Predicting the Popularity of Reddit Post

Nikisadat Abbasian | Narmin Orujova

23/03/2021

Outline

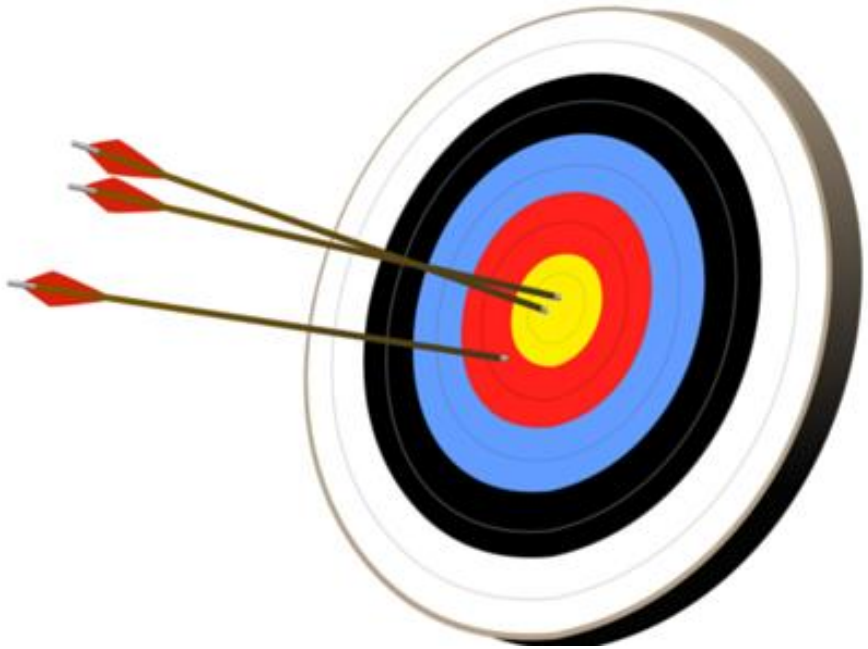
- Use Case
- Objectives
- Dataset
- Proposed Workflow
- References

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.

Objectives

- The focus is to analyze how different factors play a role in predicting how popular a post will become. These factors are the content of the post and its sentiment, the Subreddit of the post, author of the post, and the time of day the post was created.
- We are going to use python's MapReduce framework for cleaning and profiling the data and Hive to query the dataset interactively and derive some information about it.
- We are going to implement a simple model to establish a baseline. This model will always predict the mean number of upvotes.
- Next, we are going to predict the number of upvotes by training the ML Models, such as Decision Tree and Random Forest and compare it to baseline.

