College of Engineering and Mathematical Sciences Faculty Publications

College of Engineering and Mathematical Sciences

The emotional arcs of stories are dominated by six basic shapes

Andrew J. Reagan
University of Vermont

Lewis Mitchell
The University of Adelaide

Dilan Kiley
University of Vermont

Christopher M. Danforth
University of Vermont

Peter Sheridan Dodds
University of Vermont

Recommended Citation

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE

Open Access

CrossMark

The emotional arcs of stories are dominated by six basic shapes

Andrew J Reagan[1*], Lewis Mitchell[2], Dilan Kiley[1], Christopher M Danforth[1] and Peter Sheridan Dodds[1]

*Correspondence:
andrew.reagan@uvm.edu
[1] University of Vermont, 85 South
Prospect St, Burlington, VT 05405,
USA
Full list of author information is
available at the end of the article

Abstract

Advances in computing power, natural language processing, and digitization of text now make it possible to study a culture's evolution through its texts using a 'big data' lens. Our ability to communicate relies in part upon a shared emotional experience, with stories often following distinct emotional trajectories and forming patterns that are meaningful to us. Here, by classifying the emotional arcs for a filtered subset of 1,327 stories from Project Gutenberg's fiction collection, we find a set of six core emotional arcs which form the essential building blocks of complex emotional trajectories. We strengthen our findings by separately applying matrix decomposition, supervised learning, and unsupervised learning. For each of these six core emotional arcs, we examine the closest characteristic stories in publication today and find that particular emotional arcs enjoy greater success, as measured by downloads.

Keywords: stories; sentiment mining; narratology; language; society

1 Introduction

The power of stories to transfer information and define our own existence has been shown time and again [1–5]. We are fundamentally driven to find and tell stories, likened to *Pan Narrans* or *Homo Narrativus*. Stories are encoded in art, language, and even in the mathematics of physics: We use equations to represent both simple and complicated functions that describe our observations of the real world. In science, we formalize the ideas that best fit our experience with principles such as Occam's Razor: The simplest story is the one we should trust. We tend to prefer stories that fit into the molds which are familiar, and reject narratives that do not align with our experience [6].

We seek to better understand stories that are captured and shared in written form, a medium that since inception has radically changed how information flows [7]. Without evolved cues from tone, facial expression, or body language, written stories are forced to capture the entire transfer of experience on a page. An often integral part of a written story is the emotional experience that is evoked in the reader. Here, we use a simple, robust sentiment analysis tool to extract the reader-perceived emotional content of written stories as they unfold on the page.

We objectively test aspects of the theories of folkloristics [8, 9], specifically the commonality of core stories within societal boundaries [4, 10]. A major component of folkloristics is the study of society and culture through literary analysis. This is sometimes referred to as *narratology*, which at its core is 'a series of events, real or fictional, presented to the

Springer

reader or the listener' [11]. In our present treatment, we consider the plot as the 'backbone' of events that occur in a chronological sequence (more detail on previous theories of plot are in Appendix A in Additional file 1). While the plot captures the mechanics of a narrative and the structure encodes their delivery, in the present work we examine the emotional arc that is invoked through the words used. The emotional arc of a story does not give us direct information about the plot or the intended meaning of the story, but rather exists as part of the whole narrative (*e.g.*, an emotional arc showing a fall in sentiment throughout a story may arise from very different plot and structure combinations). This distinction between the emotional arc and the plot of a story is one point of misunderstanding in other work that has drawn criticism from the digital humanities community [12]. Through the identification of motifs [13], narrative theories [14] allow us to analyze, interpret, describe, and compare stories across cultures and regions of the world [15]. We show that automated extraction of emotional arcs is not only possibly, but can test previous theories and provide new insights with the potential to quantify unobserved trends as the field transitions from data-scarce to data-rich [16, 17].

The rejected master's thesis of Kurt Vonnegut - which he personally considered his greatest contribution - defines the *emotional arc* of a story on the 'Beginning-End' and 'Ill Fortune-Great Fortune' axes [18]. Vonnegut finds a remarkable similarity between Cinderella and the origin story of Christianity in the Old Testament (see Figure S1 in Appendix B in Additional file 1), leading us to search for all such groupings. In a recorded lecture available on YouTube [19], Vonnegut asserted:

