

Pokemon-Go Data Analysis

Group Members:

Lokesh Kanagala	#800957960
Vidyasagar Kanugula	#800957543
Virinchi Ande	#800970447

Description:

Predicting where a Pokemon can appear in the future, is one of the interesting challenge, on which a Data Scientist can work on. There have been many competitions/challenges on public platforms like Kaggle, where users work on Pokemon-Go datasets to predict Pokemon appearances.

Our project analyzes a Pokemon sightings dataset consisting of roughly 2,93,000 historical appearances of Pokémon's having latitude longitude coordinates, appearedLocalTime, Weather, PokemonType etc.

Source Dataset Size: 402 MB

Source Dataset Link: <https://www.kaggle.com/semioniy/predictemall>

We have developed a machine learning algorithm to predict where Pokemon can appear in future.

Tasks Involved:

K-Means Clustering:

Each record in the dataset consists of a <latitude-longitude> pair that describes the location where a Pokemon is found. But, many of those latitude-longitude pairs represents same location.

For example, The following latitude-longitude pairs < 47.826955, -117.597654>, <47.826888, -117.596629>, <47.827543, -117.597317> represent the same location "Nine Mile Falls, WA USA".



Row(Zone=5, lat=47.826834, long=-117.5978)
 Row(Zone=5, lat=47.826955, long=-117.597654)
 Row(Zone=5, lat=47.826888, long=-117.596629)

To tackle this challenge, we used K-means clustering on latitude and longitude pairs and have clustered the locations into 100 zones.

Naïve-Bayes:

Given the climate, day and time, our model predicts the zone in which a Pokemon is likely to appear.

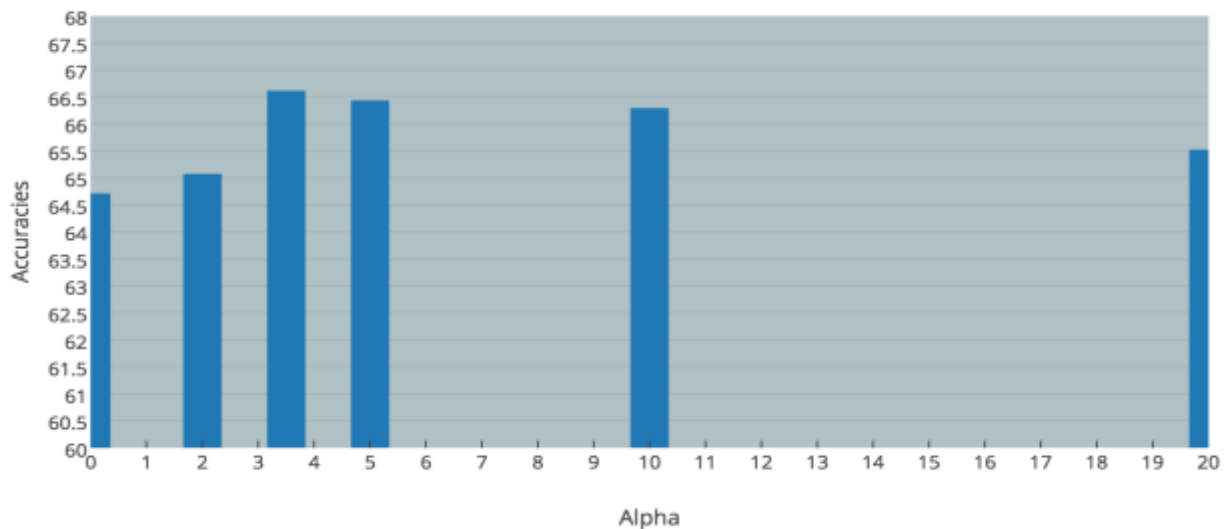
Formulation of the hypothesis:

$$P(\text{zone} \mid \text{time, day}) = P(\text{day} \mid \text{zone}) * P(\text{time} \mid \text{zone}) * P(\text{zone})$$

$$P(\text{day} \mid \text{zone}) = [\text{count}(\text{zone} \cap \text{day}) + \alpha] / [\text{count}(\text{zone}) + \alpha * \text{dayVocabulary}]$$

$$P(\text{climate} \mid \text{zone}) = [\text{count}(\text{zone} \cap \text{climate}) + \alpha] / [\text{count}(\text{zone}) + \alpha * \text{climateVocabulary}]$$

$$P(\text{time} \mid \text{zone}) = [\text{count}(\text{zone} \cap \text{time}) + \alpha] / [\text{count}(\text{zone}) + \alpha * \text{timeVocabulary}]$$



After testing the model with different alpha values, we chose $\alpha=3.5$ since, it gave the highest accuracy of 66.3%.

