Project #1 : Understanding Fan Fiction Authors

Tentative Due Date:

April 5th

High-Level Overview:

Alice enjoys reading and writing Fan Fiction within the online community An Archive of Our Own [AO3]. She has noticed that some authors generate a large and loyal following in online fandoms. She aspires to gain popularity in these communities. She has enlisted your team to help her to understand current trends in what successful authors are doing in Fan Fiction communities.

Deliverables:

The key deliverable is an oral presentation (including presentation slides). This presentation should provide a cohesive narrative that addresses at least one (and possibly several) of the questions below. The oral presentation should be roughly 15-20 minutes long.

In addition to the oral presentation, you should submit any code that you wrote (R, Python, etc). Therefore, please save scripts that you write for various tasks (e.g., reformatting the data). Additionally, I would like you to save any visualizations or preliminary analysis that you do during your investigation even if it doesn't make it into the final presentation.

During the course of the project, your team will also be asked to provide several status reports and updates.

Questions:

The following are general questions that Alice has which can guide your work. She doesn't expect you to answer all of these questions, but hopefully these questions can help you understand what she is interested in.

A) How do you identify a successful fanfiction author? What indications are provided by the online fanfic community (AO3) to tell which authors are successful?

B) Are there high-level factors that are correlated with success? Do successful authors tend to write longer works or shorter works? Do they tend to work across multiple fandoms or within a single fandom? Do they tend to be prolific or create only a small number of works? … etc?

C) If you look at successful authors' work over time, do they tend to gradually write increasingly popular works over time? Or does success tend to manifest as a breakout hit early in their career? Are there any patterns in early-career authors that might correlate with success later on?

D) When you dig deeper into the text, are their stylistic factors that are correlated with success? Do successful authors tend to use a broader vocabulary than their less successful peers? Do successful authors write more (or less) complex sentences? Do they produce works with more (or less) dialogue? … etc?

Text for Analysis:

The best source (in my opinion) of modern fan fiction is An Archive of Our Own https://archiveofourown.org/

1) I found this article helpful in understanding how interaction with An Archive of Our Own can be automated – https://medium.com/nerd-for-tech/mining-fanfics-on-ao3-part-1-data-collection-eac8b5d7a7fa – (Google search can yield additional information on this topic).

IMPORTANT: There are a number of fan fiction stories that include explicit adult content. Explicit content is clearly marked and the site and the site's advanced search functionality easily allows you to search for Fan Fiction with particular "RATING". You are strongly encouraged to analyze Fan Fiction that is written for "General Audience" and/or "Teen".

2) You are welcome to select a single fandom or use several related fandoms. You probably don't want to try and find trends across unrelated fandoms. (That is, it is probably not helpful to select a combination of Little Women fanfic and Star Trek fanfic … However, it might be interesting to look at data across several historical fandoms or several science fiction fandoms)

Note: The project is probably most interesting if you select fan fiction examples from a fandom that at least some of your team members are familiar with.

3) For this project, it is perfectly acceptable to use moderate-sized sample corpus from the archive. (Although please put some careful thought into your sampling methodology). That being said, your code and methodology should be built in such a way that I could easily apply your methodology to a large corpus of data later on (after the project). Therefore, you should use Spark for your initial processing at the start of your data pipeline to make it easy for me to apply your methodology to a larger corpus of text in the future. That is, your general methodology should be to start with Spark and then use other technologies (e.g., R or Python on your laptop) once you have sampled, filtered, or summarized the data.