> 'There is no reason why the simple shapes of stories can't be fed into computers, they are beautiful shapes.'
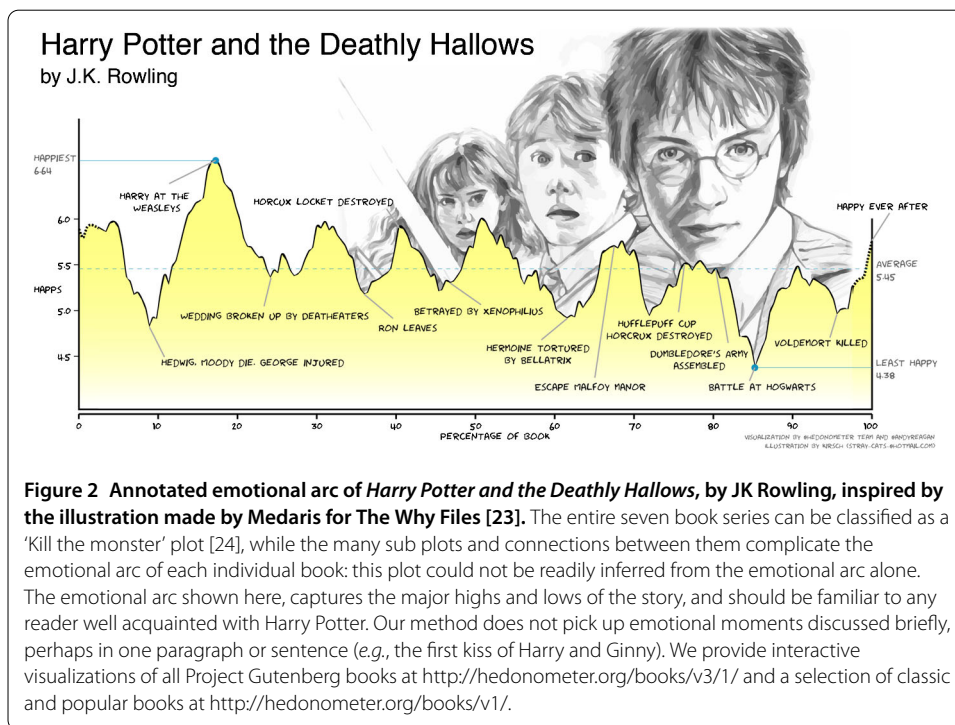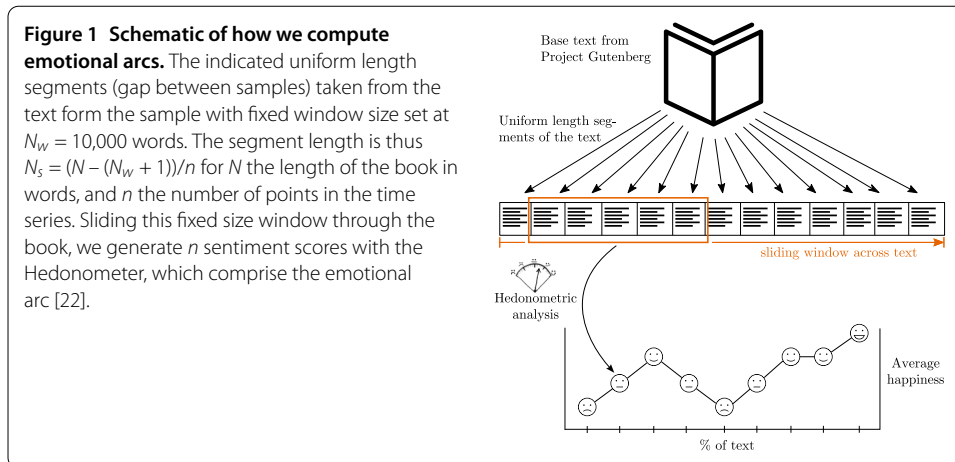
For our analysis, we apply three independent tools: matrix decomposition by singular value decomposition (SVD), supervised learning by agglomerative (hierarchical) clustering with Ward's method, and unsupervised learning by a self-organizing map (SOM, a type of neural network). Each tool encompasses different strengths: the SVD finds the underlying basis of all of the emotional arcs, the clustering classifies the emotional arcs into distinct groups, and the SOM generates arcs from noise which are similar to those in our corpus using a stochastic process. It is only by considering the results of each tool in support of each other that we are able to confirm our findings.

We proceed as follows. We first introduce our methods in Section 2, we then discuss the combined results of each method in Section 3, and we present our conclusions in Section 4. A graphical outline of the methodology and results can be found as Figure S2 in Appendix B in Additional file 1.

2 Methods

2.1 Emotional arc construction

To generate emotional arcs, we analyze the sentiment of 10,000 word windows, which we slide through the text (see Figure 1). We rate the emotional content of each window using our Hedonometer with the labMT dataset, chosen for lexical coverage and its ability to generate meaningful word shift graphs, specifically using 10,000 words as a minimum necessary to generate meaningful sentiment scores [20, 21]. We emphasize that dictionary-based methods for sentiment analysis usually perform worse than random on individual sentences [20, 21], and although this issue can be resolved by using a rolling average of

Figure 1 Schematic of how we compute emotional arcs. The indicated uniform length segments (gap between samples) taken from the text form the sample with fixed window size set at $N_W = 10,000$ words. The segment length is thus $N_s = (N - (N_W + 1))/n$ for N the length of the book in words, and n the number of points in the time series. Sliding this fixed size window through the book, we generate n sentiment scores with the Hedonometer, which comprise the emotional arc [22].

Figure 2 Annotated emotional arc of *Harry Potter and the Deathly Hallows*, by JK Rowling, inspired by the illustration made by Medaris for The Why Files [23]. The entire seven book series can be classified as a 'Kill the monster' plot [24], while the many sub plots and connections between them complicate the emotional arc of each individual book: this plot could not be readily inferred from the emotional arc alone. The emotional arc shown here, captures the major highs and lows of the story, and should be familiar to any reader well acquainted with Harry Potter. Our method does not pick up emotional moments discussed briefly, perhaps in one paragraph or sentence (*e.g.*, the first kiss of Harry and Ginny). We provide interactive visualizations of all Project Gutenberg books at http://hedonometer.org/books/v3/1/ and a selection of classic and popular books at http://hedonometer.org/books/v1/.

sentences scores, it begets a basic misunderstanding of similar efforts [12]. In Figure 2, we show the emotional arc of *Harry Potter and the Deathly Hallows*, the final book in the popular Harry Potter series by JK Rowling. While the plot of the book is nested and complicated, the emotional arc associated with each sub-narrative is clearly visible. We analyze the emotional arcs corresponding to complete books, and to limit the conflation of multiple core emotional arcs, we restrict our analysis to shorter books by selecting a maximum number of words when building our filter. Further details of the emotional arc construction can be found in Appendix C in Additional file 1.