Output Screenshots:

```

17/05/02 15:20:22 INFO scheduler.TaskSetManager: Finished task 193.0 in stage 950.0 (TID 61263) in 5 ms on mba-hs1.u
17/05/02 15:20:22 INFO scheduler.TaskSetManager: Starting task 186.0 in stage 950.0 (TID 61266, mba-hs4.uncc.edu, pa
17/05/02 15:20:22 INFO scheduler.TaskSetManager: Starting task 198.0 in stage 950.0 (TID 61267, mba-hs1.uncc.edu, pa
17/05/02 15:20:22 INFO scheduler.TaskSetManager: Finished task 184.0 in stage 950.0 (TID 61264) in 5 ms on mba-hs4.u
17/05/02 15:20:22 INFO scheduler.TaskSetManager: Finished task 195.0 in stage 950.0 (TID 61265) in 5 ms on mba-hs1.u
17/05/02 15:20:22 INFO scheduler.TaskSetManager: Starting task 188.0 in stage 950.0 (TID 61268, mba-hs4.uncc.edu, pa
17/05/02 15:20:22 INFO scheduler.TaskSetManager: Finished task 186.0 in stage 950.0 (TID 61266) in 5 ms on mba-hs4.i
17/05/02 15:20:22 INFO scheduler.TaskSetManager: Finished task 198.0 in stage 950.0 (TID 61267) in 5 ms on mba-hs1.i
17/05/02 15:20:22 INFO scheduler.TaskSetManager: Starting task 189.0 in stage 950.0 (TID 61269, mba-hs4.uncc.edu, pa
17/05/02 15:20:22 INFO scheduler.TaskSetManager: Finished task 188.0 in stage 950.0 (TID 61268) in 5 ms on mba-hs4.i
17/05/02 15:20:22 INFO scheduler.TaskSetManager: Starting task 191.0 in stage 950.0 (TID 61270, mba-hs4.uncc.edu, pa
17/05/02 15:20:22 INFO scheduler.TaskSetManager: Finished task 189.0 in stage 950.0 (TID 61269) in 4 ms on mba-hs4.i
17/05/02 15:20:22 INFO scheduler.TaskSetManager: Starting task 194.0 in stage 950.0 (TID 61271, mba-hs4.uncc.edu, pa
17/05/02 15:20:22 INFO scheduler.TaskSetManager: Finished task 191.0 in stage 950.0 (TID 61270) in 4 ms on mba-hs4.i
17/05/02 15:20:22 INFO scheduler.TaskSetManager: Starting task 196.0 in stage 950.0 (TID 61272, mba-hs4.uncc.edu, pa
17/05/02 15:20:22 INFO scheduler.TaskSetManager: Finished task 194.0 in stage 950.0 (TID 61271) in 5 ms on mba-hs4.i
17/05/02 15:20:22 INFO scheduler.TaskSetManager: Starting task 199.0 in stage 950.0 (TID 61273, mba-hs4.uncc.edu, pa
17/05/02 15:20:22 INFO scheduler.TaskSetManager: Finished task 196.0 in stage 950.0 (TID 61272) in 5 ms on mba-hs4.i
17/05/02 15:20:22 INFO scheduler.TaskSetManager: Finished task 199.0 in stage 950.0 (TID 61273) in 5 ms on mba-hs4.i
17/05/02 15:20:22 INFO cluster.YarnScheduler: Removed TaskSet 950.0, whose tasks have all completed, from pool
17/05/02 15:20:22 INFO scheduler.DAGScheduler: ResultStage 950 (collect at /users/lkanagal/projfiles/naivebayes.py:
17/05/02 15:20:22 INFO scheduler.DAGScheduler: Job 323 finished: collect at /users/lkanagal/projfiles/naivebayes.py
The Zone in which you can find a Pokemon is 44
Top five zones where you can find a pokemon [44, 66, 74, 83, 85]
17/05/02 15:20:23 INFO spark.SparkContext: Invoking stop() from shutdown hook
17/05/02 15:20:23 INFO ui.SparkUI: Stopped Spark web UI at http://192.168.150.204:4043
17/05/02 15:20:23 INFO cluster.YarnClientSchedulerBackend: Interrupting monitor thread
17/05/02 15:20:23 INFO cluster.YarnClientSchedulerBackend: Shutting down all executors
17/05/02 15:20:23 INFO cluster.YarnClientSchedulerBackend: Asking each executor to shut down
17/05/02 15:20:23 INFO cluster.YarnClientSchedulerBackend: Stopped
17/05/02 15:20:23 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
17/05/02 15:20:23 INFO storage.MemoryStore: MemoryStore cleared
17/05/02 15:20:23 INFO storage.BlockManager: BlockManager stopped
17/05/02 15:20:23 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
17/05/02 15:20:23 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator
17/05/02 15:20:23 INFO spark.SparkContext: Successfully stopped SparkContext
17/05/02 15:20:23 INFO util.ShutdownHookManager: Shutdown hook called
17/05/02 15:20:23 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-0aa85c50-b288-488c-9f5a-21792b5326b
d006
17/05/02 15:20:23 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-0aa85c50-b288-488c-9f5a-21792b5326b

```

Input: Filename, Time, Weekday, Climate

Output:

The Zones in which you can find a Pokemon is 44, 67, 74, 83, 85

Member Contributions:

Lokesh Kanagala: K-Means , Naïve-Bayes.

Vidyasagar Kanugula: K-Means, Naïve-Bayes.

Virinchi Ande: Data Preprocessing, K-Means.