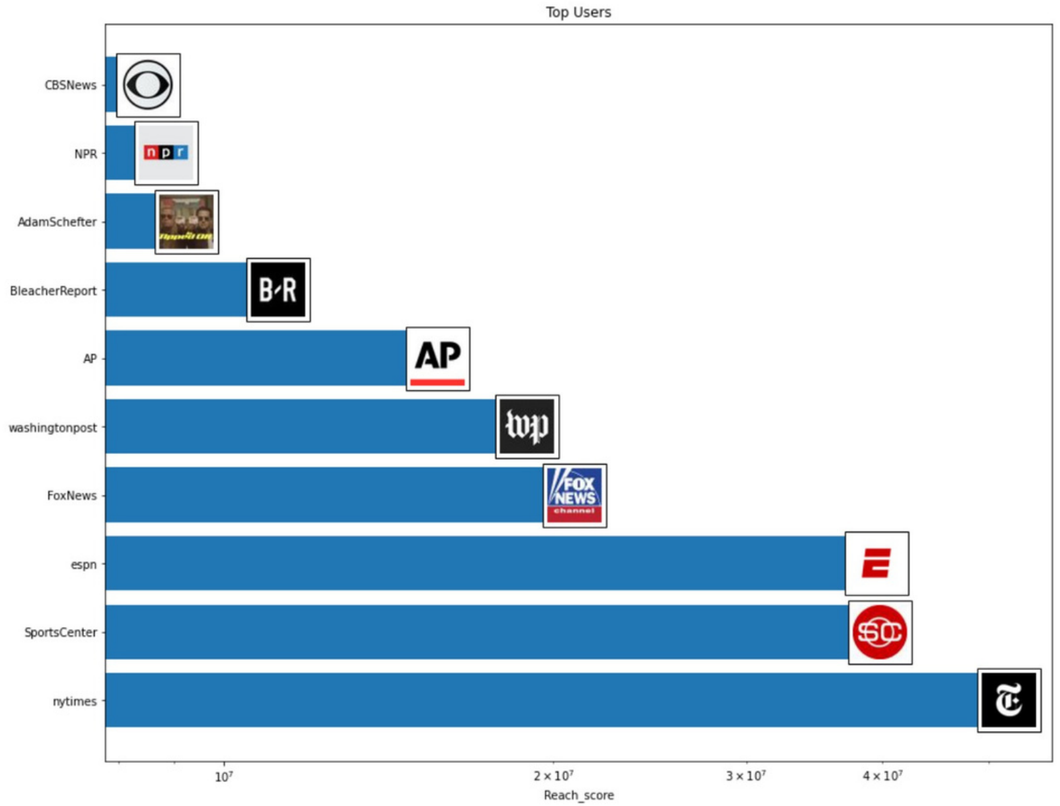
1,052 Following **20.8K** Followers

Reach of 19.7k

Analysis 2: Influencer Analysis

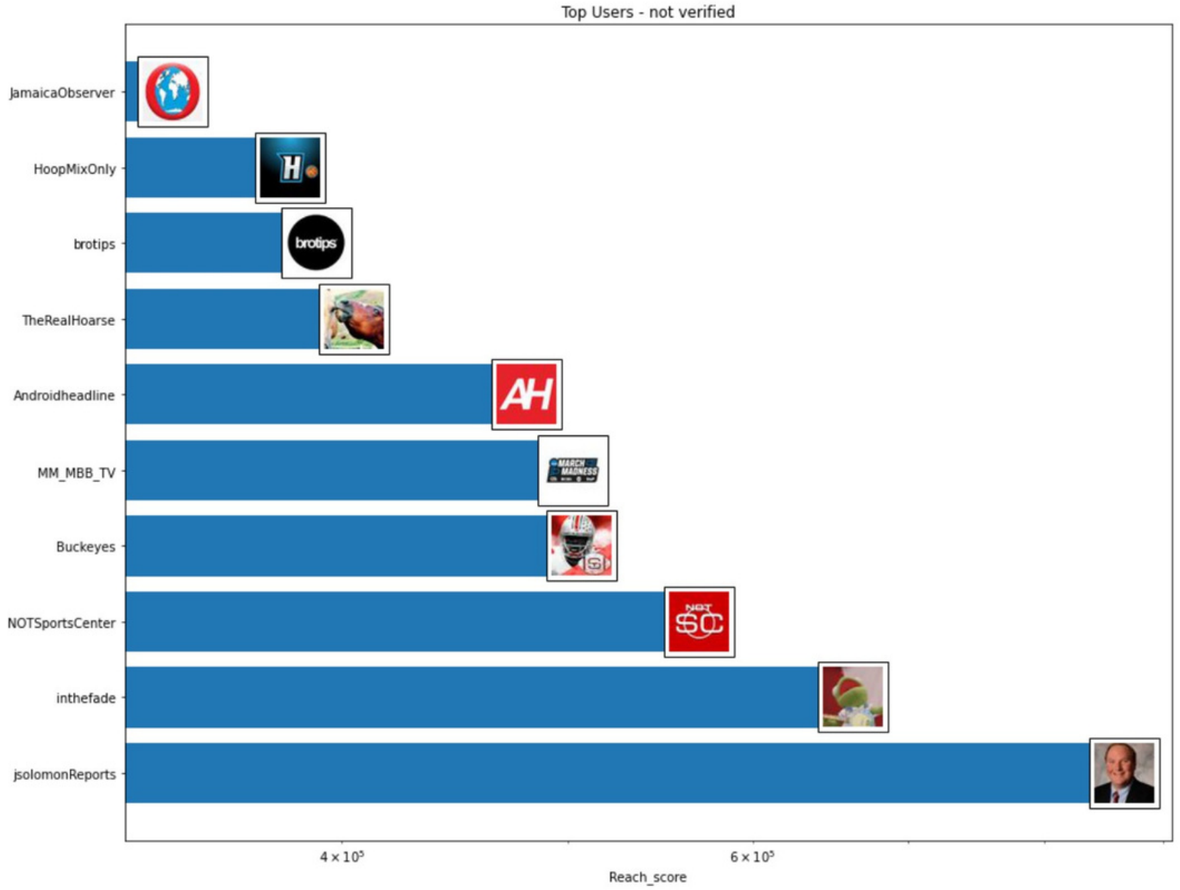
Filter: verified user

	screen_name	Reach_score
0	nytimes	52215515
1	SportsCenter	39738061
2	espn	39526933
3	FoxNews	20893563
4	washingtonpost	18935457
5	AP	15659661
6	BleacherReport	11206268
7	AdamSchefter	9241041
8	NPR	8837140
9	CBSNews	8523156



