Pokemon Go On Twitter Analysis

Geo-location Analysis

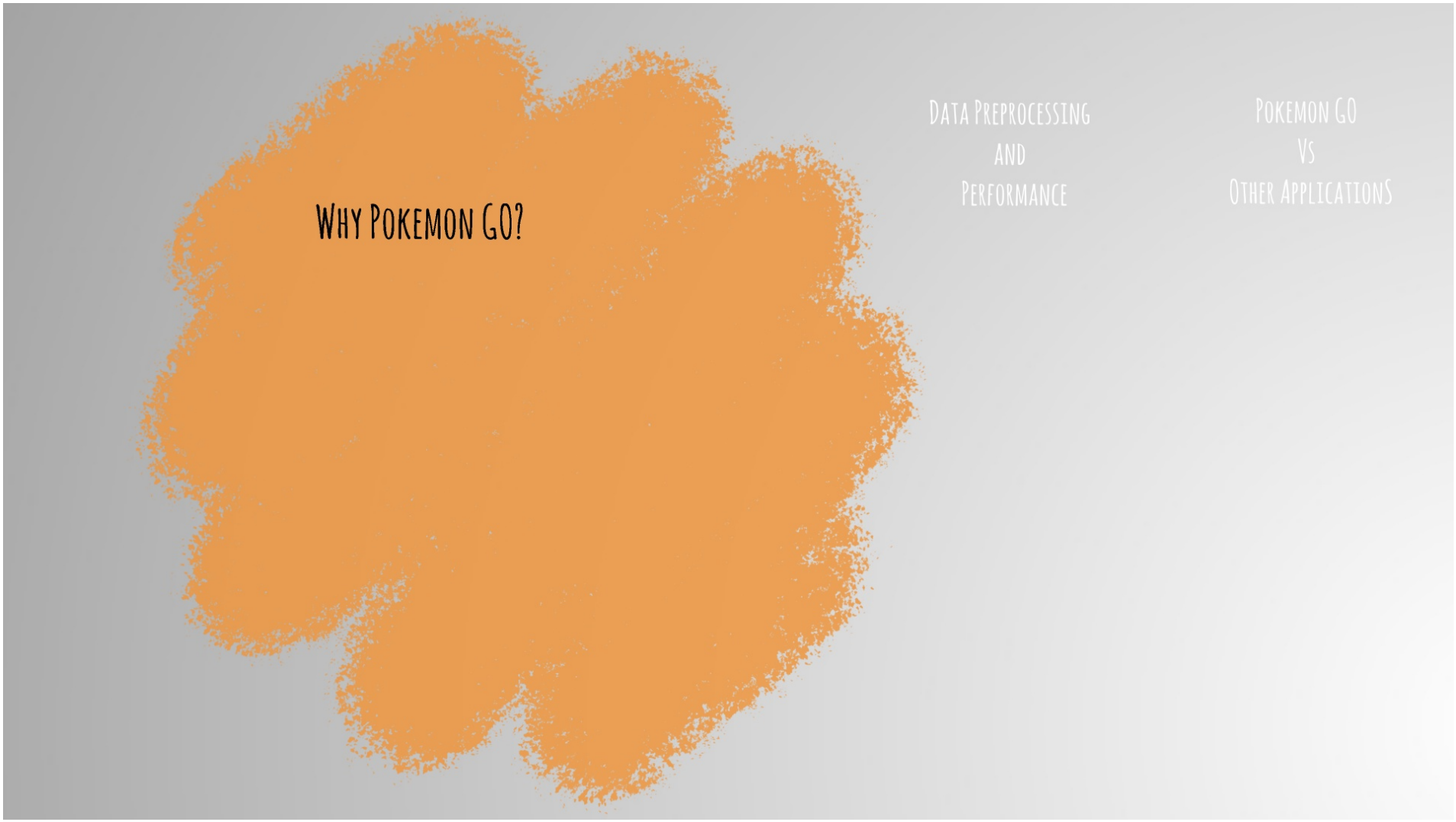
LDA

Why Pokemon GO?

Pokemon GO Facts & Predictions

Sentiment Analysis

- *Anit Gupta*
- *Anil Ravuru*
- *Shivani Bhoite*
- *Alejandro Benitez*
- *Umaraj Potla*


Why Pokemon GO?
Data Preprocessing
and
Performance
Pokemon GO
Vs
Other Applications


Why Pokemon GO?
Major events in 2016
• The U.S. Presidential Election
• The Zika Virus
• The continued threat of ISIS
• The 2016 Summer Olympics in Rio de Janeiro, Brazil
Data Preprocessing and Performance
Pokemon GO Vs Other Applications

Why Pokemon GO?

Major events in 2016

- The U.S. Presidential Election
- The Zika Virus
- The continued threat of ISIS
- The 2016 Summer Olympics in Rio de Janeiro, Brazil

At the same time Nintendo launched a gaming app which revolutionized mobile gaming platform.

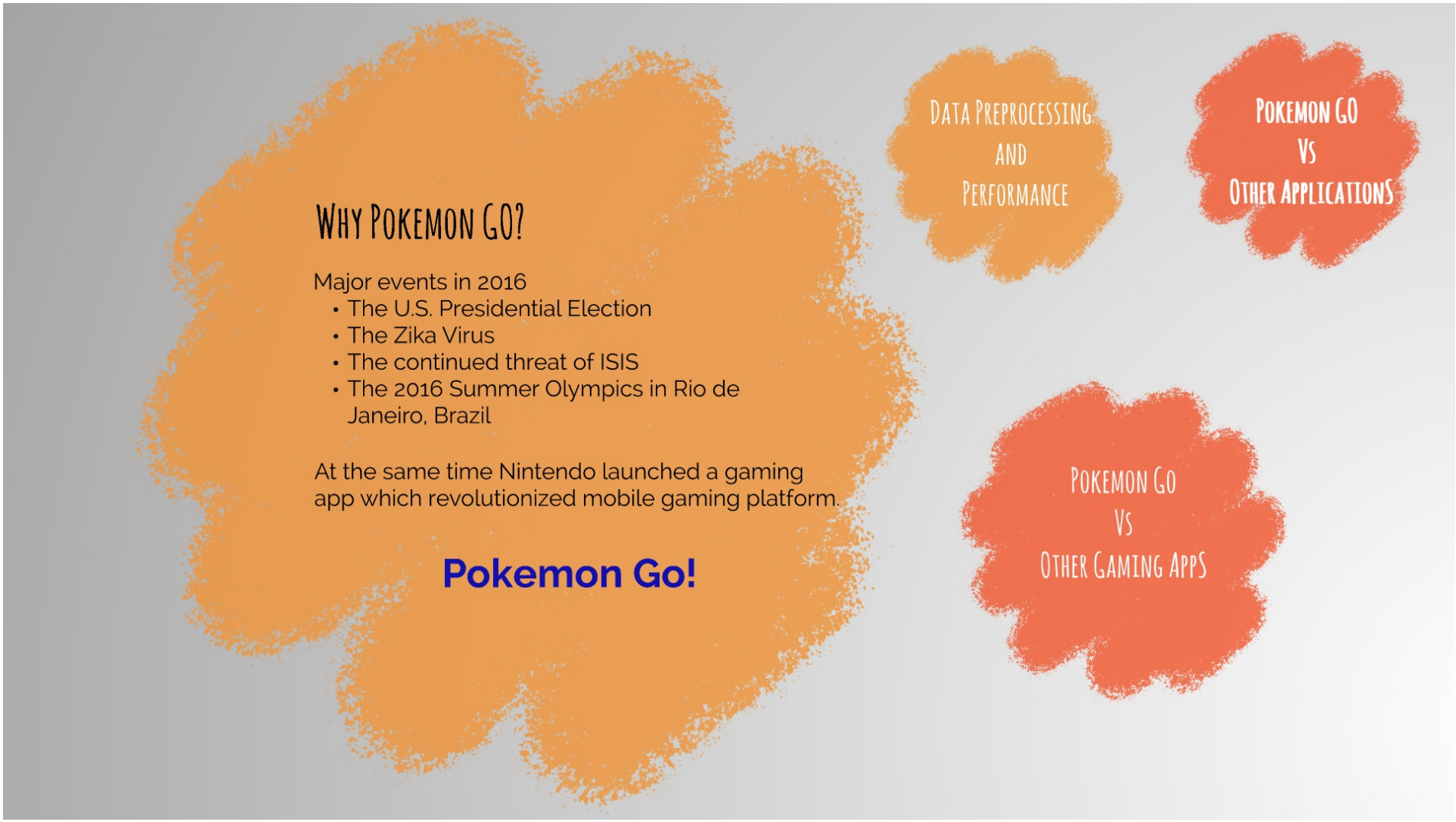
Why Pokemon GO?

