How do Fanfiction Authors Find Favorites?

An Analysis of Game of Thrones Fandom

By Apurva Saksena and Ajinkya Sheth

Introduction

The fanfiction community is huge and growing. It's an intricate network of authors, reviewers, and readers contributing to the creation of some form of contemporary culture.

At the University of Washington, we are a group of researchers studying the fanfiction community and exploring the informal learning taking place there.

We were particularly interested in authors who are user-favorites on Fanfiction.Net. When a user favorites an author, there are certain characteristics of the author that the user finds intriguing. It could be that the story is very interesting or the style of writing of the author fascinates the reader. We aim to find which authors have been favorited the most and what factors correlate with a user favoriting an author.

This blog post explores the connection between users (in a particular fandom) and the authors that they have favorited on FanFiction.net. We us a metric to measure this relationship and try to find out how it correlates to other factors such as:

1. Number of stories and chapters published by the authors
2. Number of reviews received for the published stories
3. Total number of words written by the author

4. Number of favorites received

In our analysis below, we have used the PageRank algorithm on authors in the “Game of Thrones” fandom on Fanfiction.net. Each author has at least one favorite author, and we have exploited this detail for our analysis.

Motivation

Both of us are huge Game of Thrones fans. The exciting season 8 finale and the massive popularity of GoT on social media motivated us to explore this fandom. Our current goal is to analyze authors that have been favorited by users, and which features might have earned them favorites in the GoT fandom on Fanfiction.net. This analysis can pave the way for building a recommendation engine for users on Fanfiction.net.

Dataset

Our dataset has been scraped from Fanfiction.net. For the analysis, we used two primary tables - Story and Author_favorites. The ‘Story’ table contains data about the stories - including but not limited to a unique story identifier, user id, fandom id, number of reviews, number of followers, and so on. The ‘Author_favorites’ table contains data about the users and their favorited authors. Because the data in these tables were humongous, we limited our scope to the Game of Thrones fandom. We used a cluster of the data by only retrieving the data that consisted of stories written in the “Game of Thrones” fandom.

The dataset we used was formed by combining the User Favorite table, Fandom table, and Stories table. This gave us a table consisting of User IDs and their Favorited Author IDs, both belonging to the Game of Thrones fandom on Fanfiction.net.

Method and Process

PageRank is a billion-dollar algorithm which made Google what it is. Whilst the most popular application of PageRank is web search, it can be exploited in other areas as well. The web is a gigantic graph interconnected by the weblinks. And PageRank assigns a score of importance by calculating the ‘inlinks’ to a website. In our case, we have considered the dataset of users and their favorited authors as a form of a graph: Many users favorite authors and these users could be authors themselves who have been favorited by other users. Hence, every author will have none, one or more users who favorite them. And thus we can assign a score of *‘connectedness’* to the authors by using PageRank.

A visual representation of the graph is shown below. The blue dot at the center represents a user and the yellow dots represent the favorited users as well as the favorited users who have favorited other favorited users! When there are no out-links, the graph stops traversing.