Analysis 2: Influencer Analysis

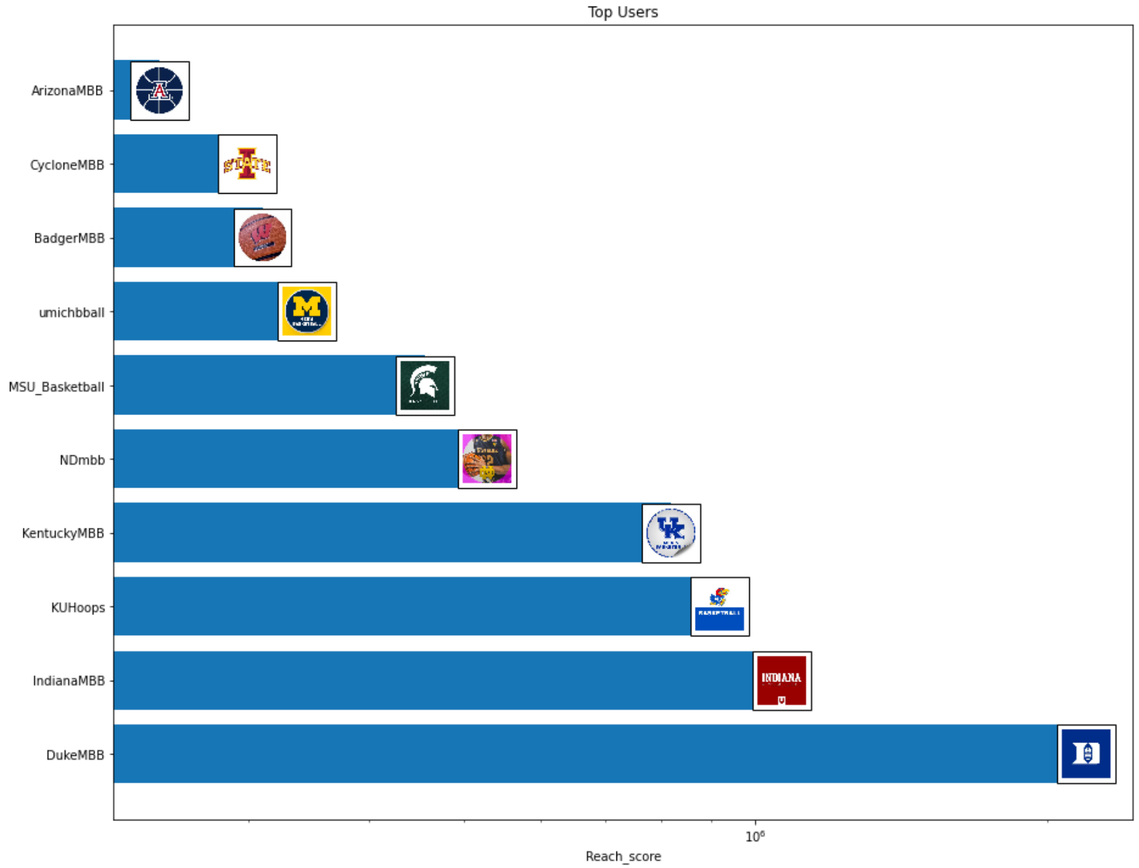
Filter: non-verified user



Analysis 2: Influencer Analysis

Most influential official team accounts

	screen_name	Reach_score
0	DukeMBB	2194734
1	IndianaMBB	1064513
2	KUHoops	917568
3	KentuckyMBB	818431
4	NDmbb	529113



Analysis 3: Sentiment Analysis - SparkML

Popular emojis regarding this subject

	emoji	total
0	🏌️‍♀️	10673
1	🏀	7602
2	🔥	3117
3	🏆	3051
4	🚨	2737

Better approach to understand Tweets' sentiment...

Analysis 3: Sentiment Analysis - SparkML

Sentiment classification model

Sentiment140

80% Training

20% Test

Transformer

- RegexTokenizer
- StopWordsRemover
- CountVectorizer
- StringIndexer

Estimator

- RandomForestClassifier

Evaluator

Model accuracy: **0.711**

Pipeline

Fit

MarchMadness Tweets

Analysis 3: Sentiment Analysis - SparkML

Classification Prediction Outcome

Positive screenshot

text
5 for Iowa??? If you have a spot In NCAA tournament locked in, there's no point in even playing in the conference to...
Looking at these teams you might say boring but college basketball fans see this and think upset city. Thrilled for.....
Spending my Saturday at the cigar lounge watching the games #NC #golf #MarchMadness https://t.co/xrK8JGsmbp
This dude is going to act like MSU didn't make about 4-5 tourneys that they shouldn't have made.
RT @UConnNetch: This is a very subtle way of saying we have one week to make navy throwbacks
@keepsit1000 This is the NCAA. They have consistently shown their inability to do the right thing. Some ADs and ins.....

