Functional Clustering of Fictional Narratives Using Vonnegut Curves

Shan Zhong[1*] and David B. Hitchcock[2]

[1*]School of Information Engineering, Zhejiang Ocean University, Zhoushan, 316000, Zhejiang, China.
[2]Department of Statistics, University of South Carolina, 1600 Hampton Street, Columbia, 29208, SC, U.S.A.

*Corresponding author(s). E-mail(s): zhongshan@zjou.edu.cn;
Contributing authors: hitchcock@stat.sc.edu;

Abstract

Motivated by a public suggestion by the famous novelist Kurt Vonnegut, we clustered functional data that represented sentiment curves for famous fictional stories. We analyzed text data from novels written between 1612 and 1925, and transformed them into curves measuring sentiment as a function of the percentage of elapsed contents of the novel. We employed sentence-level sentiment evaluation and nonparametric curve smoothing. Our clustering methods involved finding the optimal number of clusters, aligning curves using different chronological warping functions to account for phase and amplitude variation, and implementing functional K-means algorithms under the square root velocity framework. Our results revealed insights about patterns in fictional narratives that Vonnegut and others have suggested but not analyzed in a quantitative way.

Keywords: Functional Data Clustering, Text Sentiment, SRVF

1 Introduction

Kurt Vonnegut, a famous American writer, proposed during a (non-mathematical) public lecture in 2010, "The fundamental idea is that stories have shapes which can be drawn on graph paper" (Kurt Vonnegut, 2010). Vonnegut's idea drew a small amount of attention from journalists, e.g., in a post

on the *Washington Post* website (Swanson, Ana, 2015) and in an article by Johnson, Stephen (2022), characterizing novels using Vonnegut's idea from a literary point of view rather than a mathematical or statistical one. Since Vonnegut's original lecture, Reagan et al (2016) used natural language processing, along with principal component analysis, hierarchical clustering of sentiment functions, and clustering via self-organized maps, to assess Vonnegut's ideas via a formal quantitative analysis. In this article we will propose a similar statistical treatment of sentiment curves for a set of classic novels, but we will use approaches specifically designed for functional data. The idea of "Vonnegut curves" that describe the ebbs and flows of a novel's plot brings to mind the statistical field of functional data analysis (FDA), which is designed to analyze datasets consisting of curves, or functions (see Ramsay and Silverman (2005) for a classic introduction to FDA and Kokoszka and Reimherr (2017) for a recent treatment). We employed sentiment extraction methods together with FDA techniques, to provide a technical analysis of Vonnegut's idea that stories have shapes, and to determine whether all novels can be clustered into few groups with respect to these shapes.

With the rapid development of modern language processing models, much effort has been put into analyzing text data from a quantitative point of view. These efforts include learning context from large datasets of narratives written by humans for the purpose of story generation (Fan et al, 2018) or summarization (Allahyari et al, 2017), Dialogue Agents for Information Access (Dhingra et al, 2016), using social media posts for author identification (Madigan et al, 2005), mining opinions from posted reviews to do book recommendations (Sohail et al, 2017), and so on.

In our analysis, we separated a set of English or English-translated novels into ordered sentences and transformed the pure text in these sentences into numerical values using the Google Sentiment API (Google, 2023). The Sentiment API estimates the overall emotional opinion within the text based on positive or negative attitudes. In short, the Google API associates a sentiment score with each sentence or phrase (in general, this fraction of the novel that is assigned a score could be longer, like a chapter, but we chose to use a smaller unit for sentiment scoring). After that, we quantified these emotional or psychological rises and falls as a function of numerical sentiment versus the percentage of the novel's contents elapsed, as curves $f_i(t), i = 1, 2, \ldots, n$ where $t \in [0, 1]$, which we then plotted on graphs as Vonnegut suggested. Finally we used functional clustering methods on the obtained curves to discover the similarities and differences in the "shapes" of these novels.

2 Review of Functional Data Registration and Clustering

Clustering, a type of unsupervised learning, is fairly common, increasingly so for functional data (Tarpey, 2007; Liu and Yang, 2009; Sangalli et al, 2010; Jacques and Preda, 2014b). Overviews of functional clustering are given by

Hitchcock and Greenwood (2015) and Jacques and Preda (2014a). As opposed to clustering observed data points (or vectors), clustering functional data is usually performed by viewing each curve as an entity as a whole and clustering based on some property of the entities in the sampled functional data. Clustering is often carried out by utilizing a warping function that warps the chronological time of each curve into a time dimension that is the same across all the functional observations (stories, in our case). Then we measure the similarity between different fictional stories and use this information to create clusters of homogeneous sentiment curves.

Functional data clustering usually includes two steps: warping and clustering. The warping, a form of preprocessing, is an operation that transforms a collection of curves onto a standard comparable scale. Clustering places the curves into different groups, often based on some distance metric. Functional data clustering uses a distance metric such as $\int_{t\in[0,1]} \big(f_i(t) - f_j(t)\big)^2\, dt$ to measure the dissimilarity between curves i and j. In practice, these curves $\mathcal{F} = \{f_i(t), i = 1, 2, \ldots, n, t \in [0, 1]\}$ are often represented in a discretized way by fixed length vectors with a common set of measurement times $t_1, \ldots, t_n$.

To actually perform warping, usually we determine a "center curve" (template) first. The optimal warping for two curves is then found by estimating some nonlinear transformations (called warping functions, denoted, say $\gamma(\cdot)$) having the property that $\gamma \in \Gamma = \{\gamma : [0, 1] \to [0, 1], \gamma(0) = 0, \gamma(1) = 1\}$. An basic example of such a warping function could be $\gamma(t \mid \beta) = \frac{e^{\beta t}-1}{e^{\beta}-1}$ for some parameter β and time $t \in [0, 1]$ from one template curve, such that $\gamma(t \mid \beta)$ is the warped time that transforms the time domain of another curve. Usually one template curve is chosen from the curves and then other curves are aligned with respect to this template. Thus the warping results are based on the choice of central curve and are not unique.

A newer warping method is the Fisher-Rao (Srivastava et al, 2011) warping, which uses a different type of distance metric for alignment. The Fisher-Rao (Rao, 1945) distance metric is defined such that for any $f_1, f_2 \in \mathcal{F}$ and $\gamma \in \Gamma$, $d_{FR}\big((f_1 \circ \gamma), (f_2 \circ \gamma)\big) = d_{FR}\big(f_1, f_2\big)$. In the quotient space, the Fisher-Rao distance is represented by the Square Root Velocity Function (SRVF) defined as

$$q(t) = \big\{\dot{f}(t)/\sqrt{|\, \dot{f}(t)\,|} \text{ when } \dot{f}(t) \neq 0 \text{ and } 0 \text{ when } \dot{f}(t) = 0\big\},$$

where the dot defines the derivative operation. Under the SRVF framework, $q(t)$ is the first order derivative of $f(t)$, standardized and scaled by dividing by the square root of its absolute value, $\sqrt{|\, \dot{f}(t)\,|}$. In other words, $q(t)$ can be represented as $q(t) = \text{sign}(f'(t))\sqrt{|\, f'(t)\,|}$, where $\text{sign}(u) = 1$ if $u \geq 0$, and -1 if $u < 0$. Then we denote the SRVF of $f \circ \gamma(t)$, where $f \circ \gamma(t) \coloneqq f\big(\gamma(t)\big)$, as:

$$(q, \gamma)(t) = \text{SRVF}(f \circ \gamma(t)) = \frac{\frac{\mathrm{d}}{\mathrm{d}t} f \circ \gamma(t)}{\sqrt{|\, \frac{\mathrm{d}}{\mathrm{d}t} f \circ \gamma(t)\,|}} = (q \circ \gamma)(t)\sqrt{\dot{\gamma}(t)}$$

Then the Fisher-Rao distance for f_1 and f_2 in the SRVF representation becomes the $\mathbb{L}_2$ distance $\|q_1 - q_2\|$. One property of the SRVF representation is that for any $q_1, q_2 \in \mathbb{L}_2$, and any warping $\gamma \in \Gamma$, $\|(q_1, \gamma) - (q_2, \gamma)\| = \|q_1 - q_2\|$ (Srivastava et al, 2011). This property yields a unique center that can be used as the template for the curves chosen for alignment, thus precluding the need for manually checking and choosing one.