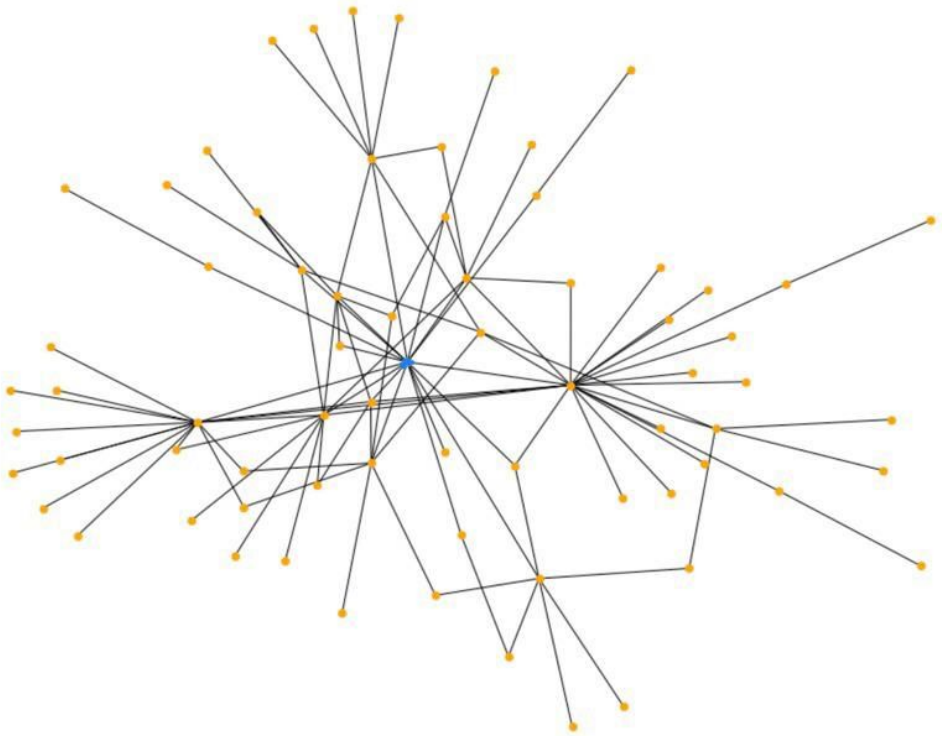


Fig 1. Network of a user (of highest pagerank) and his/her favorite authors. Blue dot represents the user with highest PageRank and yellow dots are favorited authors

This graph shows the '*connectedness*' amongst fanfiction authors. Now we attempt to determine which characteristics (features) have a good correlation with the PageRank score that we obtained. In simple words, we try to find out how closely related the PageRank score is with characteristics such as 'number of reviews', 'total words written by the author' and so on. How can this be done?

A simple way to do this is through Linear Regression. In linear regression, we plot the features against a single response and try to explain the relationship through a straight line. We are conducting our regression analysis by using metrics which depict an author's output (quantity):

1. Total words written by authors
2. Number of stories and chapters published by the author

And those depicting the recognition received (quality) by the author in the form of:

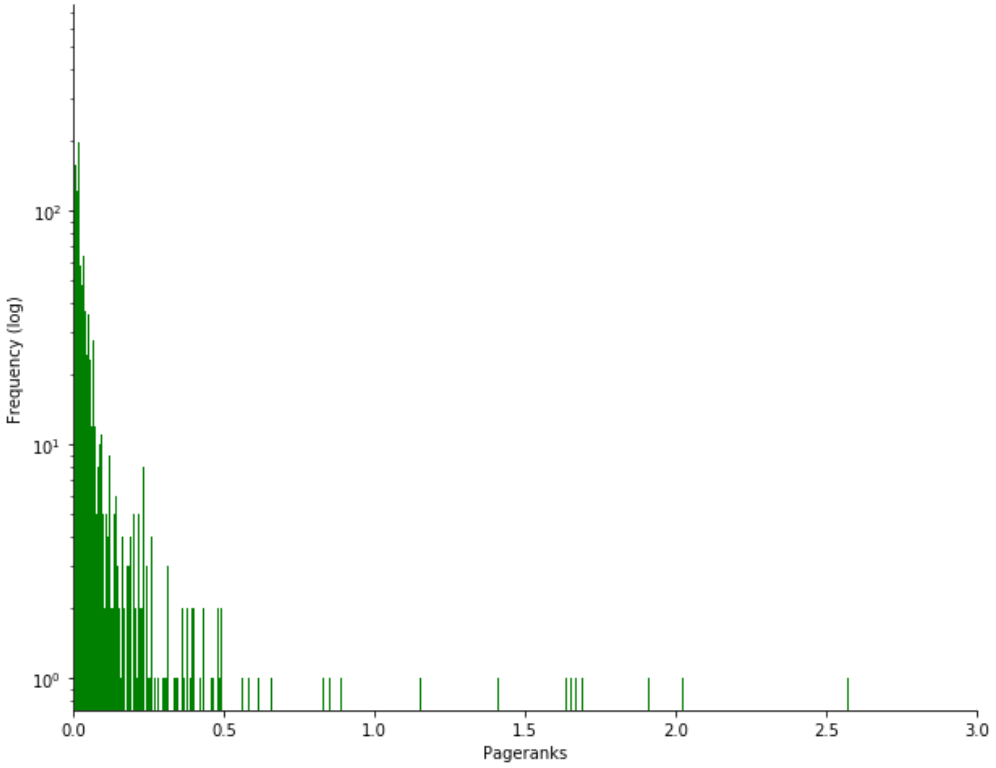
1. Number of reviews received
2. Number of times the author's stories have been favorited

The intuition behind our analysis is to discern if there is any correlation between the PageRank scores which is obtained through network analysis and the above-mentioned metrics.