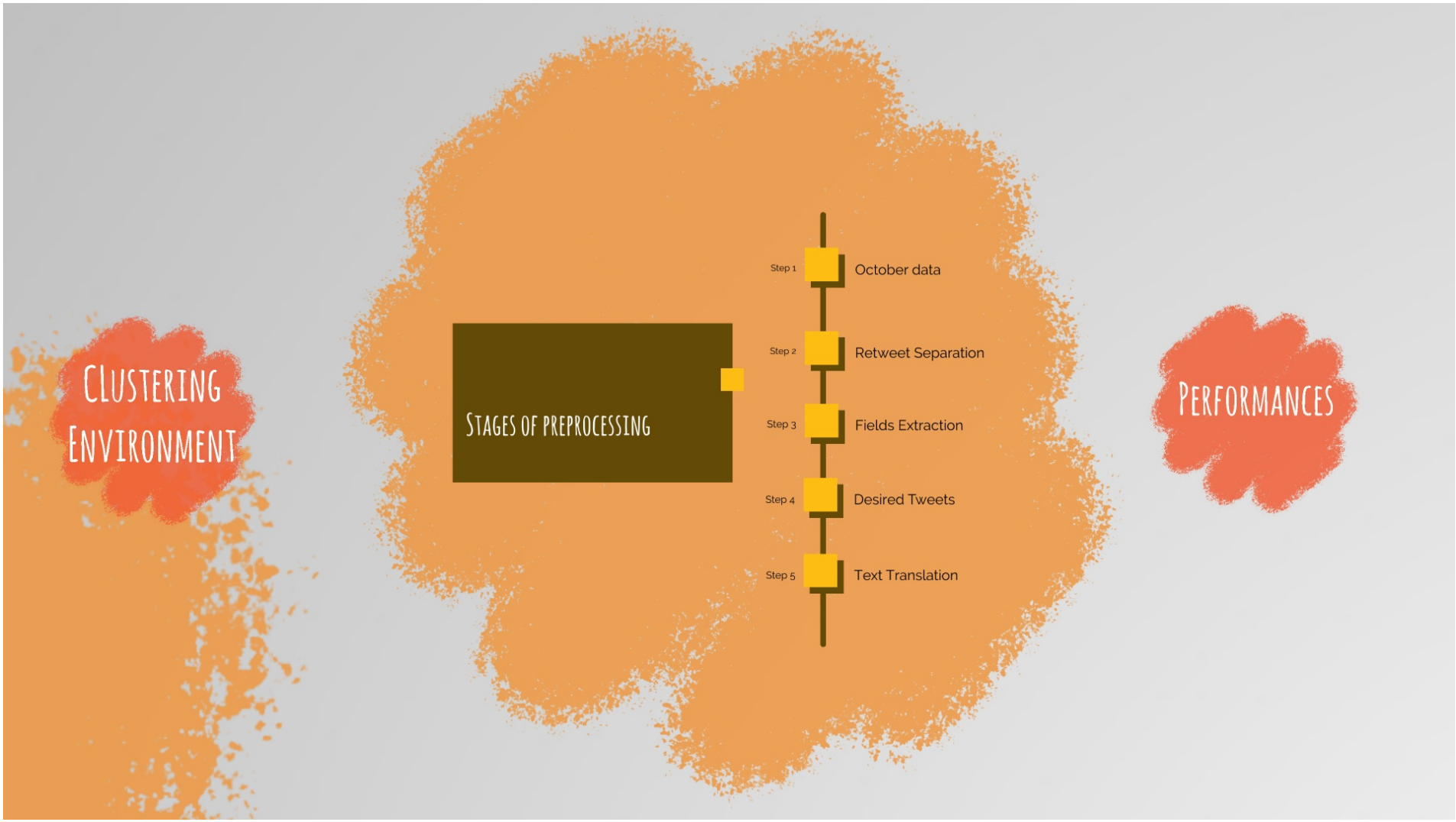
Major events in 2016

- The U.S. Presidential Election
- The Zika Virus
- The continued threat of ISIS
- The 2016 Summer Olympics in Rio de Janeiro, Brazil

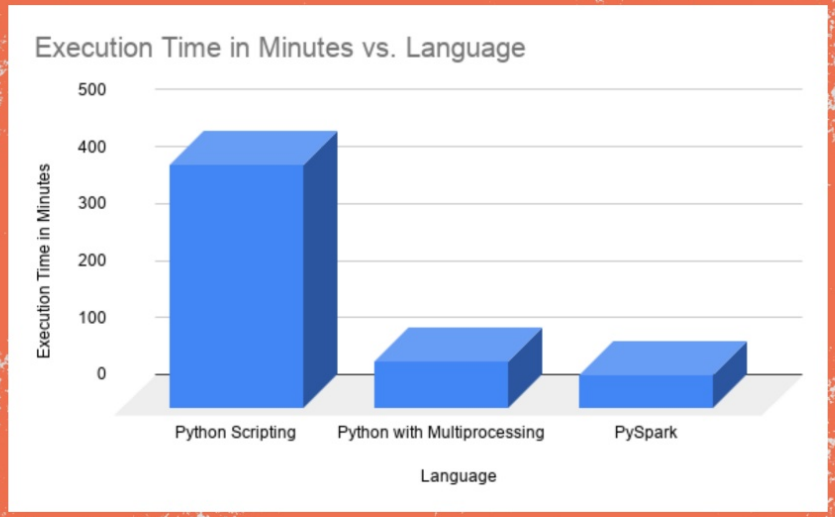
At the same time Nintendo launched a gaming app which revolutionized mobile gaming platform.

Pokemon Go!


Why Pokemon GO?
Major events in 2016
• The U.S. Presidential Election
• The Zika Virus
• The continued threat of ISIS
• The 2016 Summer Olympics in Rio de Janeiro, Brazil
At the same time Nintendo launched a gaming app which revolutionized mobile gaming platform.
Pokemon Go!
Data Preprocessing and Performance
Pokemon GO Vs Other Applications
Pokemon Go Vs Other Gaming Apps


Clustering Environment
Stages of preprocessing
Step 1
October data
Step 2
Retweet Separation
Step 3
Fields Extraction
Step 4
Desired Tweets
Step 5
Text Translation
Performances

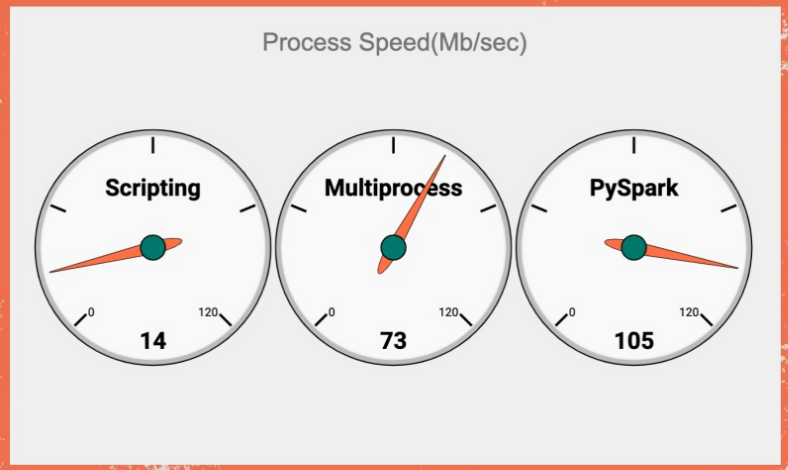
Execution Times



Descriptions

Speeds

Speed Comparisons




System Specifications
We used 8 core, 16GB RAM Linux machine for running the python script.
For Python Multiprocessing we used 8 core, 16GB RAM Linux Machine.
For PySpark, We used 3 worker nodes with 48 Cores in Azure HDInsights.