2.2 Project Gutenberg corpus

For a suitable corpus we draw on the open access Project Gutenberg data set [25]. We apply rough filters to the collection (roughly 50,000 books) in an attempt to obtain a

set of books that represent English works of fiction. We start by selecting for only English books, with total words between 20,000 and 100,000, with more than 40 downloads from the Project Gutenberg website, and with Library of Congress Class corresponding to English fiction.[a] To ensure that the 40-download limit is not influencing the results here, we further test each method for 10, 20, 40, and 80 download thresholds, in each case confirming the 40 download findings to be qualitatively unchanged. Next, we remove books with any word in the title from a list of keywords (*e.g.*, 'poems' and 'collection', full list in Appendix C in Additional file 1). From within this set of books, we remove the front and back matter of each book using regular expression pattern matches that match on 98.9% of the books included. Two slices of the data for download count and the total word count are shown in Appendix C, Figure S4 in Additional file 1. We provide a list of the book ID's which are included for download in the Online Appendices at http://compstorylab.org/share/papers/reagan2016b/, the books are listed in Table S1 in Appendix D in Additional file 1, and we attempt to provide the Project Gutenberg ID when we mention a book by title herein. Given the Project Gutenberg ID *n*, the raw ebook is available online from Project Gutenberg at http://www.gutenberg.org/ebooks/n/, *e.g.*, *Alice's Adventures in Wonderland* by Lewis Carroll, has ID 11 and is available at http://www.gutenberg.org/ebooks/11/. We also provide an online, interactive version of the emotional arc for each book indexed by the ID, *e.g.*, *Alice's Adventures in Wonderland* is available at http://hedonometer.org/books/v3/11/.

2.3 Principal component analysis (SVD)

We use the standard linear algebra technique singular value decomposition (SVD) to find a decomposition of stories onto an orthogonal basis of emotional arcs. Starting with the sentiment time series for each book b_i as row i in the matrix A, we apply the SVD to find

$$A = U \Sigma V^T = W V^T, \tag{1}$$

where U contains the projection of each sentiment time series onto each of the right singular vectors (rows of V^T, eigenvectors of $A^T A$), which have singular values given along the diagonal of Σ, with $W = U\Sigma$. Different intuitive interpretations of the matrices U, Σ, and V^T are useful in the various domains in which the SVD is applied; here, we focus on right singular vectors as an orthonormal basis for the sentiment time series in the rows of A, which we will refer to as the *modes*. We combine Σ and U into the single coefficient matrix W for clarity and convenience, such that W now represents the mode coefficients.

2.4 Hierarchical clustering

We use Ward's method to generate a hierarchical clustering of stories, which proceeds by minimizing variance between clusters of books [26]. We use the mean-centered books and the distance function

$$D(b_i, b_j) = l^{-1} \sum_{t=1}^{l} \left| b_i(t) - b_j(t) \right| \tag{2}$$

for t indexing the window in books b_i, b_j to generate the distance matrix.

2.5 Self-organizing map (SOM)

We implement a self-organized map (SOM), an unsupervised machine learning method (a type of neural network) to cluster emotional arcs [27]. The SOM works by finding the most similar emotional arc in a random collection of arcs. We use an 8×8 SOM (for 64 nodes, roughly 5% of the number of books), connected on a square grid, training according to the original procedure (with winner take all, and scaling functions across both distance and magnitude). We take the neighborhood influence function at iteration i as

$$\text{Nbd}_k(i) = \left[j \in \mathcal{N} \,|\, D(k,j) < \sqrt{N} \cdot (i+1)^\alpha \right] \qquad (3)$$

for a node k in the set of nodes $\mathcal{N}$, with distance function D given above and total number of nodes N. For results shown here we take $\alpha = -0.15$. We implement the learning adaptation function at training iteration i as $f(i) = (i+1)^\beta$, again with $\beta = -0.15$, a standard value for the training hyper-parameters.
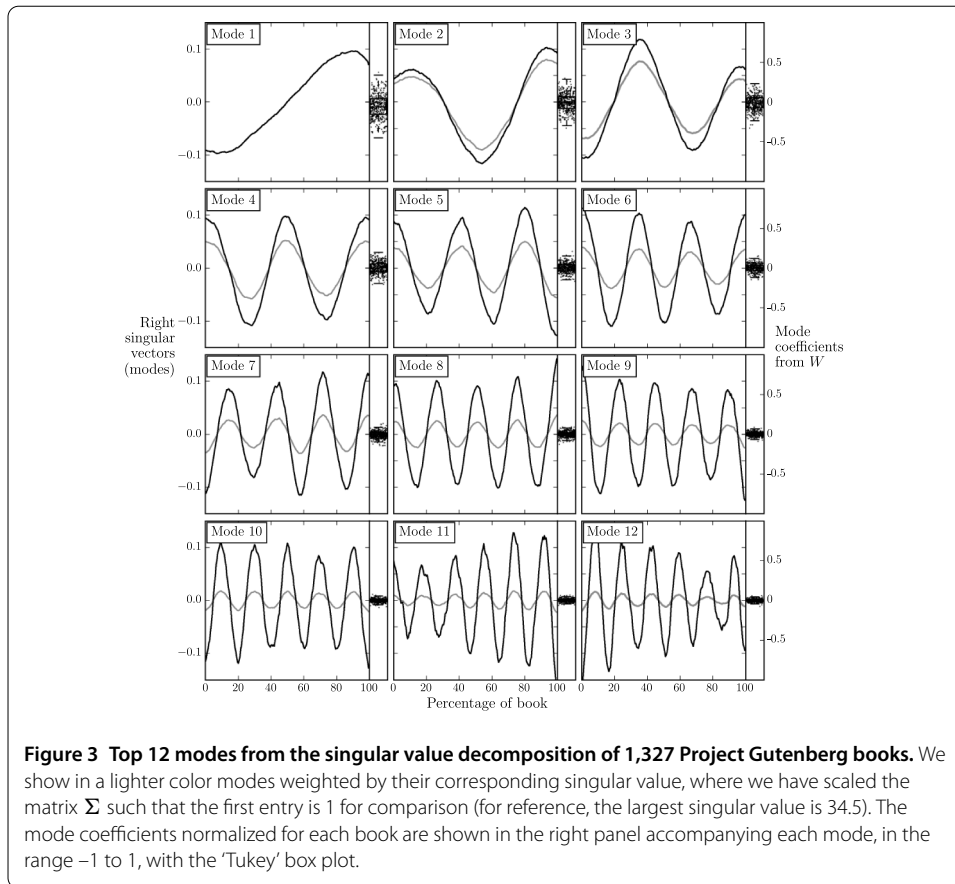
3 Results

We obtain a collection of 1,327 books that are mostly, but not all, fictional stories by using metadata from Project Gutenberg to construct a rough filter. We find broad support for the following six emotional arcs:

- 'Rags to riches' (rise).
- 'Tragedy', or 'Riches to rags' (fall).
- 'Man in a hole' (fall-rise).
- 'Icarus' (rise-fall).
- 'Cinderella' (rise-fall-rise).
- 'Oedipus' (fall-rise-fall).