Thus for the set of curves $\mathcal{F} = \{f_i(t), i = 1, 2, \ldots, n\}$ and $\gamma_1, \gamma_2, \ldots, \gamma_n \in \Gamma$, instead of finding a template curve, say, f_1 and solving $\inf_{\gamma \in \Gamma} \|(f_i \circ \gamma_i) - f_1\|$, the Fisher-Rao metric utilizes an objective function to find the Karcher mean (Karcher, 1977), which is defined as:

$$\hat{u}_q = \arg \inf_{u_q \in U_q} \sum_{i=1}^{N} \inf_{\gamma_i \in \Gamma} \|u_q - (q_i, \gamma_i)\|_2^2$$

With respect to functional K-means clustering (Sangalli et al, 2010), assume we have k template curves $\boldsymbol{\mu} = \{\mu_1(t), \ldots, \mu_k(t)\} \subseteq \mathcal{F}$ representing the "centers" of k clusters. We can then define the domain of attraction for each template $\mu_j \equiv \mu_j(t)$ as $\lambda(\mu_j, \mathcal{F}) = \{f \in \mathcal{F} : \sup_{\gamma \in \Gamma} \rho(\mu_j, f \circ \gamma) \geq \sup_{\gamma \in \Gamma} \rho(\mu_r, f \circ \gamma), \forall r \neq j, j = 1, \ldots, k\}$, i.e., the curves that are closer to template μ_j than to other templates. Hereafter, the objective function to optimize could be defined as $\sum_{j=1}^{k} \sum_{f_i \in \lambda(\mu_j, \mathcal{F})} \rho(\mu_j, f_i \circ \gamma_i)$, which is similar to the objective function of standard k-means clustering.

A simple way of finding the optimal number of clusters would then be dividing the sum of between-cluster distances by the number of clusters k for a grid of values of k, and choosing the number of clusters that produces the smallest ratio. A more sophisticated metric to optimize is the average silhouette width (Rousseeuw, 1987; Batool and Hennig, 2021).

3 Collection, Preprocessing and Summary of Data

For our data selection, we chose novels from two lists of the 100 best novels ever written, in English (McCrum, R, 2015) and in all languages (McCrum, R, 2003), respectively, selected and published by *The Guardian*, selecting for analysis those which are available in the Project Gutenberg database. Project Gutenberg (Stroube, 2003) is a volunteer effort that has digitized and stored over 60,000 e-books, including most English novels written since 1700. This resulted in a dataset of 62 well-known novels for analysis. We list the name, author, main character, and the year of publication of these novels in Appendix A, with the first ten examples summarized in Table 1. Novels published after the 1920s are not included in Project Gutenberg for copyright reasons, since these more recent works are not in the public domain. Our data set contrasts with the data studied by Reagan et al (2016), who analyzed 1327 works that were "mostly, but not all, fictional stories" (Reagan et al, 2016). They also used the Project Gutenberg database and selected their sample

based on criteria such as length, language (English), and number of downloads, whereas our smaller sample was specifically chosen to represent work considered by *The Guardian* among the best novels ever.

Table 1 Summary of the first ten selected novels.

Year	Title	Author	Main Character	First Tense
1612	Don Quixote	Miguel De Cervantes	Don Quixote	No
1678	Pilgrim's Progress	John Bunyan	None	Yes
1719	Robinson Crusoe	Daniel Defoe	Robinson Crusoe	Yes
1726	Gulliver's Travels	Jonathan Swift	Lemuel Gulliver	Yes
1749	Tom Jones	Henry Fielding	Tom Jones	No
1759	Tristram Shandy	Laurence Sterne	Tristram Shandy	Yes
1782	Dangerous Connections	Pierre Choderlos De Laclos	Epistolary format	Yes
1816	Emma	Jane Austen	Emma Woodhouse	No
1818	Frankenstein	Mary Shelley	Epistolary format	Yes
1818	Nightmare Abbey	Thomas Love Peacock	Scythrop	No

The downloaded files were in plain text format, so we preprocessed them before applying the sentiment analysis model. The preprocessing of text data is different from that of numerical data. It must transfer unstructured information into a machine-readable format, such as word vectors or frequency lists (Mikolov et al, 2013; Ramos et al, 2003). It also involves techniques such as word segmentation (Saffran et al, 1996) since documents need to be segmented into tokens first in order to count frequencies or transform to numerical vectors; named-entity recognition (NER) (Lample et al, 2016); stop-word filtering (Saif et al, 2014); and stemming (reducing words like go, went and going to their common stem). Our preprocessing steps were as follows: We first transformed all texts into a unicode format to eliminate unsupported special string types. We then removed control characters such as "\t", as well as other invalid characters, and we regularized the spacing delimiters. We separated the novel into ordered pieces by using the Python package `NLTK` (Loper and Bird, 2002), which uses an unsupervised algorithm to find the boundaries among sentences. After the novels were separated into sentences, we used the Google API-based model to infer sentiment (Google, 2023), using the predefined dictionary and other downstream tasks provided by Google. Table 2 shows an example of these extracted sentence sentiment values. The sentiment score (which takes values between -1 and 1) measures the overall sentiment, whereas the sentiment magnitude (which takes values between 0 and 1) measures the amount of emotional content. We used the sentiment magnitude times sentiment score as our sentiment variable, to identify the clearly positive and clearly negative sentences. Recall that our sentiments were analyzed based on each sentence, and the overall sentiment was calculated as a function of the sentence sentiments.

The mean length of the 62 novels we analyzed is 7,173 sentences, with an average sentence length of 23.15 words. The longest novel has nearly half a million words, while the shortest one has around 27,000 words. Among all novels, the longest sentence has a length of 256 words, while the shortest sentences have only one word. Those short word sentences usually appeared in

Table 2 Example of sentence-based sentiment, for the Jane Austen novel *Emma*. The sentiment score measures the overall sentiment, whereas the sentiment magnitude measures the amount of emotional content. The sentiment magnitude times sentiment score was our sentiment variable, from which we determined clearly positive and clearly negative sentences.

