Predicting the Popularity of Reddit Post

Nikisadat Abbasian | Narmin Orujova

26/04/2021

Outline

- Use Case
- Related Work
- Dataset
- Implementation
- Results
- References

Use Case

- Reddit is a popular social news website, where users post links, text or images which other users can up vote or down vote.
- Companies of today's world invest a lot of money in online marketing to boost their revenues. Hence, identifying the digital content that will become popular becomes a matter of foremost importance.
- Data about user reactions need to be analyzed hence companies can adapt and publish posts based on the interests of their users.

Datasets

Kaggle:

- Dataset of Reddit Comments from May 2015 to 2019
- 15 GB - 22 attributes

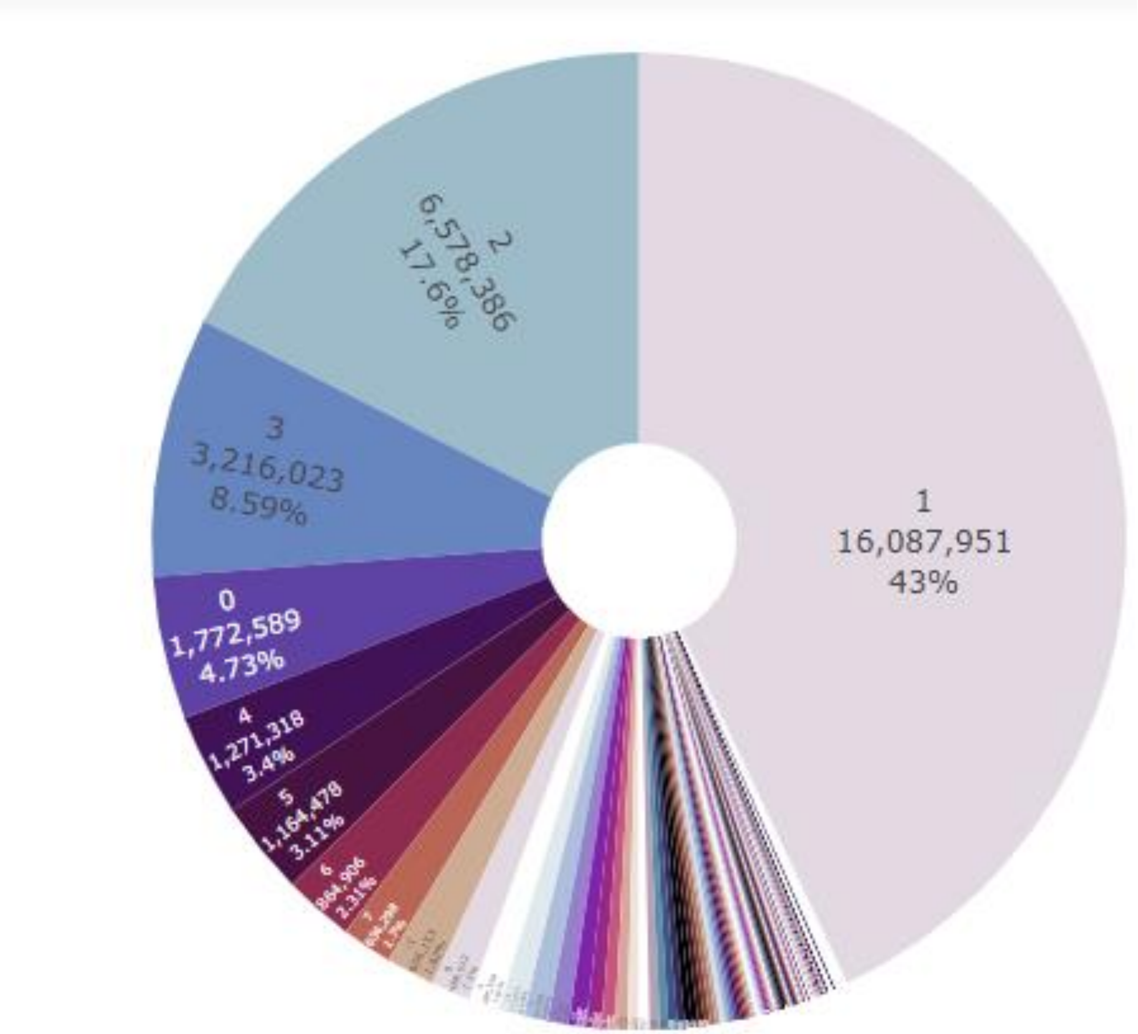
PushshiftAPI:

- Submissions Dataset for last month (5 GB)
- Comments Dataset for the last 6 months of 10 most engaged subreddits (2.5 GB)

Implementation

- Feature Exploration

  Challenges: Not being able to visualize
- Preprocessing with Map Reduce
- Feature Selection
- Baseline Model Implementation with Map Reduce
- String Indexer -Vector Assembler -Vector Indixer Pipeline
- Spark MLlib Implementation of Decision Tree and Random Forest



Results

- Decision Tree and Random Forest Models
- Accuracy Metric: RMSE

Decision Tree:

- **Time** 21/04/25 01:40:55 - 21/04/25 04:30:02 **(2h50m)**
- Root Mean Squared Error (**RMSE**) on test data = 48.8352

Random Forest:

- **Time** 21/04/25 10:33:22 - 21/04/25 15:08:26 **(4h35m)**
- Root Mean Squared Error (**RMSE**) on test data = 48.5151

References

- https://www.reddit.com/r/datasets/comments/5b56my/where_can_i_get_a_large_archive_of_reddit_posts/
- https://www.datasciencemadesimple.com/get-day-of-month-day-of-year-day-of-week-from-date-in-pyspark/
- https://www.analyticsvidhya.com/blog/2019/11/build-machine-learning-pipelines-pyspark/
- https://spark.apache.org/docs/latest/mllib-decision-tree.html
- https://spark.apache.org/docs/latest/mllib-ensembles.html#random-forests