Importantly, we obtain these same six emotional arcs from all possible arcs by observing them as the result of three methods: As modes from a matrix decomposition by SVD, as clusters in a hierarchical clustering using Ward's algorithm, and as clusters using unsupervised machine learning. We examine each of the results in this section.
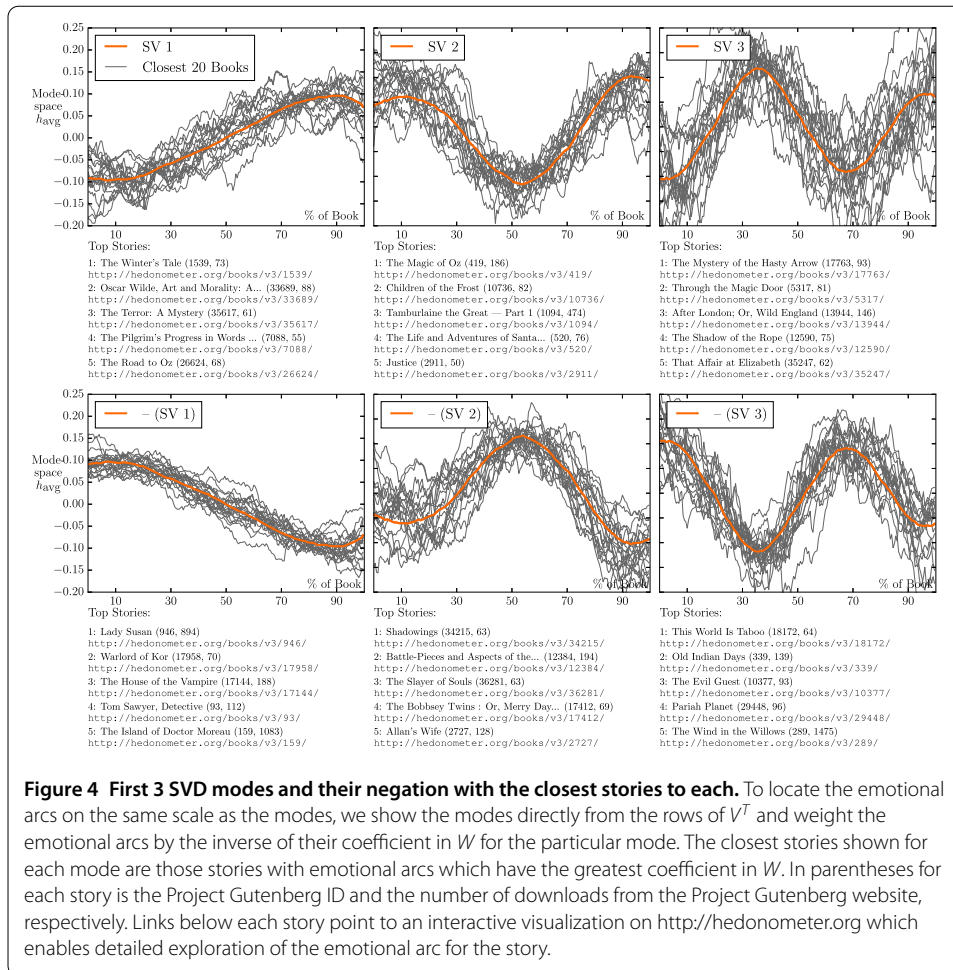
3.1 Principal component analysis (SVD)

In Figure 3 we show the leading 12 modes in both the weighted (dark) and un-weighted (lighter) representation. In total, the first 12 modes explain 80% and 94% of the variance from the mean centered and raw time series, respectively. The modes are from mean-centered emotional arcs, such that the first SVD mode need not extract the average from the labMT scores nor the positivity bias present in language [28]. The coefficients for each mode within a single emotional arc are both positive and negative, so we need to consider both the modes and their negation. We can immediately recognize the familiar shapes of core emotional arcs in the first four modes, and compositions of these emotional arcs in modes 5 and 6. We observe 'Rags to riches' (mode 1, positive), 'Tragedy' or 'Riches to rags' (mode 1, negative), Vonnegut's 'Man in a hole' (mode 2, positive), 'Icarus' (mode 2, negative), 'Cinderella' (mode 3, positive), 'Oedipus' (mode 3, negative). We choose to include modes 7-12 only for completeness, as these high frequency modes have little contribution to variance and do not align with core emotional arc archetypes from other methods (more below).

Figure 3 Top 12 modes from the singular value decomposition of 1,327 Project Gutenberg books. We show in a lighter color modes weighted by their corresponding singular value, where we have scaled the matrix Σ such that the first entry is 1 for comparison (for reference, the largest singular value is 34.5). The mode coefficients normalized for each book are shown in the right panel accompanying each mode, in the range −1 to 1, with the 'Tukey' box plot.