AzTK - Azure Toolkit

Azure-HDinsight

```
-bash: ztk: command not found
(bdp) (base) Anits-MacBook-Pro:bin agupta$ aztk spark cluster get --id mycluster

Cluster         mycluster
------------------------------------------
State:          steady
Node Size:      standard_d12_v2
Created:        2019-11-08 04:16:47
Nodes:          5
| Dedicated:    0
| Low priority: 5

|              Nodes                |       State        |       IP:Port       | Dedicated  |  Master  |
|-----------------------------------|--------------------|---------------------|------------|----------|
|tvmps_09ea51d69f8f049d29b636d44ffc07dc168bf297d33efc80c987fa414e61aa88_p| waitingforstarttask | 52.188.221.144:50004 |            |          |
|tvmps_35f5bd615f39dd92df84c188bc7bd2a3c6ad2d06696a4275d2c3411fbdab3d0b_p| waitingforstarttask | 52.188.221.144:50001 |            |          |
|tvmps_b41c5f09cb78de645aa7961285507ed1fec7f37793ed7e972030350ba45d8b78_p| waitingforstarttask | 52.188.221.144:50002 |            |          |
|tvmps_ba30c2b02c0e8bc902b3bad8379cde8f4adb8fc7d328b08bd3fef70db86be8c2_p| waitingforstarttask | 52.188.221.144:50000 |            |          |
|tvmps_bd53a926034cc5e6fd0376260d0471ea6c36e16b18bc592dae0f8b24bb5b64d_p| waitingforstarttask | 52.188.221.144:50003 |            |          |
```

Spark Master at spark://10.0.0.5:7077

URL: spark://10.0.0.5:7077
REST URL: spark://10.0.0.5:6066 *(cluster mode)*
Alive Workers: 5
Cores in use: 20 Total, 0 Used
Memory in use: 132.3 GB Total, 0.0 B Used
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (5)

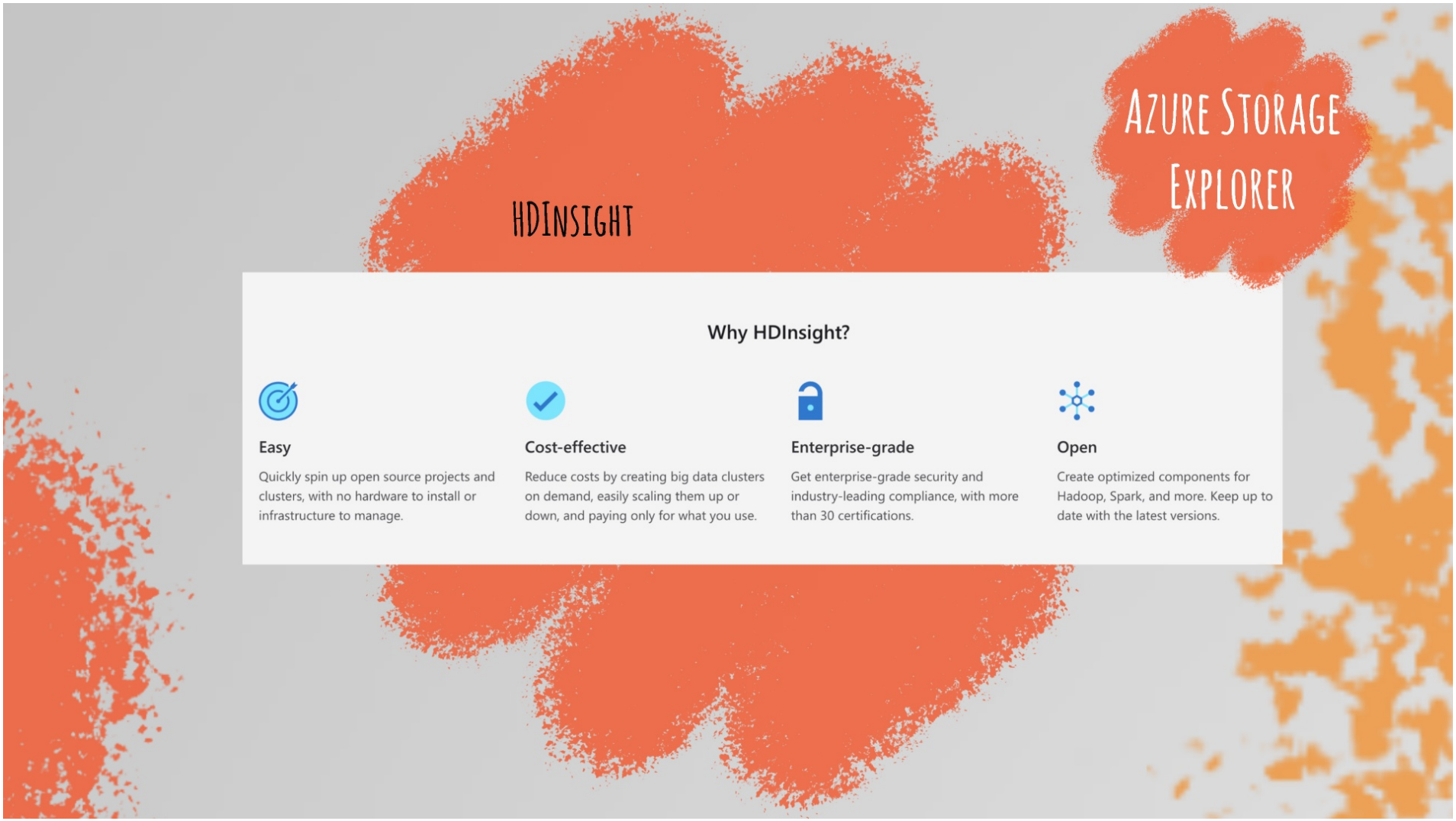
Worker Id	Address	State	Cores	Memory
worker-20191108051128-10.0.0.5-45225	10.0.0.5:45225	ALIVE	4 (0 Used)	26.5 GB (0.0 B Used)
worker-20191108051138-10.0.0.6-46661	10.0.0.6:46661	ALIVE	4 (0 Used)	26.5 GB (0.0 B Used)
worker-20191108051148-10.0.0.4-35951	10.0.0.4:35951	ALIVE	4 (0 Used)	26.5 GB (0.0 B Used)
worker-20191108051152-10.0.0.7-44897	10.0.0.7:44897	ALIVE	4 (0 Used)	26.5 GB (0.0 B Used)
worker-20191108051223-10.0.0.8-35165	10.0.0.8:35165	ALIVE	4 (0 Used)	26.5 GB (0.0 B Used)