Findings and Results (WIP)

PageRank Distribution:

The histogram below shows the distribution of PageRank scores. As expected the histogram follows the Power Law which means a small number of items are having high page rank scores while the majority of items is concentrated towards minimum scores.

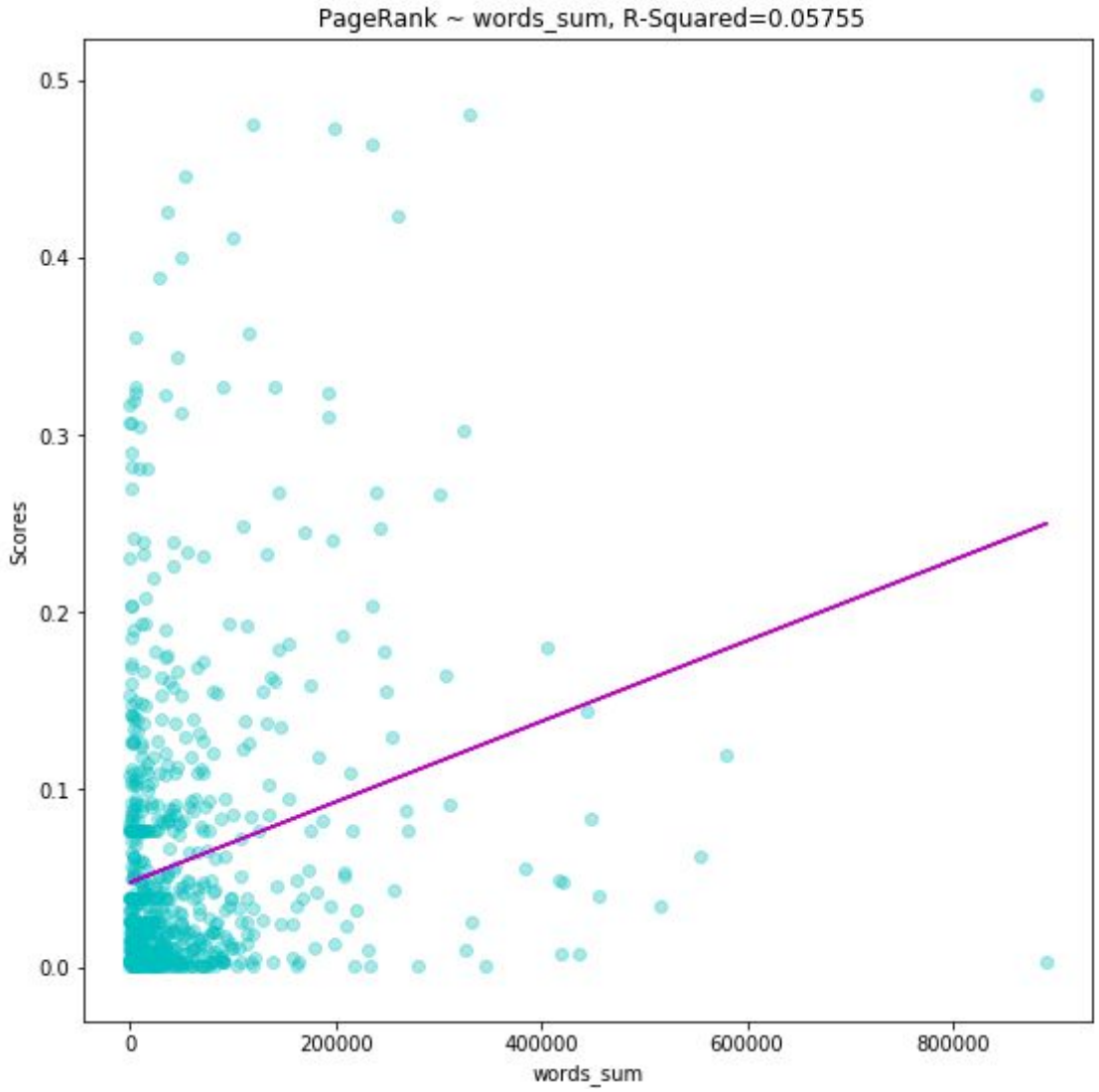


Due to the nature of our distribution, we decided to strip off all the authors having a score above 0.5, as clearly they are outliers and may not represent how the majority of the community behaves. In fact, there is a possibility of authors with high page rank scores skewing our results.