We emphasize that by definition of the SVD, the mode coefficients in W can be either positive and negative, such that the modes themselves explain variance with both the positive and negative version. In the right panels of each mode in Figure 3 we project the 1,327 stories onto each of first six modes and show the resulting coefficients. While none are far from 0 (as would be expected), mode 1 has a mean slightly above 0 and both modes 3 and 4 have means slightly below 0. To sort the books by their coefficient for each mode, we normalize the coefficients within each book in the rows of W to sum to 1, accounting for books with higher total energy, and these are the coefficients shown in the right panels of each mode in Figure 3. In Appendix E in Additional file 1, we provide supporting, intuitive details of the SVD method, as well as example emotional arc reconstruction using the modes (see Figures S5-S7 in Additional file 1). As expected, less than 10 modes are enough to reconstruct the emotional arc to a degree of accuracy visible to the eye.

We show labeled examples of the emotional arcs closest to the top 6 modes in Figure 4 and Figure S8 in Additional file 1. We present both the positive and negative modes, and the stories closest to each by sorting on the coefficient for that mode. For the positive stories, we sort in ascending order, and *vice versa*. Mode 1, which encompasses both the 'Rags to riches' and 'Tragedy' emotional arcs, captures 30% of the variance of the entire space. We examine the closest stories to both sides of modes 1-3, and direct the reader to Figure S8 in Additional file 1 for more details on the higher order modes. The two stories that have the most support from the 'Rags to riches' mode are *The Winter's Tale* (1,539) and *Oscar Wilde, Art and Morality: A Defence of 'The Picture of Dorian Gray'* (33,689). Among

Figure 4 First 3 SVD modes and their negation with the closest stories to each. To locate the emotional arcs on the same scale as the modes, we show the modes directly from the rows of V^T and weight the emotional arcs by the inverse of their coefficient in W for the particular mode. The closest stories shown for each mode are those stories with emotional arcs which have the greatest coefficient in W. In parentheses for each story is the Project Gutenberg ID and the number of downloads from the Project Gutenberg website, respectively. Links below each story point to an interactive visualization on http://hedonometer.org which enables detailed exploration of the emotional arc for the story.

the most categorical tragedies we find *Lady Susan* (946) and *Warlord of Kor* (17,958). Number 8 in the sorted list of tragedies is perhaps the most famous tragedy: *Romeo and Juliet* by William Shakespeare. Mode 2 is the 'Man in a hole' emotional arc, and we find the stories which most closely follow this path to be *The Magic of Oz* (419) and *Children of the Frost* (10,736). The negation of mode 2 most closely resembles the emotional arc of the 'Icarus' narrative. For this emotional arc, the most characteristic stories are *Shadowings* (34,215) and *Battle-Pieces and Aspects of the War* (12,384). Mode 3 is the 'Cinderella' emotional arc, and includes *Mystery of the Hasty Arrow* (17,763) and *Through the Magic Dorr* (5,317). The negation of Mode 3, which we refer to as 'Oedipus', is found most characteristically in *This World is Taboo* (18,172), *Old Indian Days* (339), and *The Evil Guest* (10,377). We also note that the spread of the stories from their core mode increases strongly for the higher modes.

