The Shape of Stories

Do stories have a shape?

The question begets two further questions.

What is a story?

In his *Quantitative Narrative Analysis* (2010:11) Franzosi asks:

> But what is a story? A story, according to the Oxford English Dictionary (OED, 1989) is
> A narrative, true or presumed to be true, relating to important events and celebrated persons of a more or less remote past; a historical relation or anecdote.
> A recital of events that have or are alleged to have happened; a series of events that are or might be narrated.

And what is an event?

We seem to be facing here an infinite regress. Franzosi continues (2010:17): But what

> is an event? Despite the central role of this concept in any definition of narrative, story, and plot, you will be hard-pressed to find a rigorous

> definition of event in the vast body of literature on narrative. The event almost seems to be taken as a given, as a commonsensical element of the language that does not need to be defined.
> … what narrative theorists have in mind when they talk about events is *human action*. Other types of events (and characters) can be part of narrative but only in a subordinate position (e.g., a heat wave and the effect that global warming may have on the probability of this type of event). They don't play a leading role but only a supportive one (e.g., in a story of older people having to be hospitalized during the heat wave).

5 Ws + H

A story, then, involves characters (typically, human) who act in time and space, the *scene* of the action, a text structure that can be summed up in the Who, What, When Where, Why, and How, or the 5 Ws of journalism + H.

So… What is a story shape?

A good answer would be: it's complicated. So, let's turn, once again, to the Oxford English Dictionary (OED) for help.

"External form or contour; that quality of a material object (or geometrical figure) which depends on constant relations of position and proportionate distance among all the points composing its outline or its external surface; a particular variety of this quality."

Visual and beautiful

Nearly all OED definitions of shape have this in common: a focus on the visual, perhaps on the beautiful. **So… however we define shape of a story, it must be visual, perhaps beautiful. We must be able to represent it.**

Story & plot

There is another OED definition, albeit obsolete, that is of particular interest in this context: "Form, order and arrangement (of words, etc.); course, order (of a story). Obsolete." The order of a story? There is an old distinction in narrative theory between story and plot. In his *From Words to Numbers,* Franzosi (2004: 56) thus introduces the distinction between story and plot: It is the Russian formalists who introduced the distinction between story versus plot in narrative (*fabula* versus *sjužet*). Building upon Aristotle's idea of plot-structure – mythos, in the master's own words: "[B]y this term 'plot-structure' I mean the organisation of the events" – Tomashevski (1965, p. 67) wrote: "Plot is distinct from story. Both include the same events, but in the plot the events are arranged and connected according to the orderly sequence in which they were presented in the work. [Continued in a note; the story is] "the action itself, … [the plot] how the reader learns of the action." A story, in other words, refers to a skeletal description of the fundamental events in their natural logical and chronological order (perhaps, with an equally skeletal listing of the roles of the characters in the story).

A *story* then moves the events chronologically; a *plot* scrambles them for rhetorical effect, to keep the audience interested, like Hitchcock starting his movies from the murder scene and then moving back to the antecedents.

How would you visualize the differences between story and plot? A question for a different time… But, just to give you a hint on a possible answer, it would involve a focus on the changing scenes (time & space), on the changing actors.

The consequences of action

For Aristotle, a story has a beginning, a middle, and an end. What happens in the middle determines whether the story is happy ending (a *comedy*) or unhappy ending (a *tragedy*). A tragedy no doubt is Oedipus's story who, after much tribulations, ends up unwittingly killing his father and marrying his mother (only the Greeks can think of such horrible stories, with Oedipus' mother ending up killing herself when she finds out the truth about her new relationship).

Sentiment analysis

Killing your father is not a good thing. Neither is sleeping with your mother. And we frown upon suicide. That is how we react to these events, *negatively*. Modern computational tools allow us to measure these reactions, to put a number to them. We can use those numbers to plot the unfolding of events, the ups and downs of stories. The technique is called **sentiment analysis**. To learn more about sentiment analysis, read the TIPS_NLP_Sentiment analysis.