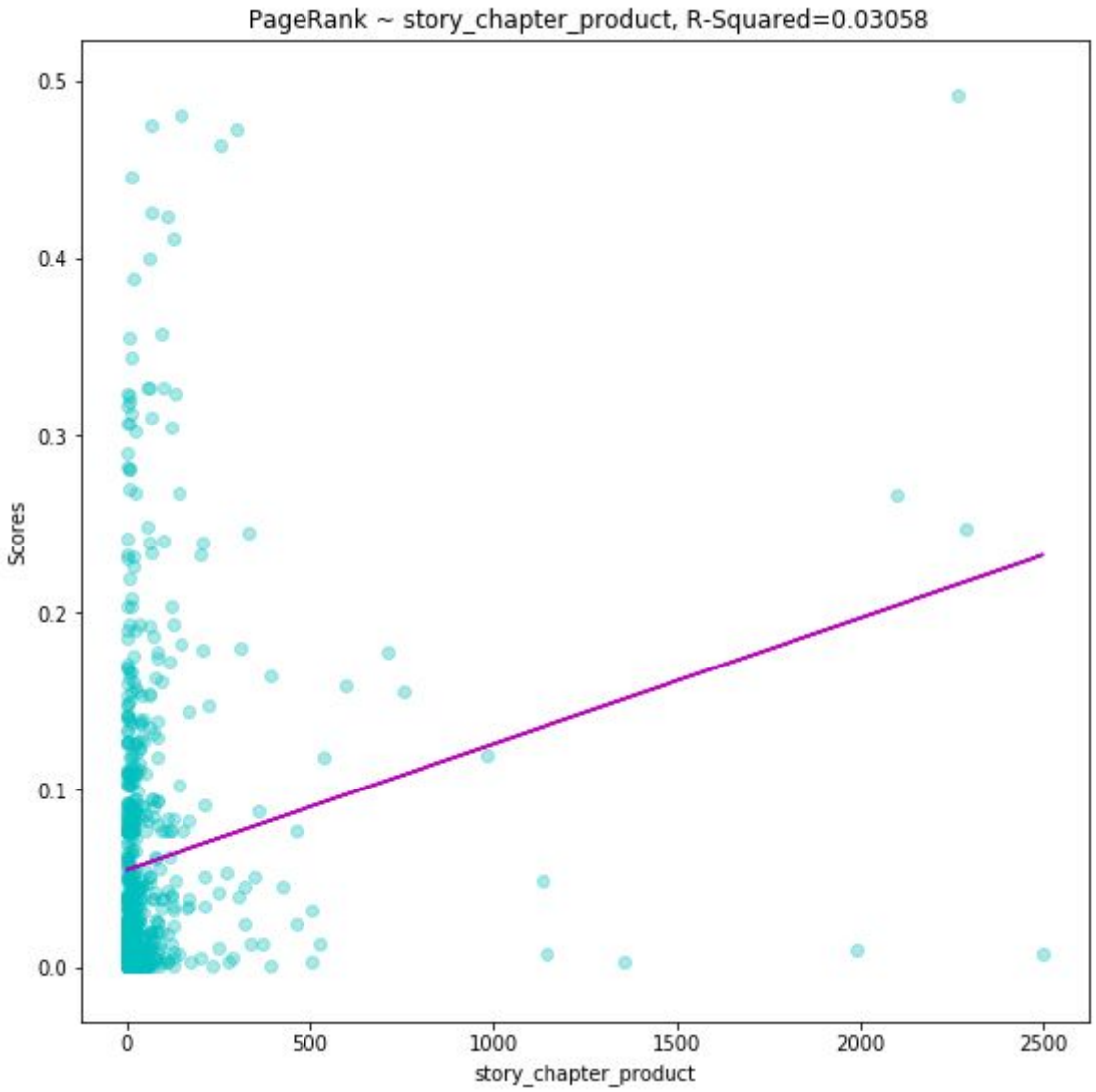
Regression Analysis

The graph below shows a plot of PageRank score against total words and the line denotes the amount of correlation between the two. A positive slope indicates a positive correlation. Please note that even a slight increase in the PageRank makes