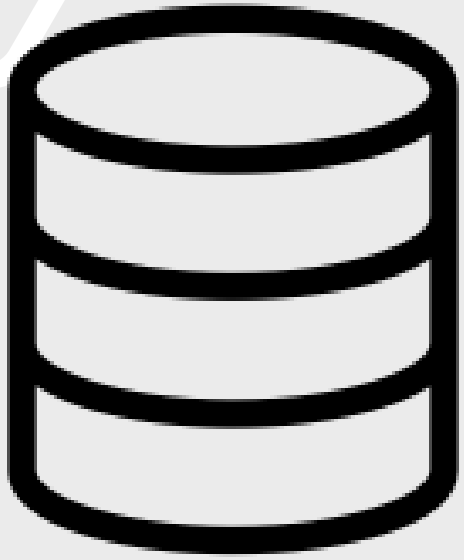


Dataset

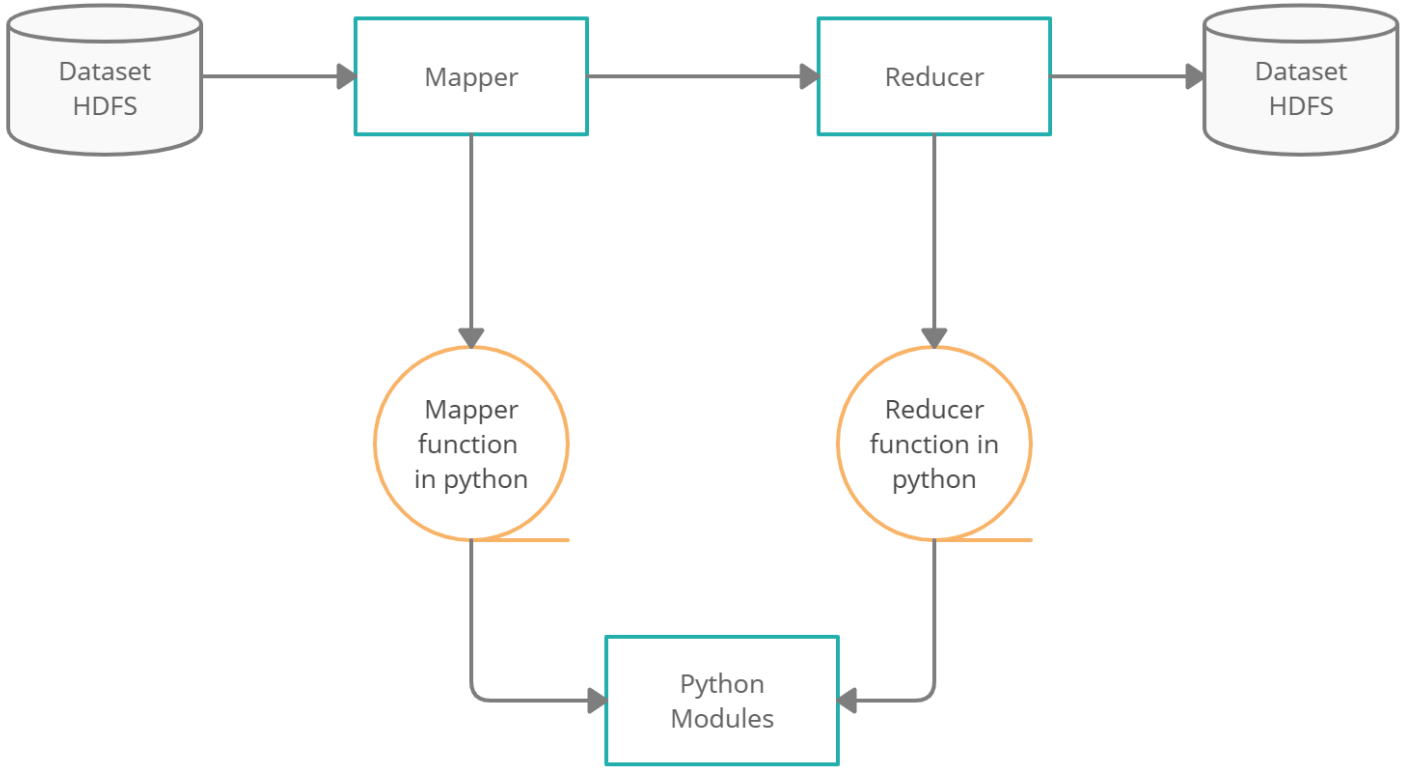
Source: Kaggle (https://www.kaggle.com/reddit/reddit-comments-may-2015)

Description: ~1.7 billion of Reddit's publicly available comments. The full dataset is an unwieldy 1+ terabyte uncompressed, so Kaggle hosts a small portion of the comments (20GB). It is a csv file that has 22 fields like subreddit_id, link_id, author, score, retrieved_on, controversiality, parent_id etc.

Target variable: score - number of upvotes



Proposed Workflow



References

https://towardsdatascience.com/predicting-reddit-comment-karma-a8f570b544fc
http://cs229.stanford.edu/proj2012/ZamoshchinSegall-PredictingRedditPostPopularity.pdf
https://github.com/mdylan2/PredictingPopularityofRedditPosts/blob/master/report.pdf
https://github.com/cjhutto/vaderSentiment