Shadowfox
Devashish Nyati 013732342
Sweta Kumari 013708201
Sweety Sojrani 013731783

Medium Stories- Analysis and Predictions

24th March 2019

OVERVIEW

Medium is one of the best places for worlds most insightful writers, masterminds, and storytellers to present the smartest takes on topics that matter. One can always find fresh thinking and unique perspectives on Medium. There are a lot of amazing articles in the field of data science, machine learning and artificial intelligence. The scope of our project is to give insights about a post before the user even publishes it.

GOALS

1. What sort of stories become prominent on Medium?
2. Get insights about stories that are similar to users topic.
3. Predict the tags that can be associated with a post?

SPECIFICATIONS

Data Sets

Name: Medium Articles (with Content)
Link: https://www.kaggle.com/aiswaryaramachandran/medium-articles-with-content
Size: 280k * 50 (approx 2GB)

The data contains all posts tagged AI, Machine Learning, Data Science or Artificial Intelligence on Medium. There are around 280000 articles in this data set. Each article has around 50 different columns associated with it. Some of the columns are:

- *postId*: Unique identifier of the post
- *text*: Content of the post
- *wordCount*: Number of words in the post
- *Title*: Title of the post

- *readingTime*: What is the time taken on average to read this article (it is based on word count)
- *recommends*: Number of Unique users who clapped for a post
- *totalClapCount*: Number of claps associated with the post

Proposed Methodologies

Methodologies for the following:

- What sort of stories become prominent on Medium - We will perform a multivariate regression of views onto multiple independent variables like word count, reading time, total clap count, and others.
- Get insights about stories that are similar to users topic - Identifying the class of the new story will be performed by K Nearest Neighbors algorithm.
- Predict the tags that can be associated with a post - Since categorizing the article body to tags will be multi-label classification Naive Bayes will be implemented.

The algorithms which we will use for the above-mentioned processes might change based on how they perform.

The data which we are using will be divided using 80-20 rule. The 80% data will be used for training and forming models and the remaining 20% will be used for testing and validation.

REFERENCES

Jekaterina Kokatjuhha (Nov 13, 2018) *How to build a data science project from scratch.* Retrieved from https://medium.freecodecamp.org/how-to-build-a-data-science-project-from-scratch-dc4f096a62a1.