Sentence	Emma Woodhouse, handsome, clever, and rich, with a comfortable home and happy disposition, seemed to unite some of the best blessings of existence; and had lived nearly twenty-one years in the world with very little to distress or vex her.
Sentiment	{sentiment score: 0.70, sentiment magnitude: 0.70 }

dialogue in forms such as “What?”, “Yes.”, etc. For the long sentences that are over 200 words, usually these appear in some narrative about the environment or due to someone talking at length. Figure 1 shows an histogram for the sentence length among the studied novels. The average sentence length for each novel is displayed with the markings along the horizontal axis. We observed that the average sentence length has become shorter over the 300 years time: The novels that have the longest average sentence length were written in the 17th or 18th century, and novels that have the shortest average sentence length were written around the 20th century.

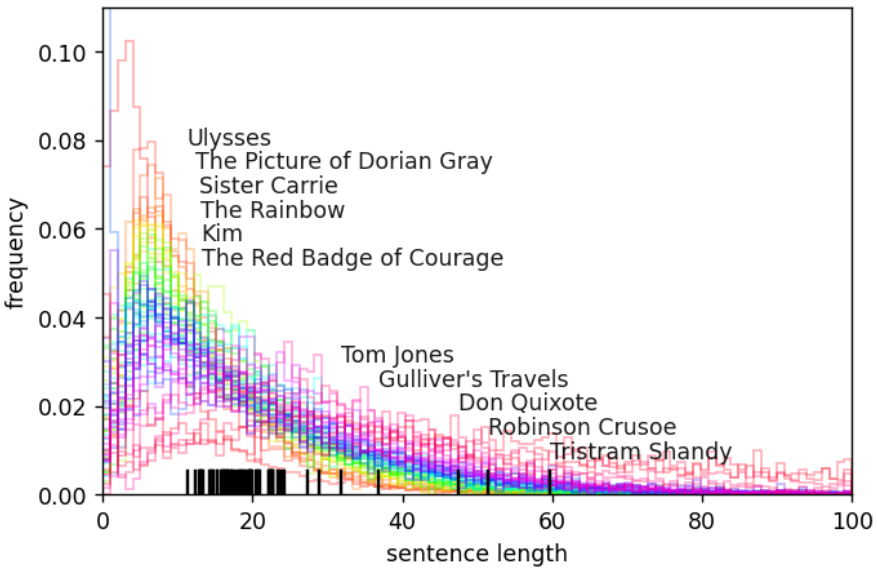

Fig. 1 Histogram of how the length of sentences was distributed among different novels, by a percentage weighting. The black marks at the bottom indicate the average sentence length for each novel. We note the five shortest and five longest novels in terms of average sentence length, indicating the trend that the average sentence length became shorter over the 300 years of time.

After the text data was transformed into numeric form, we were then able to employ familiar statistical methods.

3.1 B-spline smoothing

To smooth our raw functional data, we employed a penalized B-spline basis rather than kernel smoothing, since the B-spline approach produced more reasonable smoothing behavior, especially at the boundary regions. We first mapped our raw sentiment observations onto a grid of 500 measurement points by local averaging, creating curves of fixed length 500, and then we utilized a penalized B-spline estimator (Ruppert et al, 2003) to model the curves as:

$$Y_i(t) = f_i(t) + \epsilon_{it}, \text{ for some } f_i : [0,1] \rightarrow \mathbb{R}$$

where ϵ is independent of t with $\mathbb{E}(\epsilon) = 0$ and $\mathbb{E}(\epsilon^2) = \sigma^2$, and $f_i(t)$ was approximated with the spline basis vector $\boldsymbol{\phi}(t)$:

$$f_i(t) \approx \boldsymbol{\phi}(t)^T \boldsymbol{\beta}_i = \sum_{j=1}^{p} \beta_{ij}\phi_j(t)$$

where $\boldsymbol{\beta}_i = \beta_{i0}, \beta_{i1}, \ldots, \beta_{ip}$ are the coefficients which characterize the ith curve, and $\phi_1(t), \ldots, \phi_p(t)$ are the fixed basis functions defined in advance, generated by the Cox-de Boor recursion (De Boor, 1968) formula. We chose 100 knots (because with penalized spline methods, it is generally recommended to choose a large number of knots and achieve smoothness through the roughness penalty term) and an order-3 (cubic) B-spline basis. The penalized B-spline objective function to minimize is:

$$\hat{f}_i(t) = \arg\min_{f_i(t)} \left(\sum [Y_i - f_i(t)]^2 + \lambda \int_0^1 [f_i^{''}(t)]^2 \, dt \right)$$

where $\lambda > 0$ is the penalty term and $f_i^{''}(t)$ is the second derivative of $f_i(t)$, calculated via the `splineDesign` function in the `splines` package in R. We set λ such that the average number of local minima and maxima is around 5. The resulting smoothed curves for the 62 novels are shown in Figure 2. As we can see, the penalized B-spline approach handles the smoothing in the boundary regions of the graph well, validating the choice of B-spline smoothing to transform raw sentiment functional observations into smooth curves.

3.2 Interpreting the sentiment curves

Using data mining techniques, we illustrated the frequency of words in the novels, validated our sentiment extraction, and visualized the association between some keywords and the sentiment. We first separated the words using the `nltk` word tokenizer package in Python, then extracted keywords with a series of preprocessing techniques. We removed stopwords (from a list from `nltk`) such as "the" or "a". The term frequency inverse document frequency (TF–IDF) summarizes the frequency of occurrence across the 62 collected novels; the variant of TF-IDF we used was:

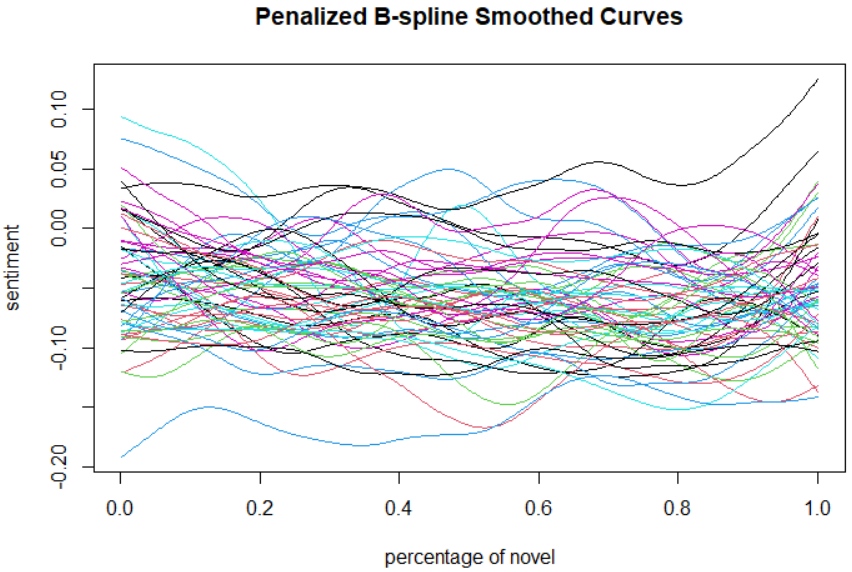

Fig. 2 The 62 sentence sentiment curves smoothed using Algorithm 1 (outlined in the Appendix) before alignment. The lengths of these curves are discretized to a fixed length of 500.

$$\text{TF-IDF}(term, novel_i) = \frac{count(term, novel_i)}{\sum_{term} count(term, novel_j)} \cdot \log\left(\frac{\text{sum of available novels}}{1+\sum_j (\text{sum of novels containing } term)} + 1\right)$$

A higher TF-IDF for a term in a novel indicates a higher importance of that term within that specific novel. In each novel we captured the words that appeared more frequently than in other novels. We removed from the frequency list words that did not appear once in any other novels or did not appear over 20 times, as well as character names and locations that were recognized using the Google API. The Pearson correlation between the standardized word frequency curve and the standardized overall sentiment curve (for the top-ranked words in the TF-IDF list) measured the association between the appearance of these higher frequency words that characterized the novel and the overall novel sentiment. We selected several feature words with the highest and lowest correlation; their summary statistics are in Table 3.