3.2 Hierarchical clustering

We show a dendrogram of the 60 clusters with highest linkage cost in Figure 5. The average silhouette coefficient is shown on the bottom of Figure 5, and the distributions of silhouette values within each cluster (see Figures S17 and S18 in Additional file 1) can be used to analyze the appropriate number of clusters [29]. A characteristic book from each cluster is shown on the leaf nodes by sorting the books within each cluster by the total
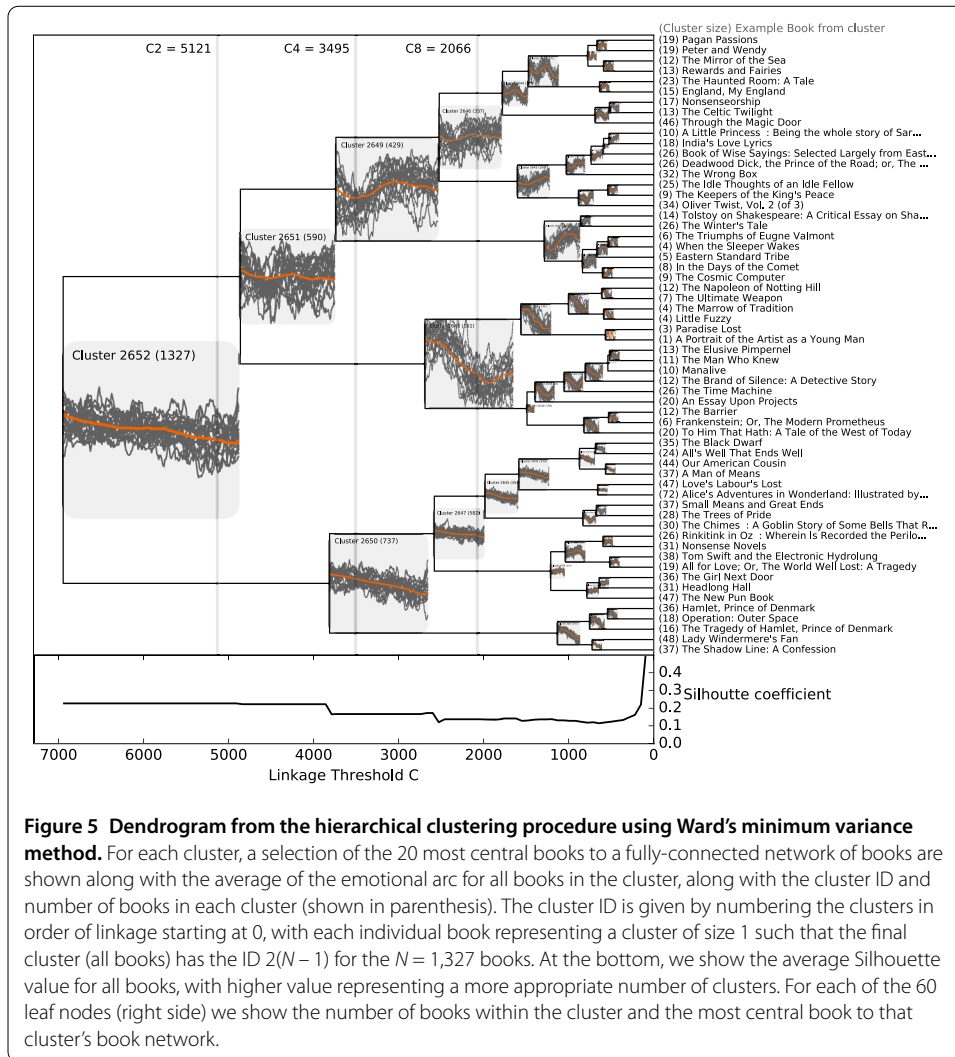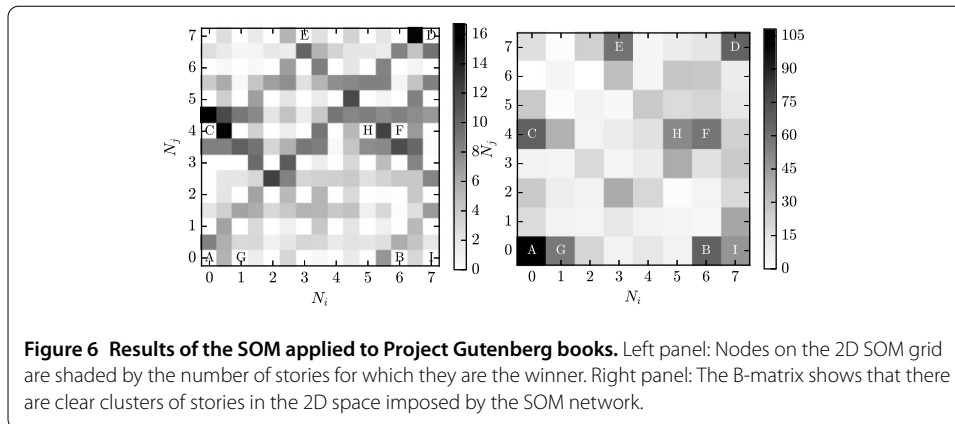
Figure 5 Dendrogram from the hierarchical clustering procedure using Ward's minimum variance method. For each cluster, a selection of the 20 most central books to a fully-connected network of books are shown along with the average of the emotional arc for all books in the cluster, along with the cluster ID and number of books in each cluster (shown in parenthesis). The cluster ID is given by numbering the clusters in order of linkage starting at 0, with each individual book representing a cluster of size 1 such that the final cluster (all books) has the ID 2(N – 1) for the N = 1,327 books. At the bottom, we show the average Silhouette value for all books, with higher value representing a more appropriate number of clusters. For each of the 60 leaf nodes (right side) we show the number of books within the cluster and the most central book to that cluster's book network.

distance to other books in the cluster (*e.g.*, considering each intra-cluster collection as a fully connected weighted network, we take the most central node), and in parenthesis the number of books in that cluster. In other words, we label each cluster by considering the network centrality of the fully connected cluster with edges weighted by the distance between stories. By cutting the dendrogram in Figure 5 at various linkage costs we are able to extract clusters of the desired granularity. For the cuts labeled C2, C4, and C8, we show these clusters in Figures S9, S11, and S15 in Additional file 1. We find the first four of our final six arcs appearing among the eight most different clusters (Figure S15 in Additional file 1).

The clustering method groups stories with a 'Man in a hole' emotional arc for a range of different variances, separate from the other arcs, in total these clusters (panels A, E, and I of Figure S16 in Additional file 1) account for 30% of the Gutenberg corpus. The remainder of the stories have emotional arcs that are clustered among the 'Tragedy' arc (32%), 'Rags to riches' arc (5%), and the 'Oedipus' arc (31%). A more detailed analysis of the results from hierarchical clustering can be found in Appendix F in Additional file 1, and this result generally agrees with other attempts that use only hierarchical clustering [12].

Figure 6 Results of the SOM applied to Project Gutenberg books. Left panel: Nodes on the 2D SOM grid are shaded by the number of stories for which they are the winner. Right panel: The B-matrix shows that there are clear clusters of stories in the 2D space imposed by the SOM network.
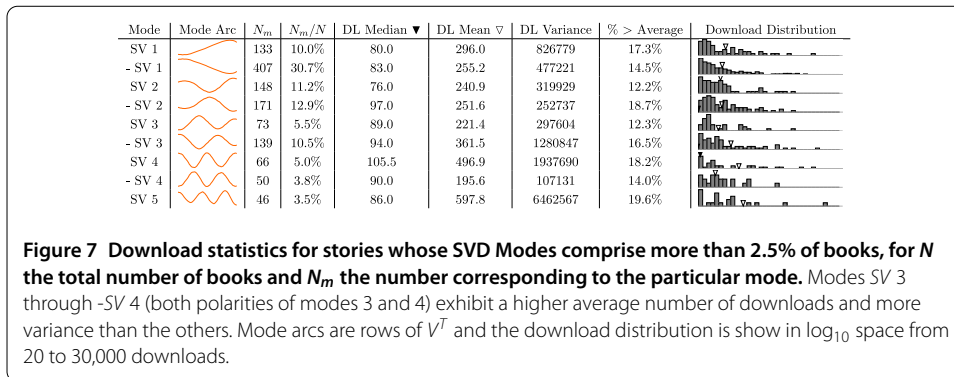
3.3 Self-organizing map (SOM)

Finally, we apply Kohonen's self-organizing map (SOM) and find core arcs from unsupervised machine learning on the emotional arcs. On the two dimensional component plane, the prescribed network topology, we find seven spatially coherent groups, with five emotional arcs. These spatial groups are comprised of stories with core emotional arcs of differing variance.

