Measuring the Influence of War on 20th Century Literature

Andre Marquez, Mark O'Shea
Rutgers University
agm86@rutgers.edu

Abstract

We are attempting to present an emiprical and analytic approach to the measurement of events on pop culture in a given restricted domain of time. In this paper we will discuss our approach applied to the measurement of the influence of war on popular literature in 20th century America. We will first discuss data collection: methods, standards, and cleaning. Then we will discuss the weighted and normalized popular elements approach for anlayzing this data. Final we discuss our conclusions about the approach, and results of the project. Note: We have not found any related work.

I. Introduction

Without a doubt, pop culture has been heavily shaped and molded by current events. There are movies like *Saving Private Ryan*, songs like *We Didn't Start the Fire*, and books like *To Kill a Mockingbird* that have roots in events that were happening around and before the time.

We want to investigate this problem and develop an approach that measures the extent to which pop culture has been influenced by current events — World War II is obviously influential, but how much so? Is it more influential than World War I, or the Great Depression? The original goal was to be able to develop this measurement for any type of historical event over any period of time, however that task was too big to be within the scope of what we can accomplish over the course of this class, so we narrowed our problem to measuring the influence of wars on literature. In order to obtain this measurement we break the problem up into three discrete steps: data preprocessing, analysis of the data, and interpretation of the analysis.

II. Related Work

After doing an extensive, though admittedly not exhaustive, search on our question, there has been no work done prior other than blog posts collecting and listing the most popular events of the 1900s based on little more than general opinion.

With that in mind, that leaves everything to be done. In order to accomplish our goal we need to lay the foundation and build up, as there is nothing for us to build upon currently.

III. Background Information

To reiterate: we are attempting to develop a measurement for the influence of 20th century wars on popular literature. In the past it has just been a opinion based, qualitative results with no real research or quantitative analysis.

IV. Proposed Approach

In order to obtain this measurement we break the problem up into three discrete steps: data preprocessing, analysis of the data, and interpretation of the analysis. In the data we've requried there to be metadata on publication date of the book, and information on the topic

of the book, (e.g. "Love", "Science Fiction", "War"). We also have a list of 20th century events (notable, and not noteworthy alike). We will use the popularity metric whose algorithm proceeds as follows:

Consider a stream of books from the dataset, sorted by publication date, limited to books published in the 1900's (including 1900). We will use an exponentially decaying window with a constant c.

- For each topic we are maintaining, multiply its score by $(1 - c)$
- Suppose a book with a topic whose score is zero is being read int he stream. Then add 1 to the score for that topic.
- if the score for each topic drops below some threshold λ, then reset that topics score to 0.

This will be computed overall books, and for individual decades as well.

Then we will attempt to reconcile the popularity scores of "War" books with those of "Romance" and "History" as we want to use those topics as a baseline since we're assuming that the occurance of books written about "Romance" or "History" will occur relatively independently of war events.

We will also compare the rise and fall of the scores to the occurence of wars. For example, we might expect more war books to be written after World War II than before World War I.

V. Experiments

As said before, we broke up this project into three discrete steps: data preproccessing, analysis of the data, and interpretation of the analysis. We used three datasets: a thorough list of historical wars obtained by scraping *Wikipedia* for wars in the 20th century, CMU's Book Summaries Corpus which contained plot summaries for 16,659 books from 1895 to 2009, and Freebase's book dataset for the metadeta consisting of publication date and topics of each of these 16,659 books.

The preproccessing aspect of the datamining was perhaps the hardest part. Many books had incorrect or even missing publication dates which we had to fix, and some books didn't have genre's associated with them. To attack the problem of a lack of publication dates, we used the Google Books API to retrieve the correct publication date of a given book. Addressing the lack of genre was a bit more difficult: we eliminated stop words using *jmlr.org*'s list of stop words, and then searched for keywords pertaining to war (e.g. 'general', 'soldier', 'war', 'fight', 'peace') to try and refine if the book belonged in the genre war.

After much work we still had to throw out a significant portion of the data leaving us with 6,579 books. We then sorted the books in ascending order by their publication date, and placed them into buckets by decade. We then ran a variation on the popularity score method, where instead of having constant λ, we had $\lambda(x)$ where x was the topic, such that and $\lambda(x)$ increased proportionally with the frequency of the book topic. This way we could somewhat normalize the results of the popularity scores of the "Romance" books and the "War" books, since Romance occured so much more frequently that it was skewing our results heavily. We then took the ratio of popularity to the number of books on that topic that were published to obtain our measurement (Note: this was progress made since our presentation. You will notice that this was necessary since in popularity, war books were dwarfed by both history and romance, but this is a result of there simply being more books published in those categories.