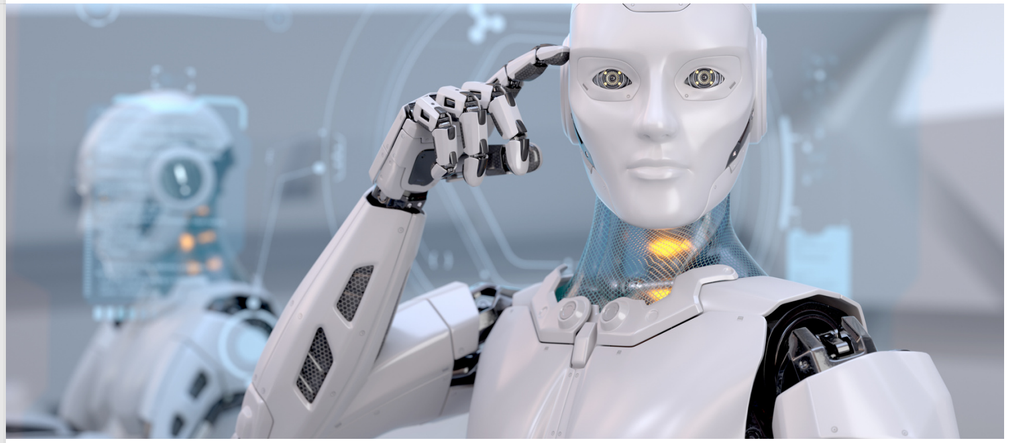
78%

Negative screenshot

text
I'd like this Tennessee team in the NCAA tournament a lot more if I didn't know who their head coach was.
@VOlthoff Rough night?!?!?!? Vicki, that's a nice 'understatement'! And… I (we were) was in the damn front row!!!… ...
RT @ItsPooda: I went here during the wrong time
I went here during the wrong time
I also just realized that I put the wrong link, here's the correct one\n\nhttps://t.co/64qFp032JL
RT @BetandWinHere: Get my #MarchMadness package: https://t.co/xTk0P9Y85N\n\nand watch my YouTube for #ncaabpicks #fr...

22%

Analysis 4: Bot Detection - SparkML



01 Balanced training dataset with labels bot or not bot.

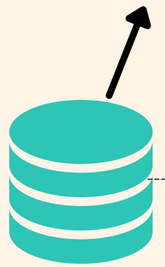
02 The classification is assigned to each Twitter account (id)

03 Data preparation to join the JSON contain accounts information, with CSV having the label for each id.

Analysis 4: Bot Detection - SparkML

Bot classification model

Bot repository
Indiana University



80% Training

20% Test

Transformer

- String Indexer
- Vector Assembler

Estimator

- Grid Search (hyperparameters)
- RandomForestClassifier

Evaluator

CV5 accuracy: **0.73**

Transformer

- String Indexer
- Vector Assembler

Estimator

- Predict using model

Evaluator

Model accuracy: **0.71**

Pipeline

Fit

MarchMadness Tweet users

Analysis 4: Bot Detection - SparkML

Prediction on our ingested Tweet users

Classification	Number of Tweet users	% of dataset
Bot	35,894	~30%
No bot	77,117	~70%

Big Data Pipeline: Serving

Source

Sentiment140

Bot Repository
by Indiana University

Ingestion

Storage

Processing

Serving

Serving Structured Data

Create DB

```
MariaDB [(none)]> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| marchmadness       |
| mysql              |
| performance_schema |
| trucks             |
+--------------------+
5 rows in set (0.003 sec)
```

Predicted Positive Sentiment

```
| 1500969500602159106 | Join me in cheering on our Huskies this week and this weekend!!

#GoHuskies
                                                |          0 |
| 1500969500736466948 | Heading into #MarchMadness tune in and text questions for Coach Wade @lsuradio

                    |          0 |
| 1500969510567882753 | The 2022 “Lawrence Welk Bubble NCAA Basketball Bracket Bonanza” Is Here! https://t.co/mjPryo0yuK

                    |          0 |
| 1500969532919267334 | My guy

                    |          0 |
+---------------------+-----------------------------------------------------------------------
------------------------------------------------------------------------------------------------
--------------------+------------+
1970 rows in set (0.002 sec)
```

```
MariaDB [marchmadness]> describe pos_tweets;
+------------+--------------+------+-----+---------+-------+
| Field      | Type         | Null | Key | Default | Extra |
+------------+--------------+------+-----+---------+-------+
| id         | varchar(30)  | NO   | PRI | NULL    |       |
| text       | varchar(800) | YES  |     | NULL    |       |
| prediction | int(11)      | YES  |     | NULL    |       |
+------------+--------------+------+-----+---------+-------+
3 rows in set (0.001 sec)
```