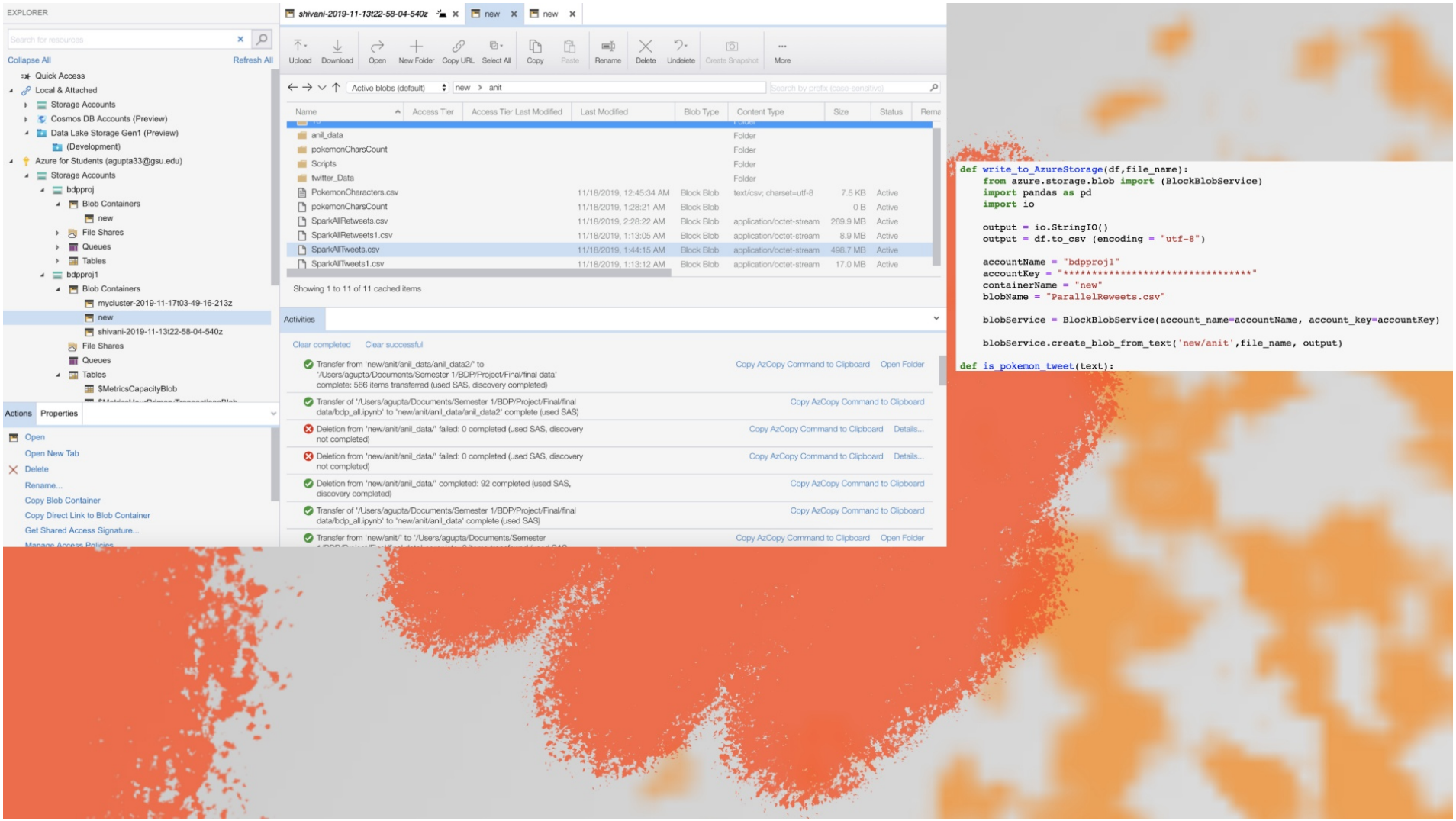
Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration

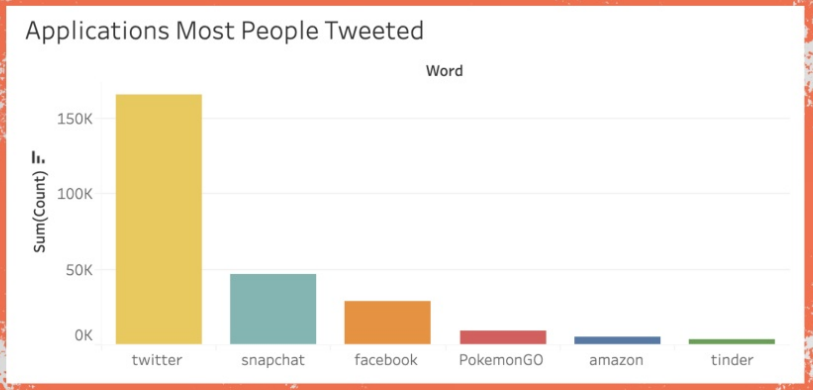
Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration


HDInsight
Azure Storage Explorer
Why HDInsight?
Easy
Quickly spin up open source projects and clusters, with no hardware to install or infrastructure to manage.
Cost-effective
Reduce costs by creating big data clusters on demand, easily scaling them up or down, and paying only for what you use.
Enterprise-grade
Get enterprise-grade security and industry-leading compliance, with more than 30 certifications.
Open
Create optimized components for Hadoop, Spark, and more. Keep up to date with the latest versions.

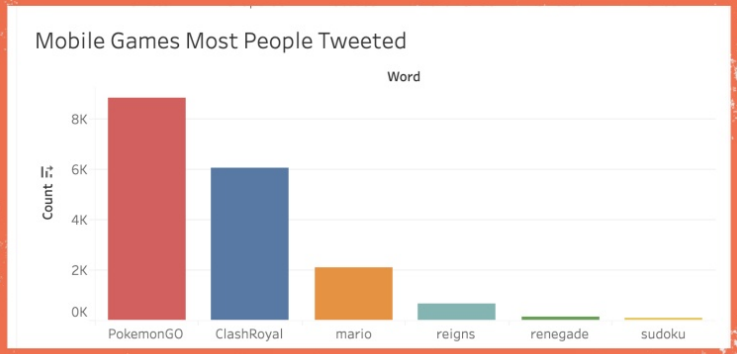

EXPLORER
Search for resources
Collapse All
Refresh All
Quick Access
Local & Attached
Storage Accounts
Cosmos DB Accounts (Preview)
Data Lake Storage Gen1 (Preview)
(Development)
Azure for Students (agupta33@gsu.edu)
Storage Accounts
bdpproj
Blob Containers
new
File Shares
Queues
Tables
bdpproj1
Blob Containers
mycluster-2019-11-17t03-49-16-213z
new
shivani-2019-11-13t22-58-04-540z
File Shares
Queues
Tables
$MetricsCapacityBlob
Actions
Properties
Open
Open New Tab
Delete
Rename...
Copy Blob Container
Copy Direct Link to Blob Container
Get Shared Access Signature...
Manage Access Policies
shivani-2019-11-13t22-58-04-540z
new
new
Upload
Download
Open
New Folder
Copy URL
Select All
Copy
Paste
Rename
Delete
Undelete
Create Snapshot
More
Active blobs (default)
new
anit
Search by prefix (case-sensitive)
Name
Access Tier
Access Tier Last Modified
Last Modified
Blob Type
Content Type
Size
Status
anil_data
pokemonCharsCount
Scripts
twitter_Data
PokemonCharacters.csv
pokemonCharsCount
SparkAllRetweets.csv
SparkAllRetweets1.csv
SparkAllTweets.csv
SparkAllTweets1.csv
Folder
11/18/2019, 12:45:34 AM
11/18/2019, 1:28:21 AM
11/18/2019, 2:28:22 AM
11/18/2019, 1:13:05 AM
11/18/2019, 1:44:15 AM
11/18/2019, 1:13:12 AM
Block Blob
text/csv; charset=utf-8
application/octet-stream
7.5 KB
0 B
269.9 MB
8.9 MB
498.7 MB
17.0 MB
Active
Showing 1 to 11 of 11 cached items
Activities
Clear completed
Clear successful
Transfer from 'new/anit/anil_data/anil_data2/' to '/Users/agupta/Documents/Semester 1/BDP/Project/Final/final data' complete: 566 items transferred (used SAS, discovery completed)
Transfer of '/Users/agupta/Documents/Semester 1/BDP/Project/Final/final data/bdp_all.ipynb' to 'new/anit/anil_data/anil_data2' complete (used SAS)
Deletion from 'new/anit/anil_data/' failed: 0 completed (used SAS, discovery not completed)
Deletion from 'new/anit/anil_data/' failed: 0 completed (used SAS, discovery not completed)
Deletion from 'new/anit/anil_data/' completed: 92 completed (used SAS, discovery completed)
Transfer of '/Users/agupta/Documents/Semester 1/BDP/Project/Final/final data/bdp_all.ipynb' to 'new/anit/anil_data' complete (used SAS)
Transfer from 'new/anit/' to '/Users/agupta/Documents/Semester
Copy AzCopy Command to Clipboard
Open Folder
Details...
def write_to_AzureStorage(df,file_name):
from azure.storage.blob import (BlockBlobService)
import pandas as pd
import io
output = io.StringIO()
output = df.to_csv (encoding = "utf-8")
accountName = "bdpproj1"
accountKey = "********************************"
containerName = "new"
blobName = "ParallelReweets.csv"
blobService = BlockBlobService(account_name=accountName, account_key=accountKey)
blobService.create_blob_from_text('new/anit',file_name, output)
def is_pokemon_tweet(text):

Pokemon GO VS Other Applications



https://public.tableau.com/profile/shivani.bhoite#!/vizhome/Book1777777/Dashboard4

Pokemon Go Vs Gaming Applications



https://public.tableau.com/profile/shivani.bhoite#!/vizhome/Book1777777/Dashboard4

Pokemon Go On Twitter Analysis

Geo-location Analysis

LDA

Why Pokemon GO?