These frequencies are plotted along with the overall sentiment in Figure 3 and Figure 4. This was done by smoothing the occurrence count for these words throughout the span of the novel with a B-spline basis. The sentiment scores make sense, as for *A Passage to India*, words like “Hindu”, “club”, and “pleasant” contribute to the positive sentiment, while “British”, “evidence”, and “mistake” occurred during periods of negative sentiment. This is sensible, since a distressing incident in the novel involves a British woman mistakenly

accusing another character of an assault, which precipitates a trial. For *Alice's Adventures In Wonderland*, words like "large" and "door" are neutral, but contributed to a negative sentiment in the parts of the novel where they appear. Note that the character Alice grows large in size and encounters an impassable door in some particularly surreal and unsettling episodes in the novel.

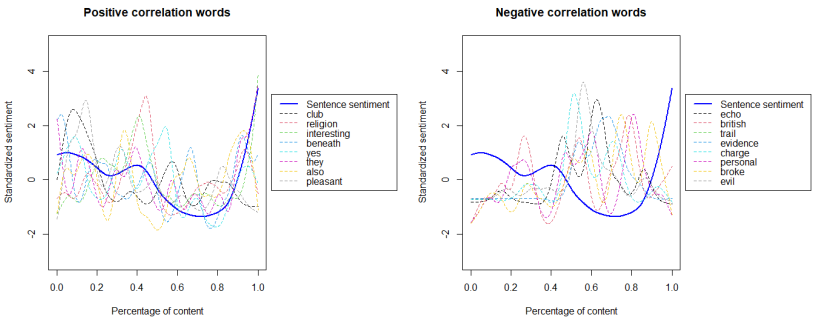

Fig. 3 The standardized sentiment curve for the novel *A Passage to India* and the selected keywords.

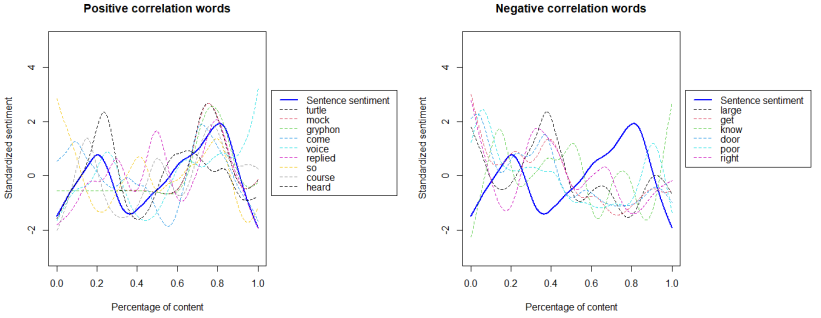

Fig. 4 The standardized sentiment curve for the novel *Alice's Adventures In Wonderland* and the selected keywords.

4 Alignment and Clustering

4.1 Warping and choosing the number of clusters

Cheng et al (2016) provided a Bayesian MCMC method to solve for the Karcher mean $u(t)$, as well as $\gamma_1(t), \gamma_2(t), \ldots, \gamma_N(t)$, by finding the posterior distribution of $\pi(u_a(t), \gamma_1(t), \gamma_2(t), \ldots, \gamma_N(t) \mid q_1, q_2, \ldots, q_N)$. The result of such warping is shown in Figure 5.

Table 3 Some high-frequency words from *A Passage to India*. TF-IDF measures how much more frequently the word appeared in the selected novel than in all 62 novels. Appearances is the number of times the word appeared in the novel. Correlation with sentiment is the Pearson correlation between the word occurrence frequency curve and the novel sentiment curve.

A Passage to India			
Keyword	TF-IDF	Appearances	Correlation with sentiment
echo	0.000238	26	-0.68
British	0.000238	26	-0.60
evidence	0.000176	24	-0.80
mistake	0.000162	24	-0.59
Keyword	TF-IDF	Appearances	Correlation with sentiment
Hindu	0.000537	23	0.30
club	0.000455	57	0.38
Professor	0.000579	38	0.57
pleasant	0.000180	26	0.65

Alice's Adventures In Wonderland			
Keyword	TF-IDF	Appearances	Correlation with sentiment
large	0.000827	33	-0.47
door	0.000752	30	-0.33
poor	0.000627	25	-0.34
Keyword	TF-IDF	Appearances	Correlation with sentiment
Turtle	0.00652	58	0.46
Come	0.000551	21	0.39
voice	0.00120	48	0.35

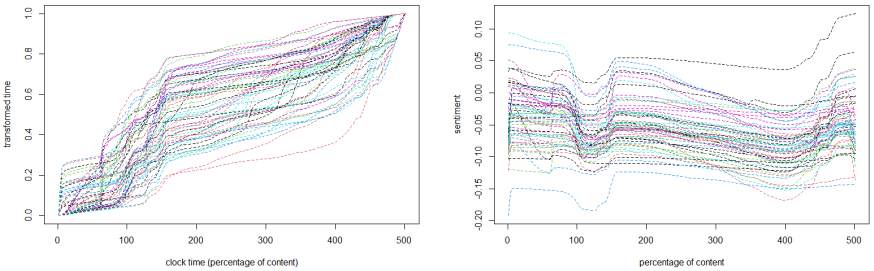

Fig. 5 Warping functions fitted by minimizing distance to the Karcher mean. Left panel: the original time on the x-axis and the transformed time on the y-axis. Right panel: warped curves via Bayesian SRVF warping by minimizing distance to the Karcher mean.

Directly minimizing the distance to the Karcher mean often oversmooths, and thus Srivastava et al (2011) suggested adding a penalizing term to find a proper γ_i, with $\mathcal{R}$ a smoothness penalty to keep the warping function close to $\gamma_i(t) = t$:

$$\arg \inf_{\gamma_i \in \Gamma} \left(\|u_Q - (q_i, \gamma_i)\|_2^2 + \lambda \mathcal{R}(\gamma_i) \right)$$

where this $\lambda\mathcal{R}(\gamma_i)$ term (Wu and Srivastava, 2011) is chosen as:

$$\lambda \int_{t=0}^{1} \mid 1 - \dot{\gamma}_i(t)^{1/2} \mid^2 \, dt.$$

Here $\dot{\gamma}_i(t)$ stands for the first derivative term of the warping function $\gamma_i(t)$.