Our "shape of stories" algorithm is fundamentally based on sentiment scores, computed via a variety of algorithms.

Story shapes at last!

"There is no reason why the simple shape of stories can't be fed into a computer. They are beautiful shapes." says Kurt Vonnegut on a widely viewed video available on YouTube. Vonnegut draws three different types of curves that he calls: "man in a hole", "boy gets girl", "Cinderella".

Reagan et al. (2016) took up Kurt Vonnegut's challenge of drawing story shapes *automatically* with a computer (a good example of Moretti's "distant reading"). They ultimately settle on six fundamental shapes:

1. Rags to Riches (rise)
2. Riches to Rags (fall)
3. Man in a Hole (fall then rise)
4. Icarus (rise then fall)
5. Cinderella (rise then fall then rise)
6. Oedipus (fall then rise then fall)

Our shape of stories tool

The NLP Suite "Shape of Stories" extraction and visualization tool is composed of Python and Java scripts combined to produce a user-friendly analytical tool. The algorithm follows Reagan et al.'s approach, with a big difference. **It is meant** for humanists and social scientists with no computational skills.

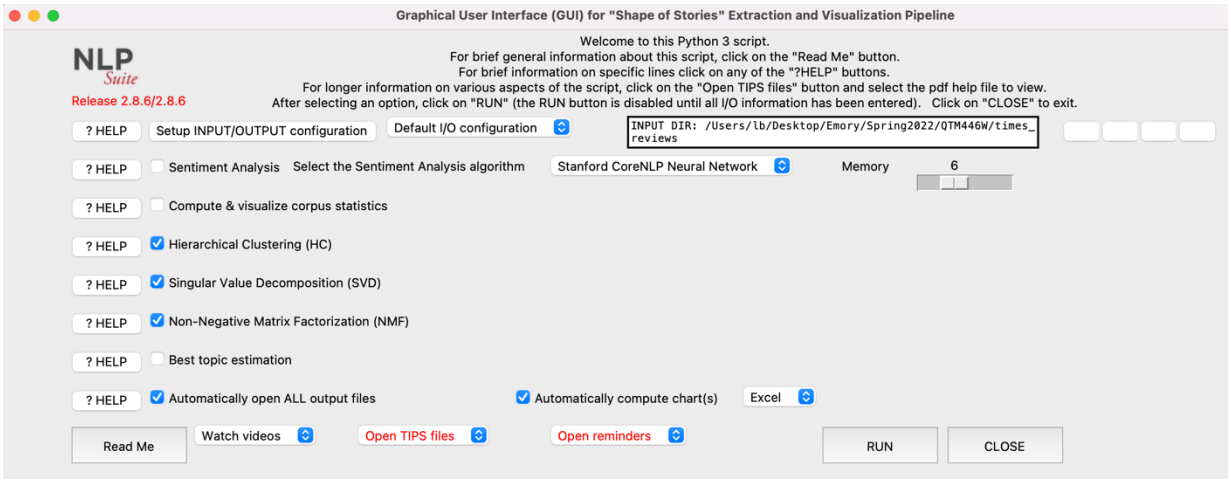
The "Shape of Stories" tool incorporates 3 different approaches to the Shape of Stories:

1. Hierarchical Clustering (HC)
2. Singular Value Decomposition (SVD)
3. Non-Negative Matrix Factorization (NMF)

Please refer to specific sections for further information on each technique. Each technique relies on previously computed sentiment score to perform the "Shape of Stories.

Graphical User Interface (GUI)

The NLP Suite GUI for the "Shape of Stories" tool is designed in the following way:



First, like other NLP Suite pipelines, it contains '? HELP' buttons that will give you access to helpful tips on the use and results of a specific script. Second, you will find scroll down menus on the lower part of the pipelines that will allow you to access any TIPS files, reminders, or videos helpful for the effective use of the tool of unexperienced users.