Pokemon GO Facts & Predictions

Sentiment Analysis

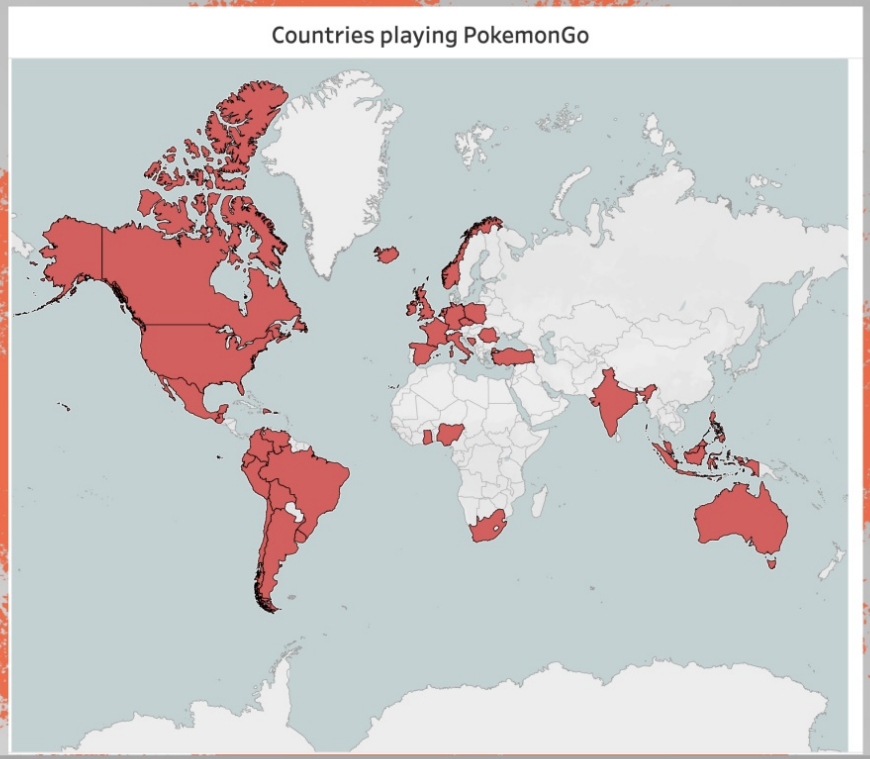
- *Anit Gupta*
- *Anil Ravuru*
- *Shivani Bhoite*
- *Alejandro Benitez*
- *Umaraj Potla*

Geo - Location Analysis

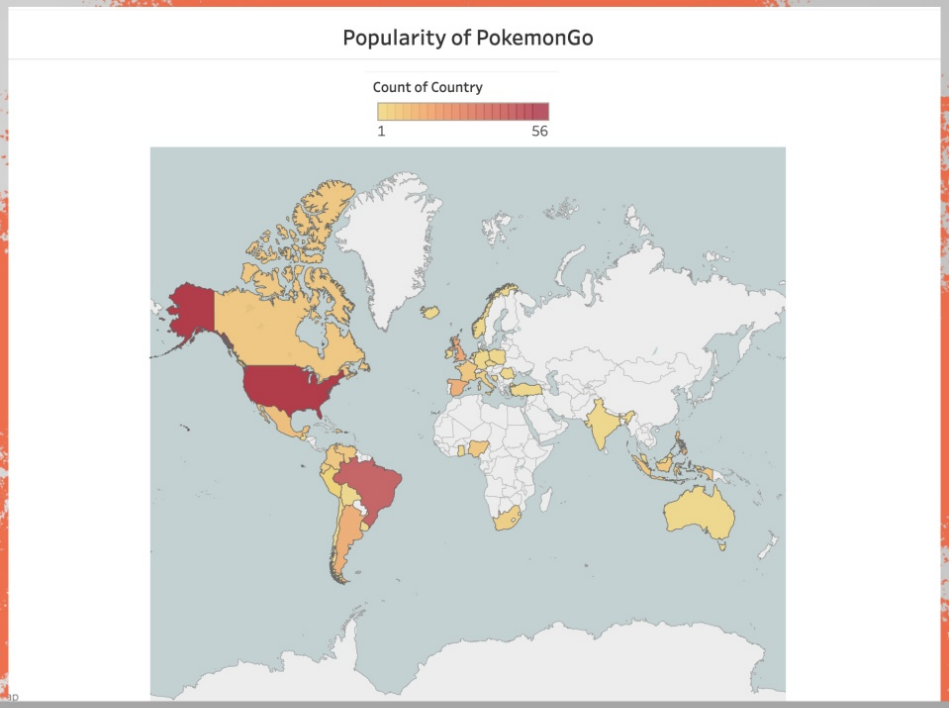
We wanted to analyze the places where people have been using the pokemonGo App.

Countries Playing Pokemon GO

Countries tweeting Most about PokemonGO



https://public.tableau.com/profile/shivani.bhoite#!/vizhome/Book1777777/3-CountriesPlayingPokemonGo



https://public.tableau.com/profile/shivani.bhoite#!/vizhome/Book1777777/3-PopularityofPokemonGo

Pokemon Go On Twitter Analysis

Geo-location Analysis

LDA

Why Pokemon GO?

Pokemon GO Facts & Predictions

Sentiment Analysis

- *Anit Gupta*
- *Anil Ravuru*
- *Shivani Bhoite*
- *Alejandro Benitez*
- *Umaraj Potla*

Sentiment Analysis

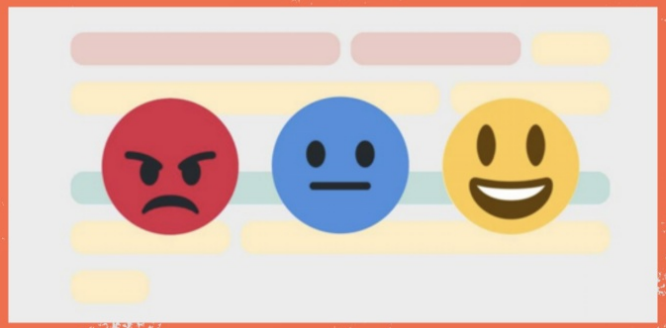
We analyzed how people are reacting towards PokemonGo!!

What's sentiment analysis?

Pokemon Go Sentiments

Sentiment analysis Definition

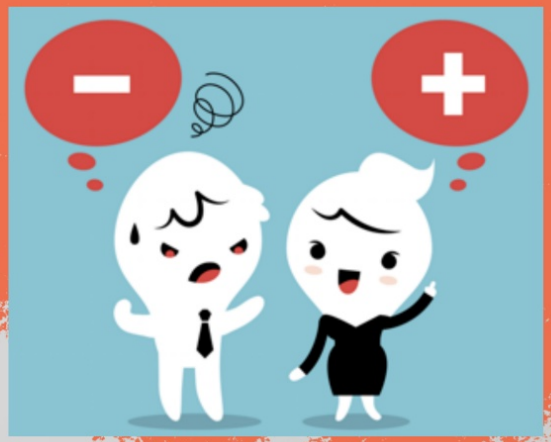
process of determining if a piece of writing is positive / negative / neutral



But Why Sentiment Analysis?

Why?? Sentiment Analysis??

- To understand the scope of PokemonGo better
- Twitter popular microblogging site
- Variety of twitter audiences – common man to celebrities



What have we used ?

What is it?

What have we used ?