In Figure 6 we see both the B-matrix to demonstrate the strength of spatial clustering and a heat-map showing where we find the winning nodes. The A-I labels refer to the individual nodes shown in Figure S19 in Additional file 1, and we observe seven spatial groups in the both panels of Figure 6: (1) A and G, (2) B and I, (3) C, (4) D, (5) E, and (6) H, and (7) F. These spatial clusters reinforce the visible similarity of the winning node arcs, given that nodes H and F are close spatially but separated by the B-matrix and contain very distinct arcs. We show the winning node emotional arcs and the arcs of books for which they are the winners in Figure S19 in Additional file 1. The legend shows the node ID, numbers the cluster by size, and in parentheses indicates the size of the cluster on that individual node. In panels A and G we see varying strengths of the 'Man in a hole' emotional arc. In panels B and I, the second largest individual cluster consists of the 'Rags to riches' arcs. In panel C, and in panel F, we find the 'Oedipus' emotional arc, with a more pronounced positive start and decline in panel C. In panel D we see the 'Icarus' arc, and in panel E and panel H we see the 'Tragedy' arc. Each of these top stories are all readily identifiable, yet again demonstrating the universality of these story types.

3.4 Null comparison

There are many possible emotional arcs in the space that we consider. To demonstrate that these specific arcs are uniquely compelling as stories written by and for *homo narrativus*, we consider the true emotional arcs in relation to their most suitable comparison: the book with randomly shuffled words ('word salad') and the resulting text from a 2-gram Markov model trained on the individual book itself ('nonsense'). We chose to compare to 'word salad' and 'nonsense' versions as they are more representative of a null model: written stories that are without coherent plot or structure to generate a coherent emotional arc, which is not true of a stochastic process (*e.g.*, a random walk model or noise). Examples of the emotional arc and null emotional arcs for a single book are shown in Figure S20 in Additional file 1, with 10 'word salad' and 'nonsense' versions. Sampled text using each method is given in Appendix C in Additional file 1. We re-run each method on

Mode	Mode Arc	N_m	N_m/N	DL Median ▼	DL Mean ▽	DL Variance	% > Average	Download Distribution
SV 1		133	10.0%	80.0	296.0	826779	17.3%	
- SV 1		407	30.7%	83.0	255.2	477221	14.5%	
SV 2		148	11.2%	76.0	240.9	319929	12.2%	
- SV 2		171	12.9%	97.0	251.6	252737	18.7%	
SV 3		73	5.5%	89.0	221.4	297604	12.3%	
- SV 3		139	10.5%	94.0	361.5	1280847	16.5%	
SV 4		66	5.0%	105.5	496.9	1937690	18.2%	
- SV 4		50	3.8%	90.0	195.6	107131	14.0%	
SV 5		46	3.5%	86.0	597.8	6462567	19.6%	

Figure 7 Download statistics for stories whose SVD Modes comprise more than 2.5% of books, for *N* the total number of books and *N_m* the number corresponding to the particular mode. Modes *SV* 3 through -*SV* 4 (both polarities of modes 3 and 4) exhibit a higher average number of downloads and more variance than the others. Mode arcs are rows of V^T and the download distribution is show in $\log_{10}$ space from 20 to 30,000 downloads.

the English fiction Gutenberg Corpus with the null versions of each book and verify that the emotional arcs of real stories are not simply an artifact. The singular value spectrum from the SVD is flatter, with higher-frequency modes appearing more quickly, and in total representing 45% of the total variance present in real stories (see Figures S22 and S25 in Additional file 1). Hierarchical clustering generates less distinct clusters with considerably lower linkage cost (final linkage cost 1,400 vs 7,000) for the emotional arcs from nonsense books, and the winning node vectors on a self-organizing map lack coherent structure (see Figures S26 and S29 in Appendix H in Additional file 1).

3.5 The success of stories

To examine how the emotional trajectory impacts success, in Figure 7 we examine the downloads for all of the books that are most similar to each SVD mode (for additional modes, see Figure S3 in Appendix B in Additional file 1). We find that the first four modes, which contain the greatest total number of books, are not the most popular. Along with the negative of mode 2, both polarities of modes 3 and 4 have markedly higher median downloads, while we discount the importance of the mean with the high variance. The success of the stories underlying these emotional arcs suggests that the emotional experience of readers strongly affects how stories are shared. We find 'Icarus' (-SV 2), 'Oedipus' (-SV 3), and two sequential 'Man in a hole' arcs (SV 4), are the three most successful emotional arcs. These results are influenced by individual books within each mode which have high numbers of downloads, and we refer the reader to the download-sorted tables for each mode in Appendix E in Additional file 1.

4 Conclusion

Using three distinct methods, we have demonstrated that there is strong support for six core emotional arcs. Our methodology brings to bear a cross section of data science tools with a knowledge of the potential issues that each present. We have also shown that consideration of the emotional arc for a given story is important for the success of that story. Of course, downloads are only a rough proxy for success, and this work may provide an outline for more detailed analysis of the factors that impact meaningful measures of success, *i.e.*, sales or cultural influence.

Our approach could be applied in the opposite direction: namely by beginning with the emotional arc and aiding in the generation of compelling stories [30]. Understanding the emotional arcs of stories may be useful to aid in constructing arguments [31] and teaching common sense to artificial intelligence systems [32].

Extensions of our analysis that use a more curated selection of full-text fiction can answer more detailed questions about which stories are the most popular throughout time, and across regions [10]. Automatic extraction of character networks would allow a more detailed analysis of plot structure for the Project Gutenberg corpus used here [11, 33, 34]. Bridging the gap between the full text stories [35] and systems that analyze plot sequences will allow such systems to undertake studies of this scale [36]. Place could also be used to consider separate character networks through time, and to help build an analog to Randall Munroe's Movie narrative charts [37].