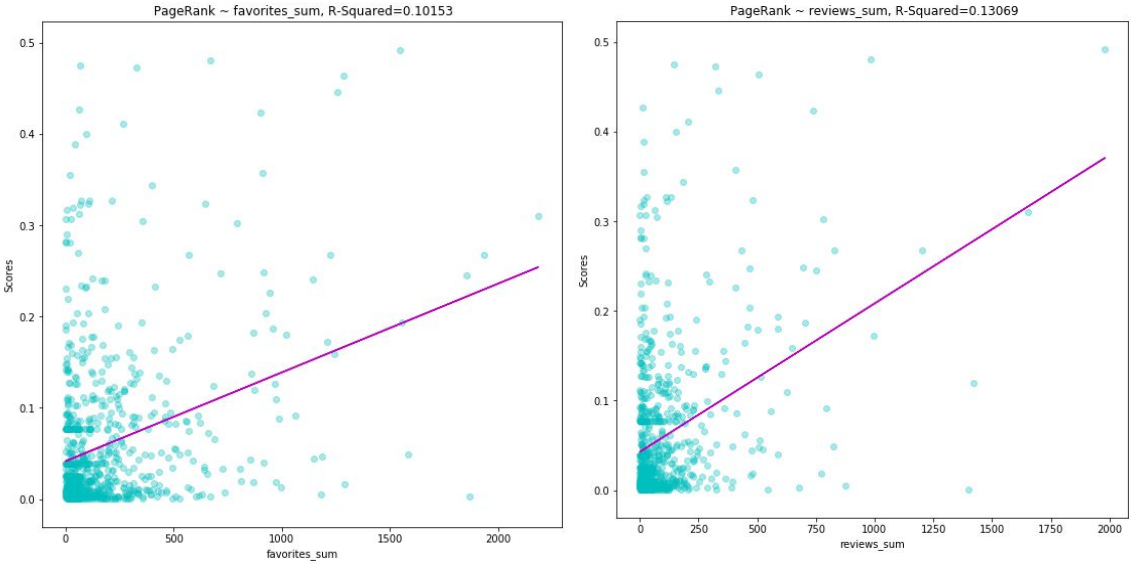
a big difference. We can safely discern that as the authors increase their output, their score improves.



The second metric we used to measure an author's output is Story-Chapter product which is the number of stories multiplied by the number of chapters. The reasoning behind multiplying both is that author adopt different styles for structuring their content. One author may have a story with multiple chapters, another one may write multiple stories with one chapter each. The plot below depicts positive correlation yet again.



Running regression over other variables yields the following output:



Our initial assumptions were correct and the features we selected are all indicators of a good PageRank score. However, which one of these is the best predictor?

Enter p-value. P-value helps to determine the level of significance of our results. In statistics, a p-value <0.05 typically indicates the trend is statistically significant. P-value only helps to infer significance which means all the variables we included in our study are important predictors for the page rank score. What p-value does tell us is how important these variables. To know which feature is better predictor, we use another metric called r-square. R-square helps to know the degree of correlation between two expected output and the actual output. It is conceived in terms of percentage.

The p-values obtained for the above features are as follows:

Features	Reviews	Favorites	Total Words	Story-Chapter Product
p-value	4.06E-27	4.12E-21	2.28E-12	3.79E-07
r-squared	0.13679	0.10153	0.05755	0.03058

Based on our analysis, it's safe to conclude the number of reviews received by the author indicates a higher probability of that author being favorited often.

Conclusion

In our analysis, we used four features, two of which Total Words and Story-Chapter product indicate the output (quantity) of an author while the other two; the number of reviews received and number of times the author's works have been favorited indicate the quality of an author's work. These features have been plotted against the page rank score which indicates the degree of an author's presence in the community. Through data science and statistical analysis, we were able to discern that the quality of works and feedback received by an author is a better indicator than the output.

Future Work

Our analysis can help pave the way for a recommendation engine for new users. This recommendation engine would leverage the PageRank algorithm to recommend authors to a user which he/she would most likely favorite. Just like Google and Amazon recommend products to users, our recommendation engine would suggest authors for users depending on the fandom they like. To build a recommendation engine as effective as Google or Amazon would require tons of optimization and fine-tuning, hence we have kept this as future work for this project.

As for fanfiction enthusiasts ourselves, we want to connect with the community so that they can help us in our analysis. Inputs from the community are always encouraged as this would help us make a better recommendation engine. So please comment on your views on the following questions:

- What would you like to see recommended? We aim to recommend the Authors, but are open to suggestions!
- What parameters do you think would affect the action of a user favoriting an author? Do you think it's just the story or could it be the number of reviews, genre or style of writing? Comment below! Our analysis indicates the number of reviews, however, it will be interesting to see if our analysis is aligning with what the community thinks.
- Which other fandoms do you want us to explore?