Predicted Negative Sentiment

```
MariaDB [marchmadness]> select * from neg_tweets limit 10;
+---------------------+-----------------------------------------------------------------------
----------------------------------------------+------------+
| id                  | text
                                              | prediction |
+---------------------+-----------------------------------------------------------------------
----------------------------------------------+------------+
| 1500754058939879424 | RT @pbtips_: I no even get the strength
The 1-cut dey pain me
Na to sleep as i dey like this                                                  |          1 |
| 1500754802686451713 | @HighlifeHeaden Facts. With us coming back we’ll get both of em on schedule in the near future. NCAA g
onna want us… https://t.co/kUJt9KOYqH          |          1 |
| 1500755815673831424 | RT @gvsuwbb: Onto the NCAA Tournament!

The Lakers come in as the No. 2 seed in the Midwest Region with a matchup against the No. 7 Wayne…       |          1 |
| 1500756294222823427 | No. 12 Iowa beats No. 14 Indiana for Big Ten title, NCAA bid | Sports - KIMT: DAILY DATA; The Latest L
ocal COVID-19… https://t.co/rWUQygJKXV          |          1 |
| 1500756591192199171 | One to go for Big East Tournament Championship win #20, then resting/waiting for notification of NCAA
```

```
MariaDB [marchmadness]> describe neg_tweets;
+------------+--------------+------+-----+---------+------
| Field      | Type         | Null | Key | Default | Extra
+------------+--------------+------+-----+---------+------
| id         | varchar(30)  | NO   | PRI | NULL    |
| text       | varchar(800) | YES  |     | NULL    |
| prediction | int(11)      | YES  |     | NULL    |
+------------+--------------+------+-----+---------+------
3 rows in set (0.006 sec)
```

MariaDB

Serving Structured Data

MariaDB

Predicted Bot Accounts

```
MariaDB [marchmadness]> select * from bot_accounts;
+---------------------+------------+
| id                  | prediction |
+---------------------+------------+
| 1413940345507823618 |          0 |
|           304679880 |          0 |
| 1488822035543465986 |          0 |
|           595492132 |          0 |
| 1476165846057644034 |          0 |
| 1322954410016866304 |          0 |
|          1109928390 |          0 |
| 1482798733947547649 |          0 |
|  977266330180210688 |          0 |
| 1436677892788920324 |          0 |
```

```
MariaDB [marchmadness]> describe bot_accounts;
+------------+------------+------+-----+---------+-------+
| Field      | Type       | Null | Key | Default | Extra |
+------------+------------+------+-----+---------+-------+
| id         | bigint(20) | YES  |     | NULL    |       |
| prediction | double     | NO   |     | NULL    |       |
+------------+------------+------+-----+---------+-------+
2 rows in set (0.001 sec)
```

Predicted Real Accounts

```
MariaDB [marchmadness]> select * from real_accounts limit 20;
+---------------------+------------+
| id                  | prediction |
+---------------------+------------+
|  709161467401019392 |          1 |
| 1019798480850685952 |          1 |
|           117332555 |          1 |
|          2147599638 |          1 |
|          1571923273 |          1 |
|            33831655 |          1 |
|            26070516 |          1 |
|          1877794296 |          1 |
|            25133703 |          1 |
|           400124515 |          1 |
|          1420568371 |          1 |
|           556507428 |          1 |
|            67803863 |          1 |
```

```
MariaDB [marchmadness]> describe real_accounts;
+------------+------------+------+-----+---------+-------+
| Field      | Type       | Null | Key | Default | Extra |
+------------+------------+------+-----+---------+-------+
| id         | bigint(20) | YES  |     | NULL    |       |
| prediction | double     | NO   |     | NULL    |       |
+------------+------------+------+-----+---------+-------+
2 rows in set (0.001 sec)
```

March Madness Data Architecture

Source

Sentiment140

Bot Repository
by Indiana University

Ingestion

Storage

Processing

Serving

Thank You!

Group C: March Madness Twitter Data Analysis with Big Data Technologies