Analysing the Shape of Stories using Natural Language Processing

Luca Davies

Abstract

This project proposed within is an experimental endeavour aiming to map and graph the emotional and sentimental ups and downs in fictional stories to allow further study such as comparison between different chapters or sections of the text and categorising the text in full.

The project will begin with the development of an application that may conduct the analysis. It will use NLP techniques including Sentiment Analysis and any others necessary to most effectively plot the emotional variations in the text. It will also include a visualiser to display this data on several scales from the whole text, to specific subsections. Results from the project will explore how well the application was able to identify the curve of emotions in stories and if known texts analysed produced the expected data.

1. Introduction

Stories are everywhere, fictional texts saturate the modern world narrative adverts, to films, to books. All fictional texts will include emotions: attached to a situation, a character, an object, anything essentially and these emotions are used to shape how a reader (in the case of books) feels as they read. Some stories are even specifically written to evoke particularly strong feelings in a reader toward the extreme. Writer Kurt Vonnegut suggested that all stories may follow a small number of basic story archetypes based upon how the emotions and overall sentiment in a text plays out during its course. Specifically, he proposed that one could plot on a graph, the emotions in a story against progression through the text. (I.e. x-axis as beginning to end and y-axis as positive to sad).

Vonnegut proposed seven or eight story types in common literature (detailed below). He posed the questions "*Do all stories fit into these types?*" and further *"Are there more types?*". This leads to further questions, asking if the graph or curve of all stories is meaningful in some way, or are there some which are too manic or too flat to hold any valuable information about the story.

This project will endeavour to answers Vonnegut's main questions and to go further and explore further topics. It will investigate the shape of stories as a whole before delving deeper, to finer granularity, looking at curves produced in relevance to a single character, or the graph of course of only a single chapter or section in the text and to test if the chapter itself conforms to Vonnegut's proposal of all stories fitting the aforementioned types. Time depending the project will also investigate if stories can be split into discrete sections of positivity/negativity or otherwise identifiable emotions.

2. Background

Previous projects relating to this project include Kurt Vonnegut's own *rejected* master's thesis from the University of Chicago [1]. This is where his idea of basic story archetype originated from and he has since gone on to lecture around the theory and topic [2]. Videos of his lecture can be found online [3]. Analysis has also been performed on characters and their interactions to produce a graph showing their narrative relations [4].

The Hedonometer Project [5] has essentially performed a very similar study and has graphed a significant catalogue of popular works of fiction.

3. The Proposed Project

3.1 Aims and Objectives

The aim of the project is to investigate “the shape of stories” in the sense proposed by Kurt Vonnegut. It will attempt to categorise stories according to his proposed types:

- Man In Hole
- Boy Meets Girl
- From Bad to Worse
- Which Way Is Up?
- Creation Story
- Old Testament
- New Testament
- Cinderella

The application produced will use existing NLP libraries to analyse the texts using sentiment analysis and similar techniques. It will also visualise this data in a graphical display to allow easier secondary human analysis of the data produced. Major objectives follow:

- Create an application to perform sentiment analysis on a whole text of a story – must track emotional ups and downs of the text and must visualise this data via a graph or similar interface.
- Perform case study: Investigate the shapes produced by a small number of whole texts and attempt to categorise them according to Kurt Vonnegut’s story types
- Perform case study: Investigate difference in data between small sections of the text (e.g. chapters) and the data for the whole text
- Perform case study: Investigate difference in data when it is filtered to include only those data points relating directly to a single character in the text
- Perform case study: Investigate if it is possible to identify discrete sections of the text strongly characterised by an emotion.
- The effectiveness of the NLP techniques and the results of these case studies will be discussed in the final paper writeup.

3.2 Methodology

The first steps will be to design and begin build an application that will perform the analysis on the text, tracking the emotional ups and downs to allow it to plot and graph the data at the end to gain a curve to associate with the analysed text. This part of the development will adopt a Waterfall-like model. This first stage will continue until such a time that the application can process an entire text to produce a simple graph

Beyond this, each case study will require further design/development work to expand upon the base program to allow the case studies (detailed in section 3.1) to be undertaken. An Agile work cycle will be adopted at this point to allow each case study to be examined individually. Each case study will be designed, developed, refined and carried out individually before moving on to the next.

The exact language and NLP tools to be used are not currently finalised. Candidates for use include Python with NLTK, Java with Apache OpenNLP and C# with CoreNLP. This will be finalised during the early design stage of the base application.