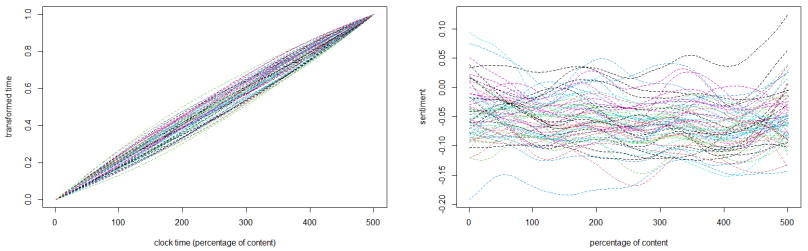

Fig. 6 Warping functions and warped curves under $\lambda = 0.1$

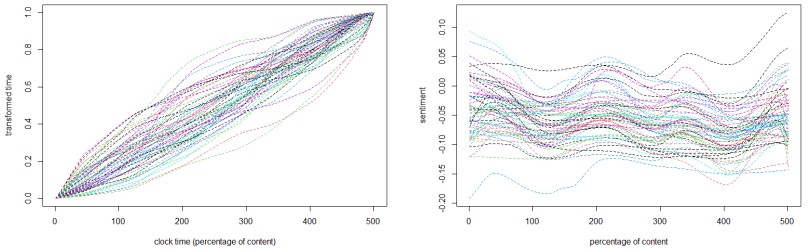

Fig. 7 Warping functions and warped curves under lambda = 0.01

We chose the value $\lambda = 0.01$, which produced a more effective alignment among a variety of values of λ (see Figures 6 and 7 for a comparison of $\lambda = 0.1$ and $\lambda = 0.01$). The partitioning of the aligned curves into clusters can be performed by minimizing metrics like the L_2 norm, where $\|f_1(t), f_2(t)\|^2 = \sqrt{\int_{t=0}^{1}[f_1(t) - f_2(t)]^2\, dt}$:

$$\arg \sum_{i=1}^{k} \inf_{u_i \in U} \sum_{j \in \text{ cluster}_i} \inf_{a,b \in \mathbb{R}} \|(f_j(t), u(t)\|_2^2$$

or by maximizing a functional version of the Pearson correlation ρ, where $\rho(f_1(t), f_2(t)) = \frac{\int_{t=0}^{1} f_1(t) f_2(t)\, dt}{\|f_1(t)\|^2 \|f_2(t)\|^2}$:

$$\arg \sum_{i=1}^{k} \max_{u_i \in U} \sum_{j \in \text{cluster}_i} \max_{a,b \in \mathbb{R}} \rho(f_j(t), u(t)).$$

Note that here $f_1(t), f_2(t), \ldots, f_N(t)$ are warped curves. We used functional K-means clustering, implemented using the R package `kma`. To choose the optimal number of clusters, we plotted the percentage of variance explained based on different numbers of clusters, as Figure 8 shows. Note that since the cluster calculated by gradient-based methods, for each number of clusters k, we ran the algorithm 10 times and selected the one with the highest percentage of variance explained or the largest within-cluster Pearson correlation. One common method in determining the optimal number of clusters is the elbow method: We can see in our example that when we have 5 clusters, adding another cluster does not yield much improvement, so we consider the optimal number of clusters to be 5.

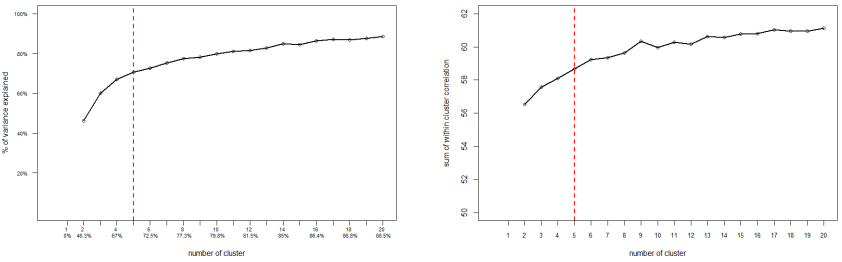

Fig. 8 Optimal number of clusters under $\lambda = 0.01$, using the L_2 distance and Pearson correlation metric.

4.2 Clustering

To carry out the clustering of our sentiment curves, we employed the following objective function, a variation on the functional K-means objective function that permits clustering and warping at the same time:

$$\arg \sum_{i=1}^{k} \inf_{u_i \in U_q} \sum_{i \in \text{cluster}_j} \Big(\inf_{\gamma_i \in \Gamma} \|u_i - (q_i, \gamma_i)\|_2^2 + \lambda \int_{t=0}^{1} |\, 1 - \dot{\gamma_i}(t)^{1/2} \,|^2 \; dt \Big)$$

where

$$(q,\gamma)(t) = \mathrm{SRVF}(f\circ\gamma(t)) = \frac{\frac{\mathrm{d}}{\mathrm{d}t}f\circ\gamma(t)}{\sqrt{|\frac{\mathrm{d}}{\mathrm{d}t}f\circ\gamma(t)|}} = (q\circ\gamma)(t)\sqrt{\dot{\gamma}(t)}$$

Figure 9 shows the curves (after warping), separated by cluster, and Figure 10 shows each functional observation's warping function, again separately by cluster:

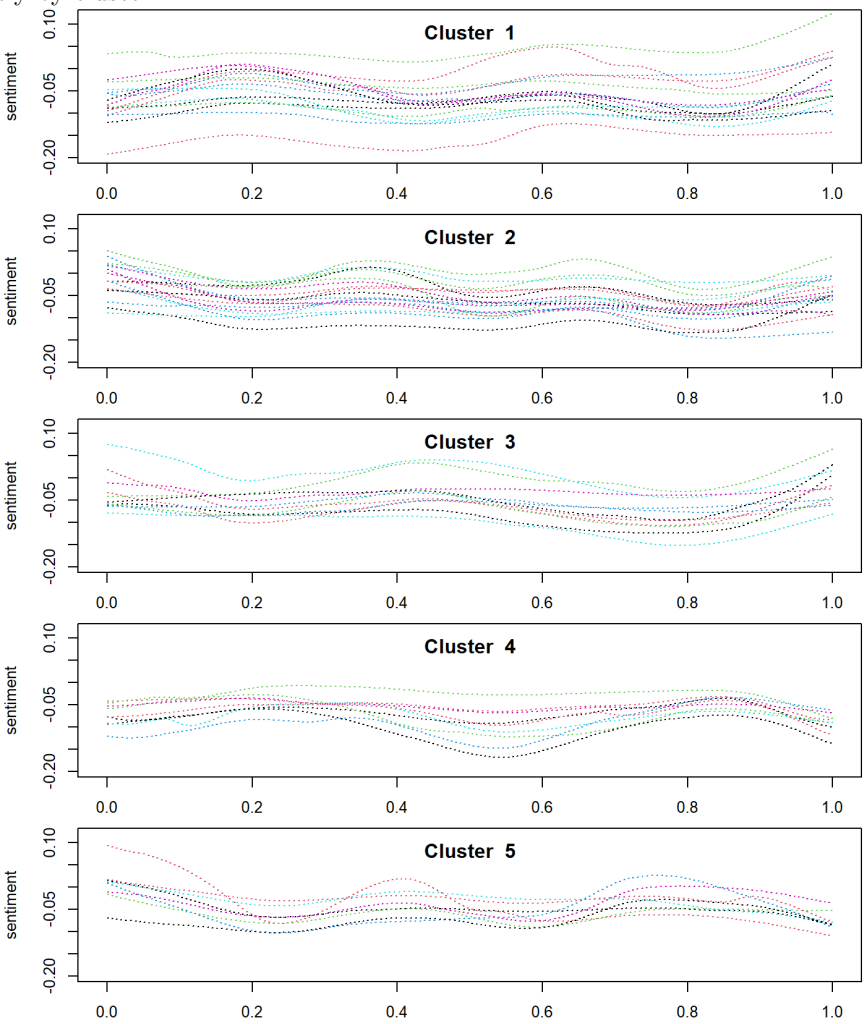

Fig. 9 The curves after warping, within each cluster, using k-means clustering with $k = 5$, $\lambda = 0.01$.