Sentiment Analysis

The first tick box (☑) available to the user is Sentiment Analysis. Sentiment analysis (also known as opinion mining or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level—whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive,

negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry", "sad", and "happy". (Wikipedia)

You will be able to choose from 5 different approaches for the computation of sentiment scores: the neural network approach by Stanford CoreNLP or the dictionary-based approaches by ANEW, hedonometer, SentiWordNet, and VADER. **For maximum accuracy please choose Stanford CoreNLP as your Sentiment Analysis Algorithm.**

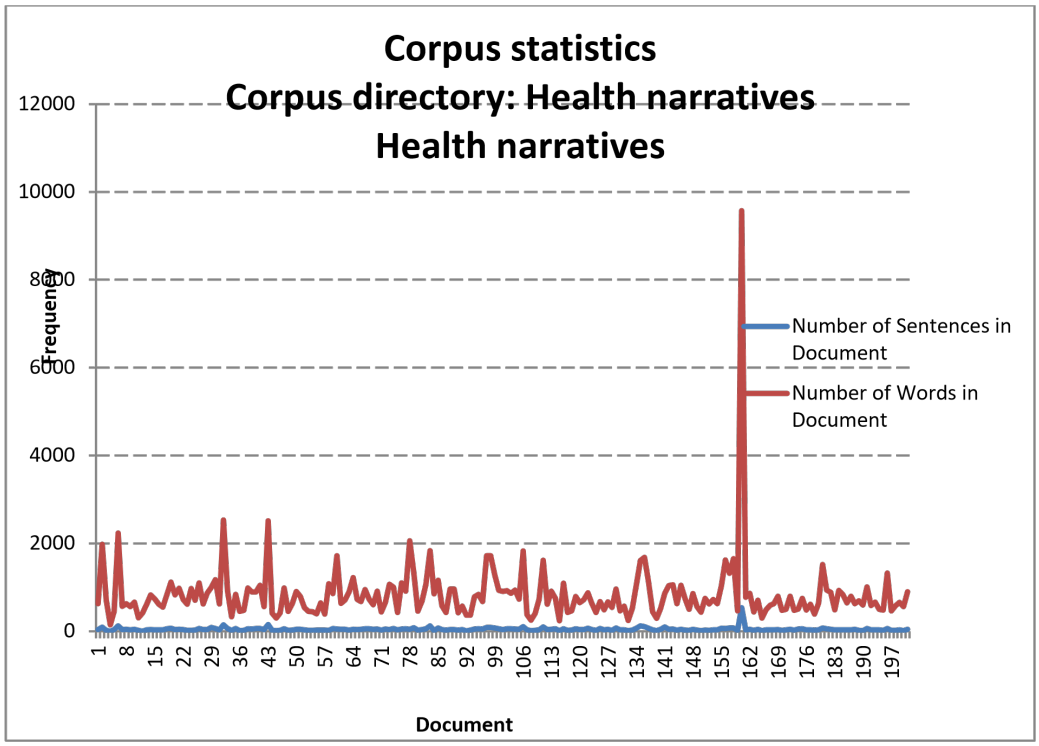
Once you have your sentiment score computed, if you wish to re-use the "Shape of Stories" tool, you can directly input the csv file created (*NLP_CoreNLP_sentiment_Dir_times_reviews.csv*) in your "Setup INPUT/OUTPUT configuration" to avoid repetitive computations by the NLP Suite.

To learn more about sentiment analysis, read the TIPS_NLP_Sentiment analysis.

Compute & Visualize Corpus Statistics

The second script available entitled "Compute & Visualize Corpus Statistics" gives you basic statistics about your corpus such as sentence and word count per text throughout your corpus. The script outputs 2 distinct excel files with word and sentence count gathered below in the excel graph.

Excel Output:



Hierarchical Clustering (HC)

Hierarchical Clustering (HC) is the first "Shape of Stories" Algorithm provided by the NLP Suite.

To learn more about Hierarchical Clustering, read the TIPS_NLP_HC.

Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) is the second "Shape of Stories" Algorithm provided by the NLP Suite.

To learn more about Singular Value Decomposition, read the TIPS_NLP_SVD.