Similarly, the exact texts for study are also yet to be chosen. Fiction will be the target (though non-fiction works may be processed given ample time constraints as a point of interest), with no particular genre of great interest, but texts that have known stories to

myself would be desirable to easily check the correct functioning of the application during testing. Project Gutenburg provides multiple formats for all available books, including plaintext which should require very little sanitation and should be somewhat easy to parse for analysis. Exact texts will be finalised during early development stages of the application.

4. Programme of Work

The project will begin mid-October 2019 and will run till March 2020. It will be split into multiple stages as follows:

- Analysis and Design – This will involve the initial stages of the Waterfall methodology and development of designs for the application using software engineering techniques and selection of the exact language, frameworks and NLP toolkits to be used. This stage will take approximately 1 week.
- Base Application Development and Testing – This will involve the latter parts of the Waterfall methodology in development of the application from the ground up using the designs from the previous stage, testing and maintaining the application beyond testing into the case studies as further modifications are made. This stage will take 4 weeks with a chance to overrun due to magnitude of stage.
- Input Data / Text Selection – This will involve selection of several (known) texts for analysis. This stage may take place anytime between commencement of the project and completion of the previous stage. This stage will take no more than 1 week combined but may be spread over longer and interlaced to take place concurrently with earlier stages.
- Investigation – This will involve use of the developed application to analyse the selected texts and attempts to categorise them into the story types as defined earlier in this document. This stage is more of a milestone and should not occupy a significant timeframe.
- Further Investigation and Case Studies:
    - Case Study 1: Whole text vs. chapters vs. sections – This will involve the development and testing of advanced features in the application to investigate small sections of text compared to the full text.
    - Case Study 2: Per character Curve Analysis – This will involve the development and testing of advanced features in the application to investigate curves produced by analysing only lines pertaining to a single character.
    - Case Study 3: Identification of Discrete Emotional Sections – This will involve the development and testing of advanced code in the application to investigate the possibility to segment a text into discrete sections based upon the primary emotions identified per section.
    - Each Case Study should take approximately 1 week, but each may run into the next in some form (e.g. Investigation itself of CS1 may run over development of code for CS2)
- Project Analysis and Evaluation – This will involve collating of all the results from each of the investigations and combining them into a single summary and discussion document. Further, reflection upon the effectiveness of the application, how well the selected texts could be classified into Vonnegut's story types and analysis of project workflow itself will feature.

A schedule for the project is shown in a Gantt chart in Figures 1 and 2 on the following pages.