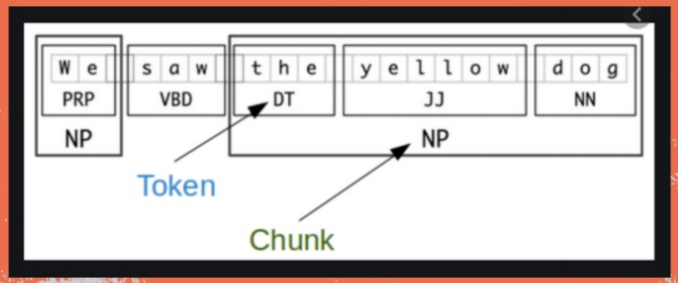
- Natural Language Toolkit (NLTK) - pen source library of tools for natural language processing

- VADER (Valence Aware Dictionary and sEntiment Reasoner)

- Chunking

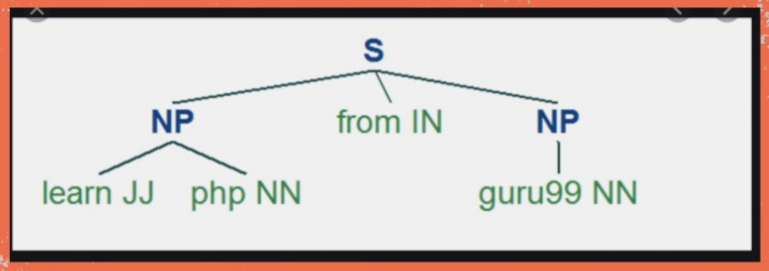
What is Chunking ?

Chunking breaks a text up into user-defined units ('chunks') that contain certain types of words (nouns, adjectives, verbs) or phrases (noun phrases, verb phrases, prepositional phrases).



GRAMMER

Grammar



Reference : https://www.guru99.com/pos-tagging-chunking-nltk.html

WHY?

WHY??

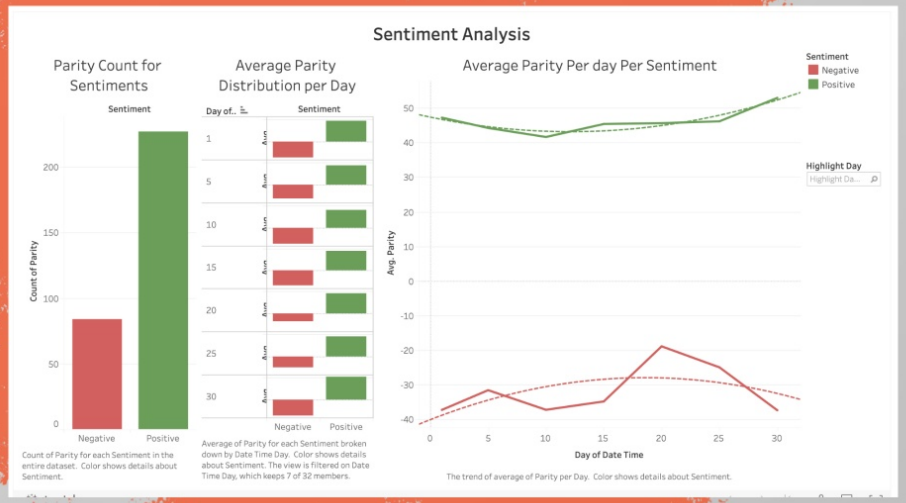
Why VADER ?

- It works exceedingly well on social media type text
- It doesn't require any training data but is constructed from a generalizable, valence-based, human-curated gold standard sentiment lexicon
- It is fast enough

Why Chunking ?

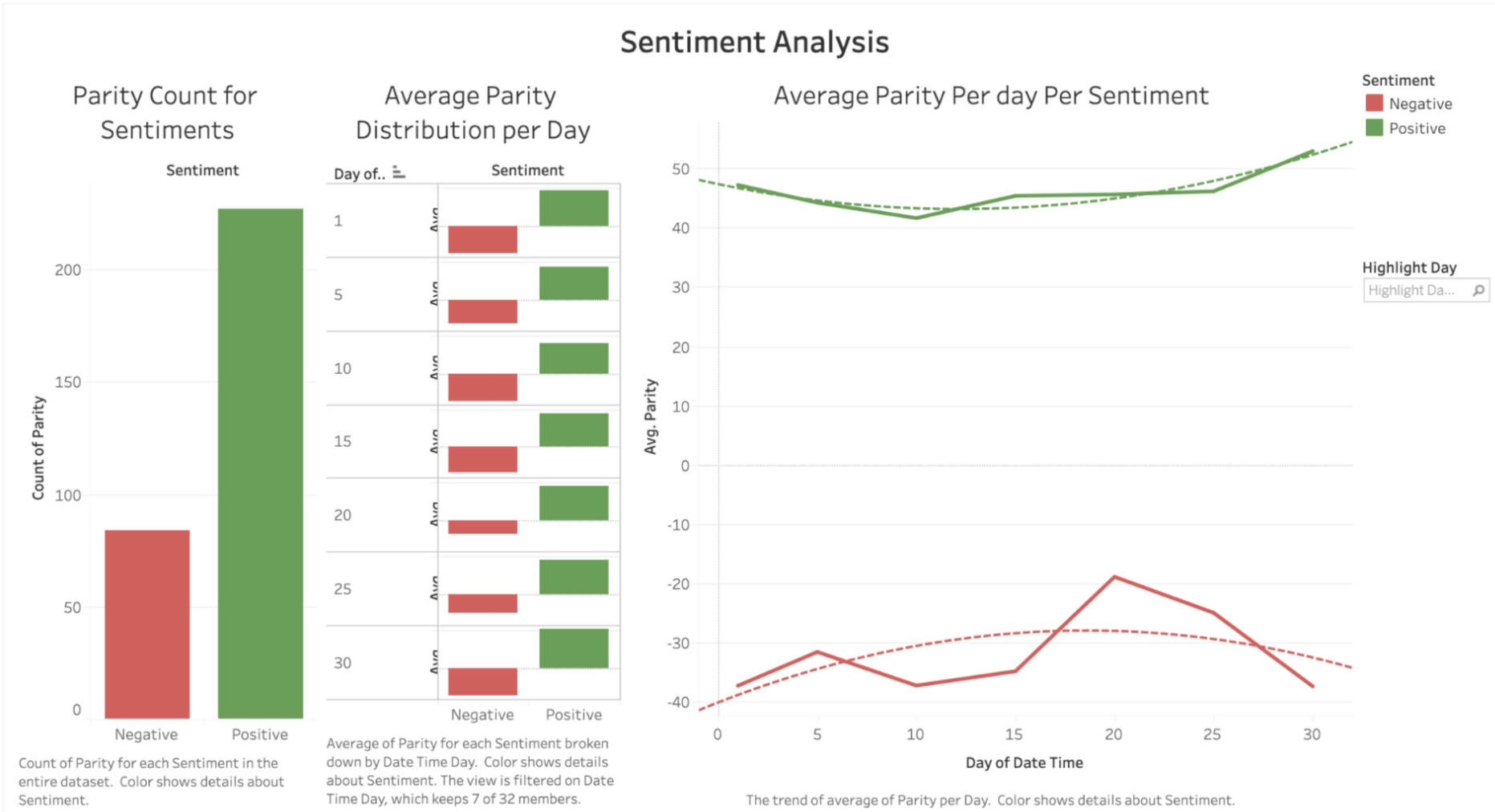
- ability to analyze the text and tag each word with its part of speech.
- allows us to pull out groups of words with set characteristics rather than selecting text by frequency

Pokemon Tweets Sentiments



Lets head to Tableau as its better viewable there

https://public.tableau.com/profile/shivani.bhoite#!/vizhome/Book1777777/Dashboard1


Sentiment Analysis
Parity Count for Sentiments
Sentiment
Count of Parity
200
150
100
50
0
Negative
Positive
Count of Parity for each Sentiment in the entire dataset. Color shows details about Sentiment.
Average Parity Distribution per Day
Day of..
Sentiment
1
5
10
15
20
25
30
Avg
Negative
Positive
Average of Parity for each Sentiment broken down by Date Time Day. Color shows details about Sentiment. The view is filtered on Date Time Day, which keeps 7 of 32 members.
Average Parity Per day Per Sentiment
Sentiment
Negative
Positive
Highlight Day
Highlight Da...
Avg. Parity
50
40
30
20
10
0
-10
-20
-30
-40
0
5
10
15
20
25
30
Day of Date Time
The trend of average of Parity per Day. Color shows details about Sentiment.