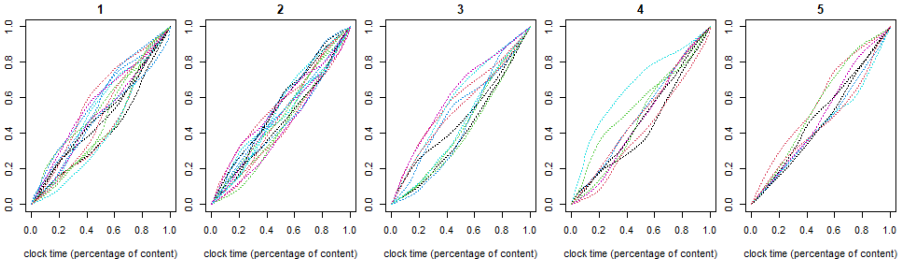

Fig. 10 The warping functions within each cluster, using k-means clustering with $k = 5$, $\lambda = 0.01$.

4.3 Interpretations of the clusters

Figures 11 and 12 display the cluster memberships of the 62 novels in terms of the five clusters. We now attempt to draw some broad conclusions about the natures of the five clusters. (Some plot spoilers follow about a few of the novels.)

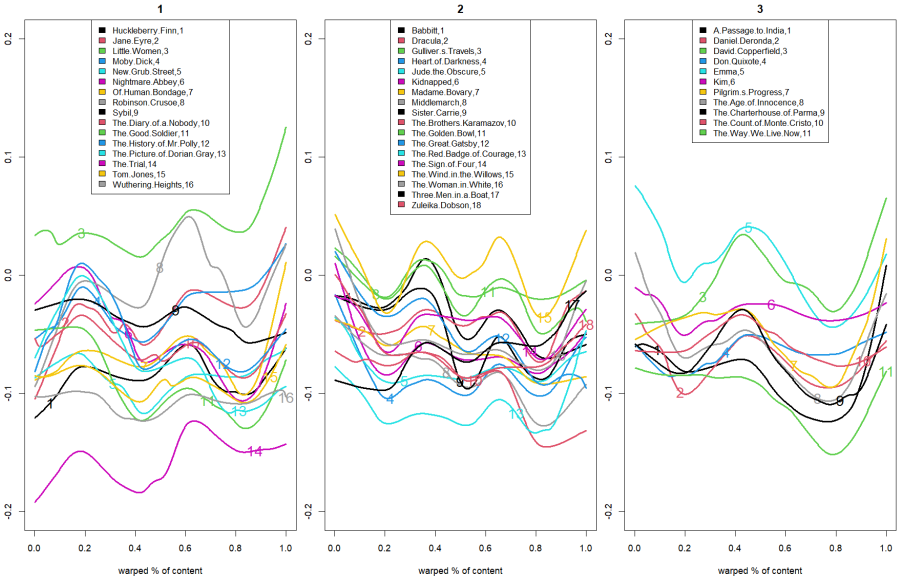

Fig. 11 The cluster memberships for the sentiment curves placed into Clusters 1, 2, and 3.

Cluster 1 contains novels that for most of the story, exhibit consistently low sentiment values, before the sentiment is abruptly improved at the end. *The Good Soldier* and *Wuthering Heights* are notable Cluster 1 novels with pessimistic and dark themes. The exception in Cluster 1 to the "late improvement" characteristic is *The Trial*, which is certainly among the most consistently negative novels in sentiment overall, as anyone who has read it will agree. Like *A Passage to India* from Cluster 3, *The Trial* features a major character being

accused of a crime, but without the exoneration at the end. An exception to the overall "low sentiment" curves in Cluster 1 is *Little Women*, perhaps the highest-sentiment novel in the entire sample, which is consistently positive yet still shows an increase at the end.

The cluster containing the most novels (18) is Cluster 2, and most of these have sentiment curves that remain nearly constant, and several of them have improvements in sentiment at the end. For example, *The Red Badge of Courage* is basically negative in sentiment for most of the novel, reflecting the shame that Henry Fleming feels at deserting his battalion during a Civil War battle. Near the end, he is wounded and this "red badge of courage" enables him to regain his pride and fight honorably, which is seen in the late rise in sentiment.

Cluster 3 contains mostly "w"-shaped curves which see the sentiment worsen, improve, worsen again, and finally improve over the course of the novel. An example of this type of novel which tends to have basically positive sentiment in general is Jane Austen's *Emma*, which is plotted at the top of the Cluster 3 curves. The up-and-down sentiment of this novel reflects Emma's social misunderstandings and unintentional insults, which she tries to recover from. Finally, marriage brings her happiness and improves the sentiment near the end of the novel. Another classic example of this pattern (except without the initial sentiment decline) is *David Copperfield*, in which David's idyllic early childhood is interrupted when his mother remarries a cruel stepfather. David experiences many difficulties during the novel, including the treachery of Uriah Heep, but at the end he finds happiness with his true love. *A Passage to India* shows a rather asymmetric w-shaped pattern: For most of the novel, Dr. Aziz is wrongly accused of a crime and tensions between natives of India and the British colonists are boiling. At the end of the novel, Aziz is exonerated and sentiment sharply improves.

Cluster 4 is a "unhappy endings" cluster, in which most of its novels display worsening sentiment at the end of the story. Cluster 4 includes a novel with one of the most famous unhappy endings in all of literature, *Anna Karenina*, as Leo Tolstoy's doomed heroine is overcome with despair. Cluster 5 also features novels with worsening sentiment at the end, although less dramatically so than in Cluster 4. Most of the sentiment curves in Cluster 5 are fairly stable, which the exception of *Frankenstein*, which from an initial high sentiment, quickly plunges into the negative range shared by the other curves in this cluster.

Overall, we see that Vonnegut's idea that literature tends to follow a few patterns is reflected fairly well in this cluster analysis. While our sample of novels from the 1600s through the early 1900s does not reflect the entire scope of literature, we do see in this sample several major plot profiles, including the dramatic shifts in sentiments seen in Cluster 3; the more subtle variation in sentiment in Clusters 1 and 2, and the mostly sad-ending "tragedies" in Cluster 4.