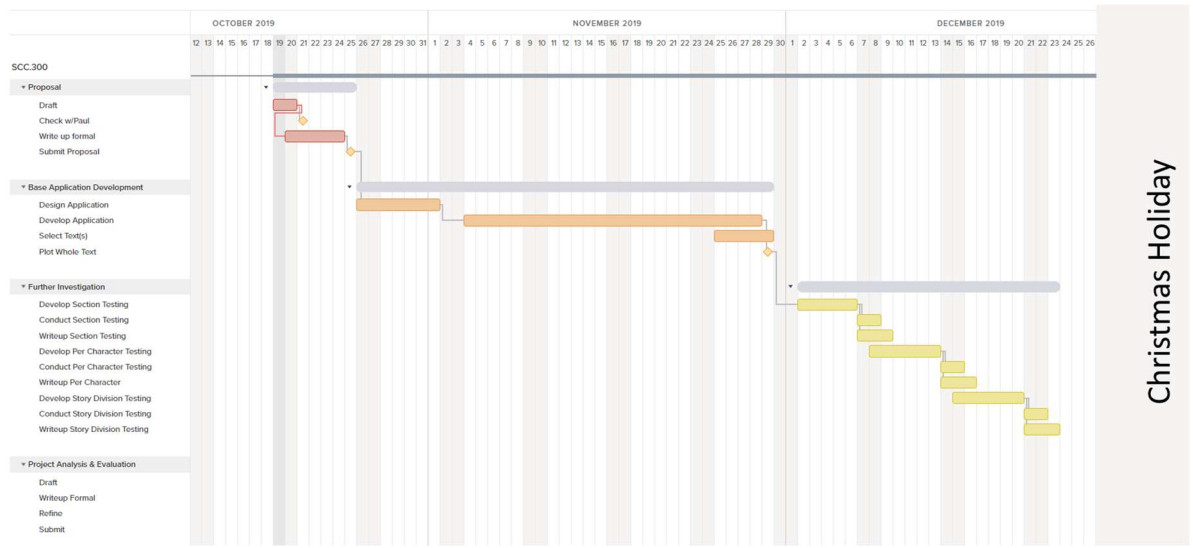

Figure 1: Project Schedule Oct-Dec 2019

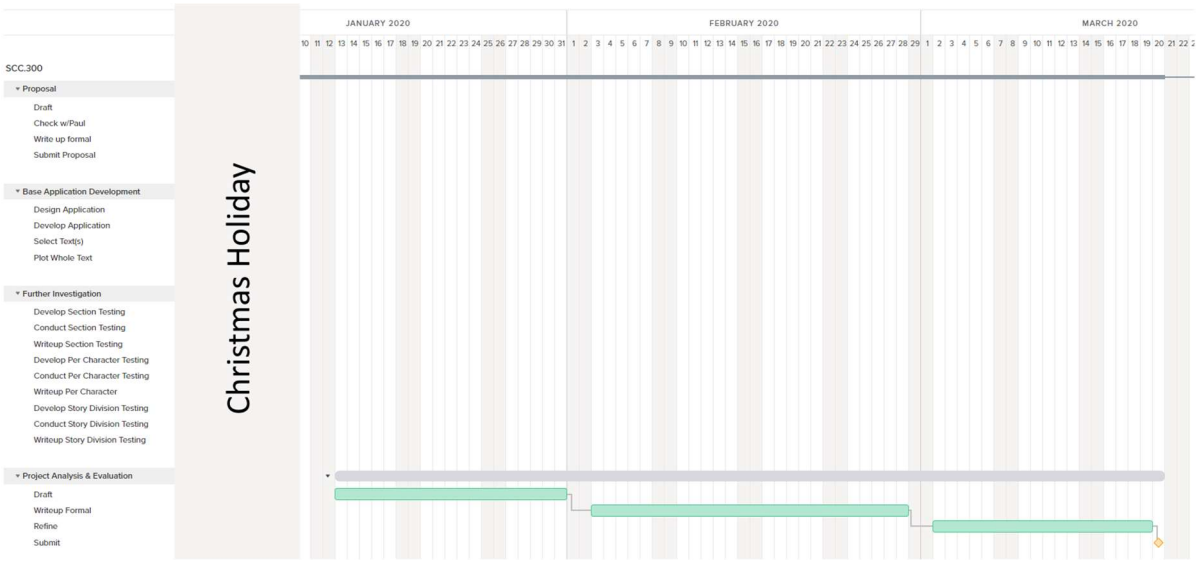

Figure 2: Project Schedule Jan-Mar 2020

5. Resources Required

Access to a software development specification computer capable of developing applications. Exact requirements for software will depend upon chosen language and toolkits.

- Python
  - Python IDE
  - NLTK and supporting libraries
- Java
  - Java IDE
  - Apache OpenNLP
- C#
  - C# IDE

- CoreNLP

Access to Project Gutenburg from which to select texts and input data.

All the above are freely available or already provisioned by Lancaster University as part-and-parcel of Undergraduate Study.

6. References

1. Kurt Vonnegut's Master's Thesis Rejected by University of Chicago
   https://whitherthebook.wordpress.com/2017/02/15/kurt-vonneguts-masters-thesis-rejected-by-university-of-chicago/
2. Kurt Vonnegut: The Shapes of Stories:
   https://fs.blog/2011/09/kurt-vonnegut-the-shapes-of-stories/
3. Kurt Vonnegut on The Shape of Stories (Short Lecture):
   https://www.youtube.com/watch?v=GOGru_4z1Vc
4. Extraction and Analysis of Fictional Character Networks: A Survey
   https://dl.acm.org/citation.cfm?doid=3362097.3344548
5. The Hedonometer Team
   http://hedonometer.org/books/v2/v2/
6. What Makes Us Happy
   https://www.turing.ac.uk/blog/what-makes-us-happy
   https://theconversation.com/what-makes-us-happy-we-analysed-200-years-of-written-text-to-find-the-answer-125252
   https://www.nature.com/articles/s41562-019-0750-z
7. Movie Story Shapes
   https://www.turing.ac.uk/blog/u-shaped-emotional-rollercoaster
   https://arxiv.org/abs/1807.02221