Number of Books on Each Topic Published per Decade			
Year	**War**	**Romance**	**Historical**
1900	1	9	9
1910	1	8	10
1920	1	15	9
1930	2	8	15
1940	8	12	15
1950	12	17	37
1960	7	11	27
1970	11	11	50
1980	8	28	74
1990	16	47	82
total	69	173	344

Weighted Popularity Scores Across the Decades			
Year	**War**	**Romance**	**History**
1900	0.9960	8.9651	8.9691
1910	1.9990	7.9690	9.9531
1920	1.0000	14.9302	8.9551
1930	1.9920	7.9681	14.9451
1940	7.9512	11.9391	14.9272
1950	11.9432	16.9362	36.8673
1960	6.9721	10.9531	26.8773
1970	10.9631	10.9541	49.7885
1980	7.9621	27.8863	73.6520
1990	15.9073	49.8095	81.6818

In the above tables, you can see the clear effects of the large number of history books and romance books in comparison to the smaller number of war books. Because of the size of our dataset, we are not sure what to make of this; it could be either the dataset omitted many war books, or just had more access to romance or history books. Either way this was the case, and as you can see the popularity scores are heavily skewed in favor of the topics with more books. (This was also true for science fiction which, like history, had many books that fell under that topic across the century.) Again, in order to remedy this, we tried to look at the ratio of popularity to number of books published, which gave us nicer results:

Ratio of Popularity to Total Number of Topic Books			
Year	**War**	**Romance**	**History**
1900	0.14435	0.05182	0.02607
1910	0.02897	0.04606	0.02893
1920	0.01449	0.08630	0.02603
1930	0.02887	0.04605	0.04345
1940	0.11523	0.06901	0.04339
1950	0.17309	0.09789	0.10717
1960	0.10105	0.06331	0.07813
1970	0.15889	0.06332	0.14473
1980	0.11539	0.16119	0.21401
1990	0.23054	0.27058	0.23745

By looking at the ratio we're able to discern some sort of spike in books on war in the 40s, 50s, and 60s. However, this interpretation should be taken with a grain of salt considering how small our data set is, and how much we tried to normalize everything during the analysis. This does however coincide with the occurance of the two largest wars: World War II (1939 - 1945), the Vietnam War (1954 - 1975), the Korean War (1950 - 1953), and the Cold War (1945 - 1990). So while we aren't able to discern just how much a singular event influenced the presence of war in literature, we can draw attention to an increase in war books within some neighborhood of time of certain wars. Again, this is all interpretation based off of a small data set, so nothing is definitive.

VI. Conclusions and Future Work

We set out on this project hoping to establish a metric for measuring the impact of a specific event on pop culture. However, such a broad problem could not be addressed in the scope of this class so we reduced it to the general impact of war on literature in the 20th century. It is difficult to say anything concrete about the metric and measurements we used, as the small data set is prone to be misleading.

One of the biggest hurdles, in fact, was the search for appropriate datasets. Though there exist many datasets on movies, music, and literature, it was very hard to find datasets that also included information pertaining to what topic the particular instance of movie, music, or literature was on. For example, it cannot be deduced from the title, composer, and song length that the song *We Didn't Start the Fire* by Billy Joel is related to war, peace, and global conflict in general, thus we absolutely needed some parameter that gave us this information in our data set.

Even though we eventually found such a dataset that contained more than 16,000 books (large enough), and had the parameters we needed, after the preprocessing the number was reduced just above 6,000, and of these 6,000 only 69 were on war. All of the problems we had, from preprocessing to interpretation of the results, stem from this lack of data. That being said, even though we lacked data, we have blazed the trail in the field of analyzing and measuring the impact of events on pop culture which is an interesting problem.

Should we be able to get our hands on appropriately sized data sets, I'm optomistic about the performance of the metric on those larger sets, which would inspire a generalization of this method across not just wars, but any kind of current event that happened.

VII. References

www.wikipedia.org

www.cs.cmu.edu/ dbamman/booksummaries.html

www.freebase.com/book

infolab.stanford.edu/ ullman/mmds/ch4.pdf