Vonnegut himself posited seven or eight "shapes of stories"; (Johnson, Stephen (2022), among other articles, reproduces a graphic by graphic designer Maya Eilam depicting the shapes) in his lecture, which he described with

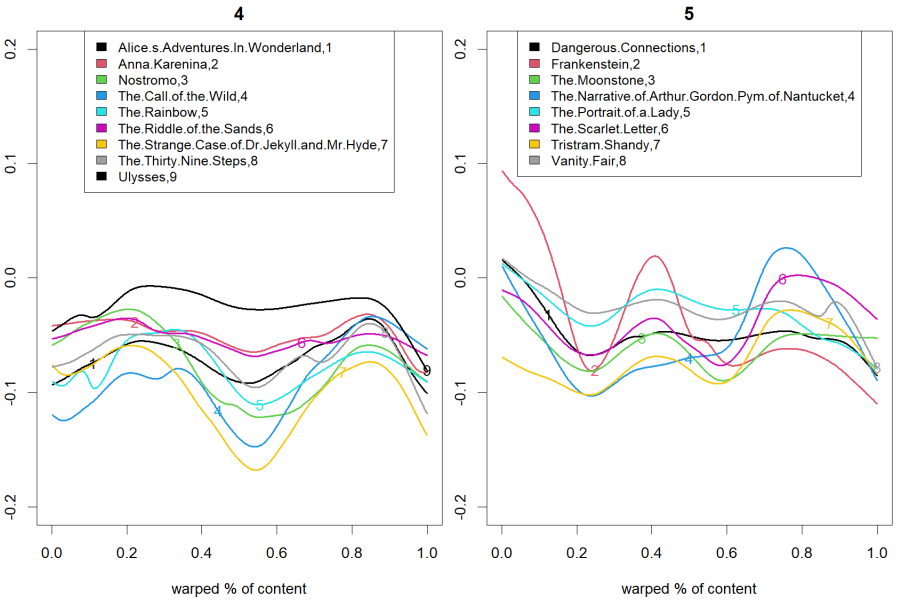

Fig. 12 The cluster memberships for the sentiment curves placed into Clusters 4 and 5

vivid names. His "Man in Hole" and "Boy Meets Girl" categories, with their oscillations in sentiment preceding a final happy ending, share characteristics with our Clusters 1 and 2. His "From Bad to Worse" and "Old Testament" categories with their unhappy endings, are similar to our Clusters 4 and 5. His categories of "Which Way Is Up?" (stories having total ambiguity) and "Creation Story" (increasingly positive sentiment/growth) have few exemplars in our data set, except perhaps the outlier *Little Women* in the latter category. Finally, Vonnegut's "New Testament" and "Cinderella" categories are in fact the same shape, and their late downfalls followed by a last-minute joyful recovery certainly mirrors our Cluster 3.

We can also compare our clustering results to those of Reagan et al (2016), who found six clusters in their data set and characterized them, following Vonnegut's terminology, as: "Rags to riches" (rise); "Tragedy/Riches to rags" (fall); "Man in a hole" (fall-rise); "Icarus" (rise-fall); "Cinderella" (rise-fall-rise); "Oedipus" (fall-rise-fall). We see that several of their categories, though not all, share shapes with our obtained clusters. Some possible reasons for the discrepancies in our results include the fact that our data set includes novels specifically chosen as the best ever, while their data set is a much broader collection of works; in addition, our representations of the smoothed sentiment functions appears to include cluster mean functions with more turning points than theirs, so that clusters with more complicated sentiment curves, such as the w-shaped functions of our Cluster 3, can be included.

5 Future Research

Future work includes investigating the potential use of our functional clustering results for book recommendation or story generation. One possible area of research is the use of these patterns to simulate new short stories with similar patterns which adhere to the sentiment curve categories we identified in our cluster analysis. This could potentially be done by using the high frequency entity words captured from the novels, along with the aligned sentiment curves, to generate a template storyline using artificial intelligence. One thing that current story generation models suffer from is the inability to adhere to coherence in a long range (Guan et al, 2020). The techniques of text summarization also may be explored further; see Allahyari et al (2017) for a brief survey on summarization. Another area of future methodological research is finding the optimal number of clusters (and performing the simultaneous warping and clustering) through the Dirichlet Process Mixture (DPM) model, which can adaptively learn the number of clusters from the data.

Appendix A Appendix

Algorithm 1 Algorithm to generate B-spline basis

Input: **K** : **int**, number of knots; **ord** : **int**, order of curves
$t \leftarrow sequence(0, 1, length = 501)$, equal step sequence from 0 to 1 with length 501
$u \leftarrow [0 : K]/K$, separate the t by K equal knots
$u \leftarrow [repeat(0, ord), u, repeat(1, ord)]$, pad the knots by the order ord to handle the boundary issues
$\mathbf{p} \leftarrow K + ord$, p is then the number of basis functions
for j **in** (0 to p) **do**
  if $u[j] \leq t \leq u[j+1]$ **then** $N[j][0](t) \leftarrow 1$, **else** $N[j][0](t) \leftarrow 0$
end for
for r **in** (1 to ord) **do**
  for j **in** (0 to p) **do**
    $N[j][r](t) \leftarrow \frac{t-u[j]}{u[j+r]-u[j]}\ N[j][r-1](t) + \frac{u[j+r+1]-t}{u[j+r+1]-u[j+1]}\ N[j+1][r-1](t)$
  end for
end for
for j **in** (0 to p) **do**
  $\phi_j(t) \leftarrow N[j][r](t)$
end for
Output: $\phi(t)$

Table A1 Full list of selected Novels

Year	Title	Author	Main Character
1612	Don Quixote	Miguel De Cervantes	Don Quixote
1678	Pilgrim's Progress	John Bunyan	None
1719	Robinson Crusoe	Daniel Defoe	Robinson Crusoe
1726	Gulliver's Travels	Jonathan Swift	Lemuel Gulliver
1749	Tom Jones	Henry Fielding	Tom Jones
1759	Tristram Shandy	Laurence Sterne	Tristram Shandy
1782	Dangerous Connections	Pierre Choderlos De Laclos	Epistolary format
1816	Emma	Jane Austen	Emma Woodhouse
1818	Frankenstein	Mary Shelley	Epistolary format
1818	Nightmare Abbey	Thomas Love Peacock	Scythrop
1838	The Narrative of Arthur Gordon Pym of Nantucket	Edgar Allan Poe	Arthur Gordon
1839	The Charterhouse of Parma	Stendhal	Fabrizio del Dongos
1844	The Count of Monte Cristo	Alexandre Dumas	Edmond Dantès
1845	Sybil	Benjamin Disraeli	Sybil
1847	Jane Eyre	Charlotte Brontë	Jane Eyre
1847	Wuthering Heights	Emily Brontë	Heathcliff
1848	Vanity Fair	William Makepeace Thackeray	Becky Sharp
1850	David Copperfield	Charles Dickens	David Copperfield
1850	The Scarlet Letter	Nathaniel Hawthorne	Hester Prynne
1851	Moby-Dick	Herman Melville	Ahab
1856	Madame Bovary	Gustave Flaubert	Emma Bovary
1859	The Woman in White	Wilkie Collins	Walter Hartright
1865	Alice's Adventures In Wonderland	Lewis Carroll	Alice
1868	Little Women	Louisa M. Alcott	Jo March
1868	The Moonstone	Wilkie Collins	Rachel Verinder
1871	Middlemarch	George Eliot	Dorothea Brooke
1875	The Way We Live Now	Anthony Trollope	Augustus Melmotte
1876	Daniel Deronda	George Eliot	Daniel Deronda
1878	Anna Karenina	Leo Tolstoy	Anna Karenina
1879	The Brothers Karamazov	Fyodor Dostoevsky	Fyodor Pavlovitch Karamazov
1881	The Portrait of a Lady	Henry James	Isabel Archer
1884	Huckleberry Finn	Mark Twain	Huckleberry Finn
1886	The Strange Case of Dr Jekyll and Mr Hyde	Robert Louis Stevenson	Utterson Gabriel John
1889	Three Men in a Boat	Jerome K. Jerome	Jerome
1890	The Picture of Dorian Gray	Oscar Wilde	Dorian Gray
1890	The Sign of Four	Arthur Conan Doyle	Sherlock Holmes
1891	New Grub Street	George Gissing	Jasper Milvain
1892	The Diary of a Nobody	George Grossmith	Charles Pooter
1893	Kidnapped	Robert Louis Stevenson	David Balfour
1895	Jude the Obscure	Thomas Hardy	Jude Fawley
1895	The Red Badge of Courage	Stephen Crane	Henry Fleming
1897	Dracula	Bram Stoker	Abraham Van Helsing
1899	Heart of Darkness	Joseph Conrad	Charles Marlow
1900	Sister Carrie	Theodore Dreiser	Carrie
1901	Kim	Rudyard Kipling	Kimball "Kim" O'Hara
1903	The Call of the Wild	Jack London	Buck
1903	The Riddle of the Sands	Erskine Childers	Carruthers
1904	Nostromo	Joseph Conrad	Nostromo
1904	The Golden Bowl	Henry James	Maggie Verver
1908	The Wind in the Willows	Kenneth Grahame	Mole
1910	The History of Mr Polly	HG Wells	Alfred Polly
1911	Zuleika Dobson	Max Beerbohm	Zuleika Dobson
1915	Of Human Bondage	W Somerset Maugham	Philip Carey
1915	The Good Soldier	Ford Madox Ford	John Dowell
1915	The Rainbow	D. H. Lawrence	Brangwens Family
1915	The Thirty-Nine Steps	John Buchan	Richard Hannay
1920	The Age of Innocence	Edith Wharton	Newland Archer
1922	Babbitt	Sinclair Lewis	George F. Babbitt
1922	Ulysses	James Joyce	Molly Bloom
1924	A Passage to India	EM Forster	Dr. Aziz
1925	The Great Gatsby	F. Scott Fitzgerald	Jay Gatsby
1925	The Trial	Franz Kafka	Józef K.