Non-Negative Matrix Factorization (NMF)

Non-Negative Matrix Factorization (NMF) is the third "Shape of Stories" Algorithm provided by the NLP Suite.

To learn more about Non-Negative Matrix Factorization, read the TIPS_NLP_NMF.

Parameters GUIs

The NLP Suite allows you to tune parameters of the algorithms in order to get personalized results answering best to your needs and type of Analysis.

To learn more about Parameter Tuning, read the TIPS_NLP_SoS_Parameters.

Best Topic Estimation

The Best Topic Estimation script is used to find the optimal number of clusters (k) to run the Shape of Stories tool.

To learn more about the Best Topic Estimation function, read the TIPS_NLP_Best Topic Estimation.

Output

The pipeline provides a number of tabular (csv files) and visual tools to help the user to identify common story shapes. Please look at specific script sections for specific details.

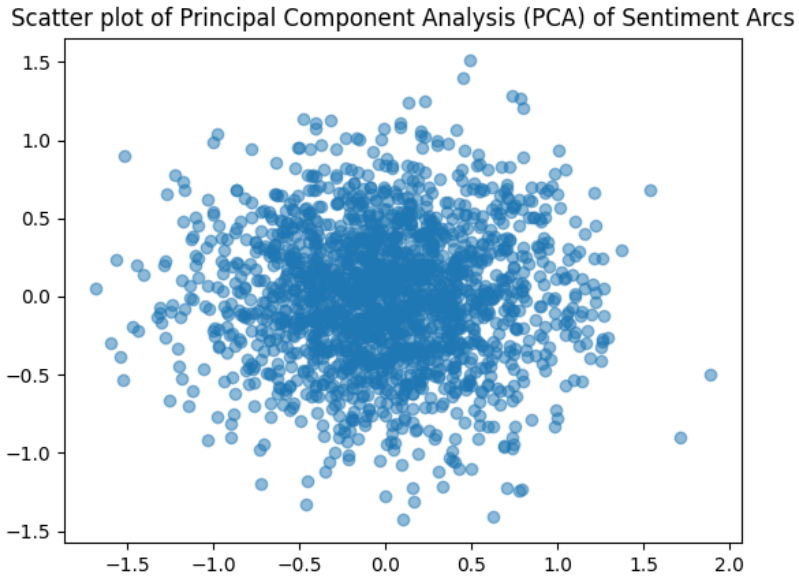
Principal Component Analysis (PCA)

Principal Component Analysis is produced by all scripts. The principal components of a collection of points in a real coordinate space are a sequence of p unit vectors, where the i-th vector is the direction of a line that best fits the data while being orthogonal to the first i-1

vectors. Here, a best-fitting line is defined as one that minimizes the average squared distance from the points to the line. It is commonly used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible (Wikipedia).

In the NLP Suite, PCA is used to reduce dimensions of n-dimensional vector and produce a 2-dimensional graph allowing us to highlight relationships between points belonging to the same cluster (Similar to Word2Vec in the approach) using the Sentiment Score.

Output graph:



References

Burrows, J.F. 1987. "Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style." *Literary and Linguistic Computing*, Vol. 2, No. 2, pp. 61-70.

Franzosi, Roberto. 2004. *From Words to Numbers*. Cambridge, UK: Cambridge University Press.

Franzosi, Roberto. 2010. *Quantitative Narrative Analysis*. Thousand Oaks, CA: Sage.

Reagan, Andrew J., Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. "The Emotional Arcs of Stories Are Dominated by Six Basic Shapes". *EPJ Data Science*, Vol. 5, No. 31, pp. 1-12.

Vonnegut, Kurt. https://www.youtube.com/watch?v=oP3c1h8v2ZQ

Vonnegut, Kurt. https://www.youtube.com/watch?v=GOGru_4z1Vc

Vonnegut, Kurt. 2005. "Here is a Lesson in Creative Writing." In: pp. 23-28, Kurt Vonnegut, *A Man Without a Country. Edited by Daniel Simon.* New York: Seven Stories Press.