Lauren Benson
CAPS research paper outline
Draft #1

I.   Introduction

The Hunger Games. The Da Vinci Code. Twilight. Fifty Shades of Grey. These novel titles are

some of the most well-known in the nation, yet the reasoning behind their success is one for

investigation. Take a moment, and come up with a longer list of the most popular contemporary

books you can think of. Do they include titles and genres like the ones I already mentioned? Full

of vampires, werewolves, sexually explicit scenes, and a lack of admirable sentence structure?

My question, upon embarking on my CAPstone journey, is to figure out why exactly the success

of literature appears to be based on the most arbitrary of criteria. I would think that the most

popular and successful books are the ones that are the most well written and perfectly

constructed, with heart-stopping plotlines and exquisitely developed characters. Of course, there

is an abundant of these sort of novels, as there is no lack for talented authors. Why is it then, that

so many of the New York Times best-selling novels are those containing obsessive love stories

about vampires?

For my project, I seek to delve into the path of literature, and what will truly determine a

book's success. I am not the first to ask this question, as there have been a plethora of studies and

surveys conducted trying to extract this secret of success. As is the procedure nowadays,

inevitably a computer must be consulted at some point when attempting to answer widespread,

ambiguous questions. These studies have attempted to quantify vast amounts of data on this

subject, process it, and make conclusions about the reasoning behind this phenomena. Some have

been moderately successful, others less so. The most notable ones have come up with original

algorithms attempting to predict the success of a pre-sold novel. This is hardly a unique concept. Predictors of success are seen all over the market, from music to businesses to advertisements. A song must have x y and z in order to appeal to an audience broad enough to be played on a popular radio station. A billboard must have these colors with this amount of words in this font in order to be noticed by the public. The attempt to analyze and conquer the mind of the public to make a living is almost a field of study on its own. What makes processing and analyzing this sort of data so difficult is the quality of the data itself. How do you quantify thoughts? How do you convert people's desires and oftentimes unconscious opinions into numbers able to be organized and studied?

Luckily, I am not the first person to ask this question. The previously stated studies and surveys pertaining to this topic are integral to my project. The first, and most notable of these studies was one conducted at Stony Brook University. In this study, samples of literature were used to determine specific criteria that makes a book successful. This criteria was then compacted and used to create statistical stylometry that is supposed to be able to predict a specific novel's success once it is released. These statistics, which claim to be 85% accurate, are supposed to analyze a novel's language use and other stylistic criteria in accordance with the desires of the general public. While these statistics are admirable, and upon further examination appear to be accurate, I have identified flaws in this study as well as countless others attempting to do the same. As an individual fiercely loyal to the great classical authors of the literature world (Hemmingway, Dickens, Fitzgerald, Ellison. . .) I oppose the idea that their success can be quantified. No computer or mathematical formula could ever hope to overcome the genius that is the exceptional human mind. Yes, computers can now write A-level essays and replicate

years of work and study in a mere hour, but they have yet to cross the bridge into human genius. Through attempting to replicate these technological studies, by creating my own program meant to statistically analyze the quality of writing, I hope to prove my own program wrong and find instead that there is no such formula to the beauty of the written language.

To begin, I will define what I believe to be the "aesthetics" of quality writing. For my purposes, I am talking about only the elite writers- those who have already mastered written the written language. Grammar rules, spelling, basic word choice and other low-level criteria do not apply to these authors, as they have already surpassed the stepping stones to being good writers. Every book that is currently published and on the market, whether or not a reader believes it to be quality, is written by an accomplished author. Only the elite ever make it to the point of having their work published, so one can safely assume that all published authors have a solid base and know how to effectively write. The market of published literature is the olympics of writing. Publishing a book is a process underrated in its difficulty given the sheer mass of published literature. That being said, the standards of quality writing still wildly vary cross genre, cross audience, and cross culture.

  II.   Overview of past studies (ie. stony brook)
III.   fieldwork explanations
        A. graphs, data sets, surveys
 IV.   Interpretations of results
  V.   Explanation of program
        A. criteria used
        B. decisions made
 VI.   Conclusions vis a vis research question
        A. What truly makes successful literature?
VII.   Conclusion

  I.   Introduction:
        ● overview of popular literature, including the classics as well as pop culture hits.

- ○ Examples: Dickens versus Stephenie Meyer: Why are they equally popular?
  - ○ Has our taste in books become worse over the years?
  - ○ Do we even individually like these books, or are we just influenced by their popularity to be convinced we like them?
  - ○ What makes a book truly good? If it is the quality of writing, then why are such terribly written books so popular?
- ● Research Question: To what extent have recent statistical "trend" studies exploring success rates of novels shifted our definition of "quality literature"?

II. Overview of past studies
- ● Rather than be the first attempt to computationally analyze literature, I plan to use past attempts to create my own conclusions
- ● Explanation of how algorithms have been used to try to quantify the quality of writing
  - ○ i.e. decide on certain criteria that deems a novel to be "successful", and programs the computer to compile all statistics related to such criteria
- ● Stony Brook Study
  - ○ explanation of study
  - ○ First real attempt to use algorithms for this purpose
- ● Other studies
  - ○ how they differ from Stony Brook
  - ○ how they are inspiring my own program
- ● Explanation of which criteria I find the most useful from these studies
  - ○ Which studies do I find useless/lack criterion validity, and which ones I find to be valid

III. Fieldwork Explanations
- ● Begin by talking about surveys
- ● Go through process of how I decided to survey people
  - ○ criteria used, questions asked
  - ○ include samples of some of the literature paragraphs I showed to people and explain why I used those particular samples
- ● Describe how I use survey data to design my program

IV. Interpretations of results
- ● depending on how many surveys I have, and how much data I have, I'll categorize my results based on whatever criteria I end up keeping
- ● discuss outliers, discuss how surveys and studies are valid

- Interpret the actual results
    - how do they respond to my research question?
    - are they actually saying what I hoped they would say?
    - Are the results disappointing or uplifting? (ie, what do people base their acceptance or rejection of literature on?
    - Which results will I decide to keep, or will I use all of them?
    - Will I have to change my survey questions/change my approach if I am not getting the results I need?

V. Explanation of program
- brief description of how writing a basic javascript program is relevant to my project
- compare/contrast my own code to that of the previous studies I analyzed
- Talk about which criteria I decided to use and why
- Is the program successful? Does it prove what I want it to prove?
- Brief description of the statistics I came up with
- Analyze such statistics
- Do these support the data I gathered from my actual surveys? As in, does the computer match what real people are saying?

VI. Conclusions vis a vis research question
- Based on my final data/statistics, what have I learned/concluded?
- Have my opinions changed?
- Have I proven anything that previous studies neglected to show?
- Do my own results negate the previous studies, or support them?
- Describe my actual conclusions: What makes successful literature based on my results, and describe its relevance to real life opinions?

VII. Conclusion
- Explain point of entire project, and why its important?