Pokemon Go On Twitter Analysis

Geo-location Analysis

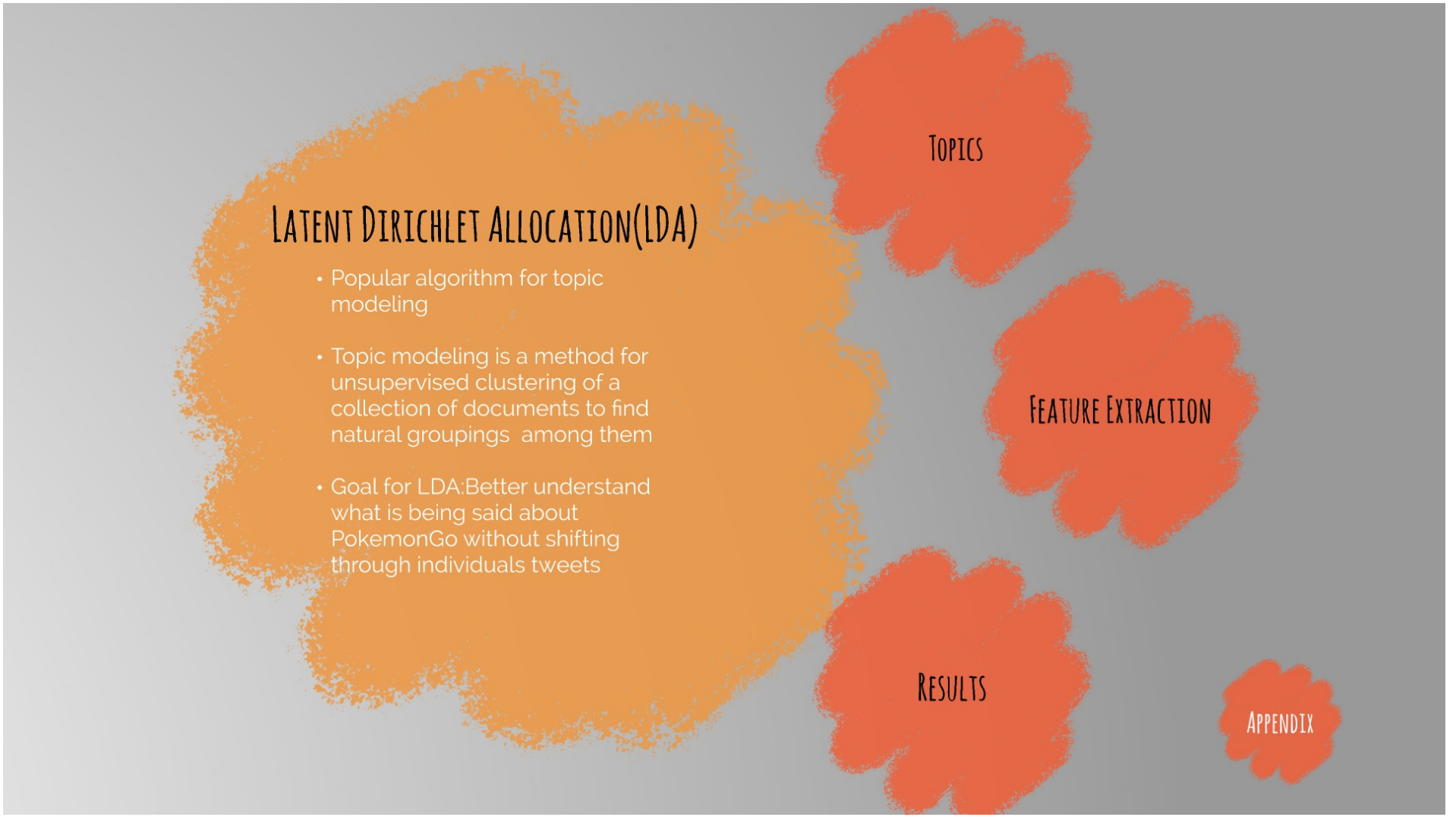
LDA

Why Pokemon GO?

Pokemon GO Facts & Predictions

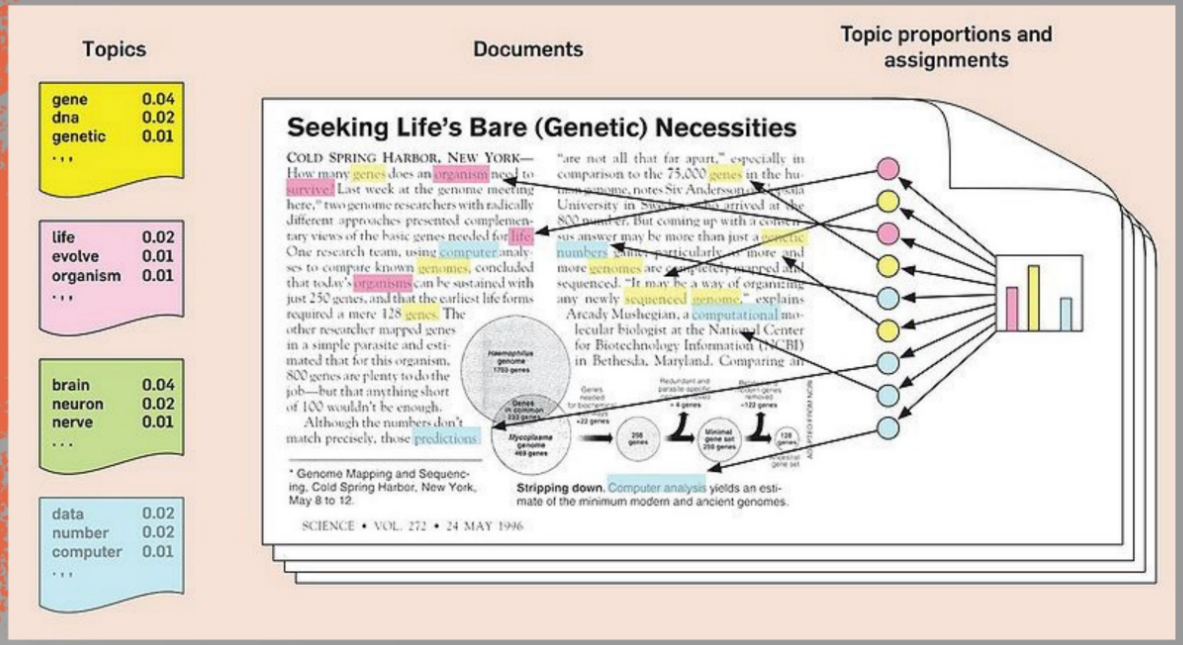
Sentiment Analysis

- *Anit Gupta*
- *Anil Ravuru*
- *Shivani Bhoite*
- *Alejandro Benitez*
- *Umaraj Potla*


Latent Dirichlet Allocation(LDA)
• Popular algorithm for topic modeling
• Topic modeling is a method for unsupervised clustering of a collection of documents to find natural groupings among them
• Goal for LDA:Better understand what is being said about PokemonGo without shifting through individuals tweets
Topics
Feature Extraction
Results
Appendix

LDA Topics

- LDA model treats each document as a mixture of topics
- Each topics is a mixture of words
- In this way, topics are allowed to overlap instead of discrete groups
- Topics are represented by keywords frequent in each topic
- Disadvantage is topics may be difficult interpret into coherent simple ideas