References

Allahyari M, Pouriyeh S, Assefi M, et al (2017) Text summarization techniques: a brief survey. **arXiv preprint:** at https://doi.org/10.48550/arXiv.1707.02268

Batool F, Hennig C (2021) Clustering with the average silhouette width. Computational Statistics and Data Analysis 158:107,190

Cheng W, Dryden IL, Huang X (2016) Bayesian registration of functions and curves. Bayesian Analysis 11(2):447–475

De Boor C (1968) On uniform approximation by splines. Journal of Approximation Theory 1(1):219–235

Dhingra B, Li L, Li X, et al (2016) Towards end-to-end reinforcement learning of dialogue agents for information access. **arXiv preprint:** at https://doi.org/10.48550/arXiv.1609.00777

Fan A, Lewis M, Dauphin Y (2018) Hierarchical neural story generation. **arXiv preprint:** at https://doi.org/10.48550/arXiv.1805.04833

Google (2023) Analyzing sentiment. https://cloud.google.com/natural-language/docs/analyzing-sentiment

Guan J, Huang F, Zhao Z, et al (2020) A knowledge-enhanced pretraining model for commonsense story generation. Transactions of the Association for Computational Linguistics 8:93–108

Hitchcock DB, Greenwood MC (2015) Clustering functional data. In: Hennig C, Meila M, Murtagh F, et al (eds) Handbook of Cluster Analysis. CRC Press, Boca Raton, p 265–288

Jacques J, Preda C (2014a) Functional data clustering: a survey. Advances in Data Analysis and Classification 8(3):231–255

Jacques J, Preda C (2014b) Model-based clustering for multivariate functional data. Computational Statistics and Data Analysis 71:92–106

Johnson, Stephen (2022) Kurt Vonnegut on the 8 “shapes” of stories. https://bigthink.com/high-culture/vonnegut-shapes

Karcher H (1977) Riemannian center of mass and mollifier smoothing. Communications on Pure and Applied Mathematics 30(5):509–541

Kokoszka P, Reimherr M (2017) Introduction to Functional Data Analysis. Chapman and Hall/CRC

Kurt Vonnegut (2010) Kurt Vonnegut on the shapes of stories. https://www.youtube.com/watch?v=oP3c1h8v2ZQ

Lample G, Ballesteros M, Subramanian S, et al (2016) Neural architectures for named entity recognition. **arXiv preprint:** at https://doi.org/10.48550/arXiv.1603.01360

Liu X, Yang MC (2009) Simultaneous curve registration and clustering for functional data. Computational Statistics and Data Analysis 53(4):1361–1376

Loper E, Bird S (2002) Nltk: The natural language toolkit. **arXiv preprint:** at https://doi.org/10.48550/arXiv.cs/0205028

Madigan D, Genkin A, Lewis DD, et al (2005) Author identification on the large scale. In: Proceedings of the 2005 Meeting of the Classification Society of North America (CSNA)

McCrum, R (2003) The 100 greatest novels of all time: The list. https://www.theguardian.com/books/2003/oct/12/features.fiction

McCrum, R (2015) The 100 best novels written in English: the full list. https://www.theguardian.com/books/2015/aug/17/the-100-best-novels-written-in-english-the-full-list

Mikolov T, Chen K, Corrado G, et al (2013) Efficient estimation of word representations in vector space. **arXiv preprint:** at https://doi.org/10.48550/arXiv.1301.3781

Ramos J, et al (2003) Using tf-idf to determine word relevance in document queries. In: Proceedings of the First Instructional Conference on Machine Learning, New Jersey, USA, pp 133–142

Ramsay JO, Silverman BW (2005) Functional Data Analysis. Springer

Rao CR (1945) Information and the accuracy attainable in the estimation of statistical parameters. Bulletin of the Calcutta Mathematical Society 37:81–91

Reagan AJ, Mitchell L, Kiley D, et al (2016) The emotional arcs of stories are dominated by six basic shapes. EPJ Data Science 5(1):1–12

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20:53–65

Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric Regression. 12, Cambridge University Press

Saffran JR, Newport EL, Aslin RN (1996) Word segmentation: The role of distributional cues. Journal of Memory and Language 35(4):606–621

Saif H, Fernandez M, He Y, et al (2014) On stopwords, filtering and data sparsity for sentiment analysis of twitter. In: Proceedings of LREC 2014,

Ninth International Conference on Language Resources and Evaluation, pp 810–817

Sangalli LM, Secchi P, Vantini S, et al (2010) K-mean alignment for curve clustering. Computational Statistics and Data Analysis 54(5):1219–1233

Sohail SS, Siddiqui J, Ali R (2017) A novel approach for book recommendation using fuzzy based aggregation. Indian Journal of Science and Technology 8(1)

Srivastava A, Wu W, Kurtek S, et al (2011) Registration of functional data using Fisher-Rao metric. **arXiv preprint:** at https://doi.org/10.48550/arXiv.1103.3817

Stroube B (2003) Literary freedom: Project Gutenberg. XRDS: Crossroads, The ACM Magazine for Students 10(1):3–3

Swanson, Ana (2015) Kurt Vonnegut graphed the world's most popular stories. https://www.washingtonpost.com/news/wonk/wp/2015/02/09/kurt-vonnegut-graphed-the-worlds-most-popular-stories/

Tarpey T (2007) Linear transformations and the k-means clustering algorithm: applications to clustering curves. The American Statistician 61(1):34–40

Wu W, Srivastava A (2011) An information-geometric framework for statistical inferences in the neural spike train space. Journal of Computational Neuroscience 31(3):725–748