We are producing data at an ever increasing rate, including rich sources of stories written to entertain and share knowledge, from books to television series to news. Of profound scientific interest will be the degree to which we can eventually understand the full landscape of human stories, and data driven approaches will play a crucial role.

Additional material

Additional file 1: Supplementary material (as Appendices A-H). (pdf)

Author details
[1]University of Vermont, 85 South Prospect St, Burlington, VT 05405, USA. [2]University of Adelaide, Adelaide, SA 5005, Australia.

Endnote
  [a] The specific classes have labels PN, PR, PS, and PZ.

References
 1. Pratchett T, Stewart I, Cohen J (2003) The science of Discworld II: the globe. Ebury Press, London
 2. Campbell J (2008) The hero with a thousand faces, 3rd edn. New World Library, Novato
 3. Gottschall J (2013) The storytelling animal: how stories make us human. Mariner Books, New York
 4. Cave S (2013) The 4 stories we tell ourselves about death.
    http://www.ted.com/talks/stephen_cave_the_4_stories_we_tell_ourselves_about_death
 5. Dodds PS (2013) Homo narrativus and the trouble with fame. Nautilus magazine.
    http://nautil.us/issue/5/fame/homo-narrativus-and-the-trouble-with-fame
 6. Nickerson RS (1998) Confirmation bias: a ubiquitous phenomenon in many guises. Rev Gen Psychol 2:175-220
 7. Gleick J (2011) The information: a history, a theory, a flood. Pantheon, New York
 8. Propp V (1968) Morphology of the folktale (1928). University of Texas Press, Austin
 9. MacDonald MR (1982) Storytellers sourcebook: a subject, title, and motif index to folklore collections for children. Gale Group, Farmington Hills
10. da Silva SG, Tehrani JJ (2016) Comparative phylogenetic analyses uncover the ancient roots of Indo-European folktales. R Soc Open Sci 3(1):150645. doi:10.1098/rsos.150645.
    http://rsos.royalsocietypublishing.org/content/3/1/150645.full.pdf
11. Min S, Park J (2016) Narrative as a complex network: a study of Victor Hugo's Les Misérables. In: Proceedings of HCI Korea
12. Jockers M (2014) A novel method for detecting plot.
    http://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/
13. Dundes A (1997) The motif-index and the tale type index: a critique. J Folklore Res 34:195-202
14. Dolby SK (2008) Literary folkloristics and the personal narrative. Trickster Press, Bloomington
15. Uther H-J (2011) The types of international folktales. A classification and bibliography. Based on the system of Antti Aarne and Stith Thompson. Part I. Animal tales, tales of magic, religious tales, and realistic tales, with an introduction. FF communications, vol 284. Finnish Academy of Science and Letters, Helsinki
16. Kirschenbaum MG (2007) The remaking of reading: data mining and the digital humanities. In: The national science foundation symposium on next generation of data mining and cyber-enabled discovery for innovation, Maryland

17. Moretti F (2013) Distant reading. Verso, New York
18. Vonnegut K (1981) Palm sunday. RosettaBooks LLC, New York
19. Vonnegut K (1995) Shapes of stories. https://www.youtube.com/watch?v=oP3c1h8v2ZQ
20. Reagan A, Tivnan B, Williams JR, Danforth CM, Dodds PS (2015) Benchmarking sentiment analysis methods for large-scale texts: a case for using continuum-scored words and word shift graphs. arXiv:1512.00531
21. Ribeiro FN, Araújo M, Gonçalves P, Gonçalves MA, Benevenuto F (2016) SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Sci 5(1):23. doi:10.1140/epjds/s13688-016-0085-1
22. Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM (2011) Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter PLoS ONE 6(12):e26752. doi:10.1371/journal.pone.0026752
23. Tenenbaum DJ, Barrett K, Medaris SV, Devitt T (2015) In 10 languages, happy words beat sad ones. http://whyfiles.org/2015/in-10-languages-happy-words-beat-sad-ones/
24. Booker C (2006) The seven basic plots: why we tell stories. Bloomsbury Academic, New York
25. Various (2010) Project Gutenberg. http://gutenberg.org
26. Ward JH Jr (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58(301):236-244
27. Kohonen T (1990) The self-organizing map. Proc IEEE 78(9):1464-1480
28. Dodds PS, Clark EM, Desu S, Frank MR, Reagan AJ, Williams JR, Mitchell L, Harris KD, Kloumann IM, Bagrow JP, Megerdoomian K, McMahon MT, Tivnan BF, Danforth CM (2015) Human language reveals a universal positivity bias. Proc Natl Acad Sci USA 112(8):2389-2394
29. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53-65
30. Li B, Lee-Urban S, Johnston G, Riedl M (2013) Story generation with crowdsourced plot graphs. In: Proceedings of the twenty-seventh AAAI conference on artificial intelligence
31. Bex FJ, Bench-Capon TJ (2010) Persuasive stories for multi-agent argumentation. In: AAAI fall symposium: computational models of narrative, vol 10, p 4
32. Riedl MO, Harrison B (2015) Using stories to teach human values to artificial agents
33. Bost X, Labatut V, Linarès G (2016) Narrative smoothing: dynamic conversational network for the analysis of TV series plots. arXiv:1602.07811
34. Prado SD, Dahmen SR, Bazzan ALC, Carron PM, Kenna R (2016) Temporal network analysis of literary texts. arXiv:1602.07275
35. Nenkova A, McKeown K (2012) A survey of text summarization techniques. In: Mining text data. Springer, Berlin, pp 43-76
36. Winston PH (2011) The strong story hypothesis and the directed perception hypothesis
37. Munroe R (2009) Movie narrative charts. http://xkcd.com/657/