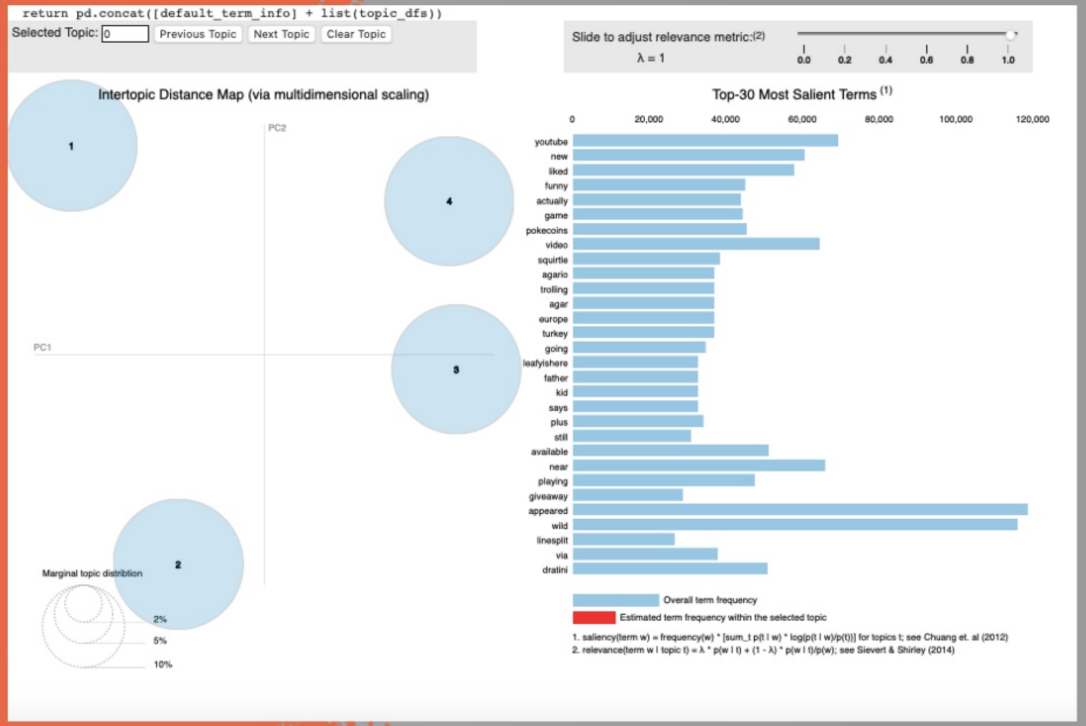


TF-IDF

- Term Frequency-Inverse Document Frequency
- Numeric feature reflecting how important a word is in a document and penalizing it for how common the word is in many documents

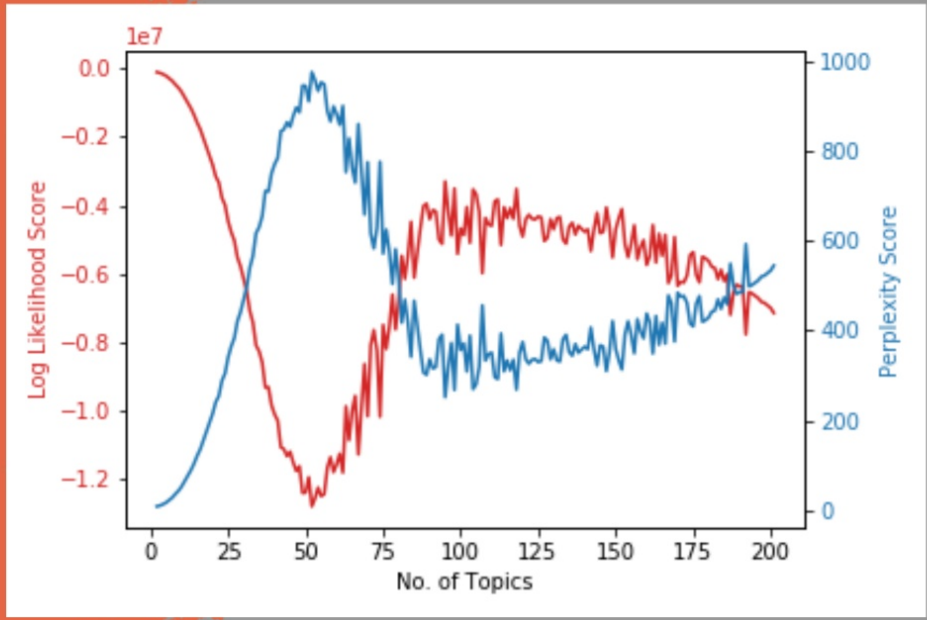
"

Go to notebook



Choosing # of Topics

- Use perplexity and log-likelihood metrics to evaluate separation of topics on subset of data
- Optimal topics is 2
- Increased Topics to 4 to break down topics further into more understandable topic-term distribution.
- More than 5 topics began to significantly overlap in topics



Pokemon Go On Twitter Analysis

Why Pokemon GO?

Geo-location Analysis

LDA

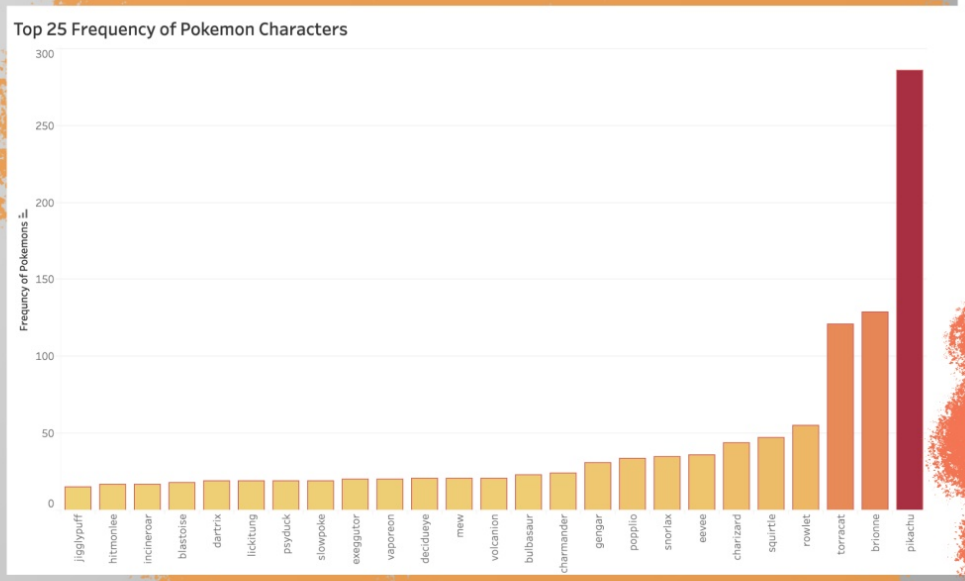
Pokemon GO Facts & Predictions

Sentiment Analysis

- *Anit Gupta*
- *Anil Ravuru*
- *Shivani Bhoite*
- *Alejandro Benitez*
- *Umaraj Potla*

TOP POKEMON CHARACTERS TWEETED

Well most tweeted pokemon character is no more a mystery!!



FUTURE PREDICTIONS ?

How many Tweets should we expect in future ?

We predicted the popularity of the game using tweets counts on hourly-daily basis.

How Did we do?

Time Series Analysis Graph

How did we do?

- Time Series Analysis by State Space Methods 'statespace' library

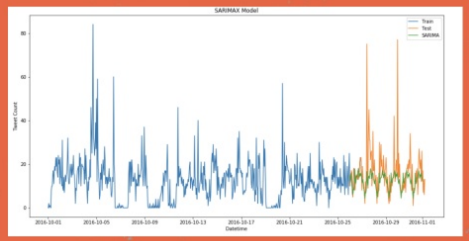
- **SARIMAX** - Seasonal Autoregressive Integrated Moving-Average

- SARIMAX works on a wider range of models by adding the estimation of additive and multiplicative seasonal effects, as well as arbitrary trend polynomials

- One-step-ahead prediction uses the true values of the training sample at each step to predict the next in-sample value.

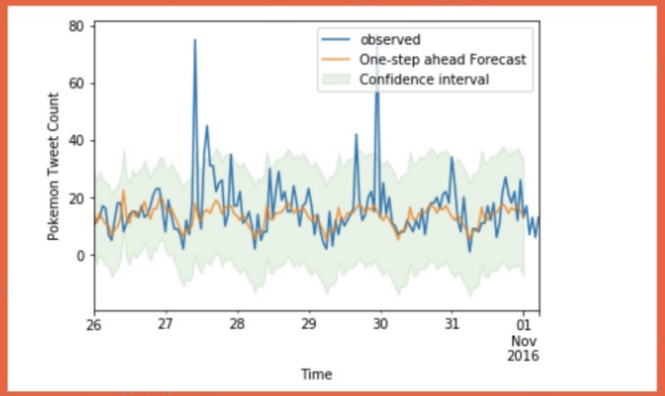
- Dynamic predictions use one-step-ahead prediction up to some point in the dataset (specified by the dynamic argument); after that, the previous predicted values are used in place of the true values for each new predicted element.

Time Series Analysis using SARIMAX Model



More PreDictions

Let's Zoom IN



Thank You!

Pokemon Go On Twitter Analysis

Geo-location Analysis

LDA

Why Pokemon GO?

Pokemon GO Facts & Predictions

Sentiment Analysis

- *Anit Gupta*
- *Anil Ravuru*
- *Shivani Bhoite*
- *Alejandro Benitez*
- *Umaraj Potla*