

~~Project Guten-Topic, meine LDA~~
~~Project Tip of the Gutenberg~~
Project Guten-bag-of-words

Jack Etheredge

search for books

- Browse Catalog
- Bookshelves
- Main Page
- Categories
- Contact Info

Project Gutenberg appreciates your donation!

[Donate](#)

- Why donate?

in other languages

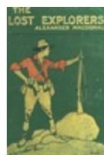
- Português
- Deutsch
- Français



Free ebooks - Project Gutenberg

[Book search](#) · [Book categories](#) · [Browse catalog](#) · [Mobile site](#) · [Report errors](#) · [Terms of use](#)

Some of the Latest Books



Welcome

Project Gutenberg offers over 57,000 free eBooks. Choose among free epub books, free kindle books, download them or read them online. You will find the world's great literature here, with focus on older works for which copyright has expired. Thousands of volunteers digitized and diligently proofread the eBooks, for enjoyment and education.

No fee or registration is required. If you find Project Gutenberg useful, please consider a small [donation](#), to help Project Gutenberg digitize more books, maintain our online presence, and improve Project Gutenberg programs and offerings. Other ways to help include [digitizing more books](#) 📁, [recording audio books](#) 📻, or [reporting errors](#).

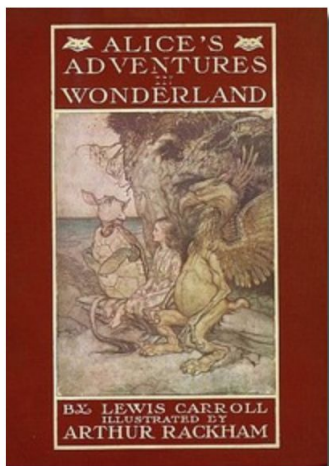


[Project Gutenberg](#)
[Mobile Site](#)



[Project Gutenberg](#) > [57,175 free ebooks](#) > [34 by Lewis Carroll](#)




















Alice's Adventures in Wonderland by Lewis Carroll



[Download](#)

[Bibrec](#)

Download This eBook

Format ?	Size	?	?	?
 Read this book online: HTML	217 kB			
 EPUB (with images)	1.3 MB			
 EPUB (no images)	144 kB			
 Kindle (with images)	2.9 MB			
 Kindle (no images)	485 kB			
 Plain Text UTF-8	173 kB			
 More Files...				

Motivation:

Project Gutenberg is a great resource for free books, but lacks easy access to:

- Plot summary

- Genre

- Characters

Aims:

ML-generated

Genre

Summary

Archetypal characters

Aims:

ML-generated

Genre

Summary

Archetypal characters

Topics (approximate Genres)

Clean

Lemmatize

TF-IDF

NMF

Topics (approximate Genres)



Fantasy stuff:

Topic 1: prince princess wa king said queen palace fairy went daughter little came day man till

Topic 14: dorothy ozma oz scarecrow wizard wa nome woodman tin said glinda magic trot billina toto

Space/sci-fi stuff:

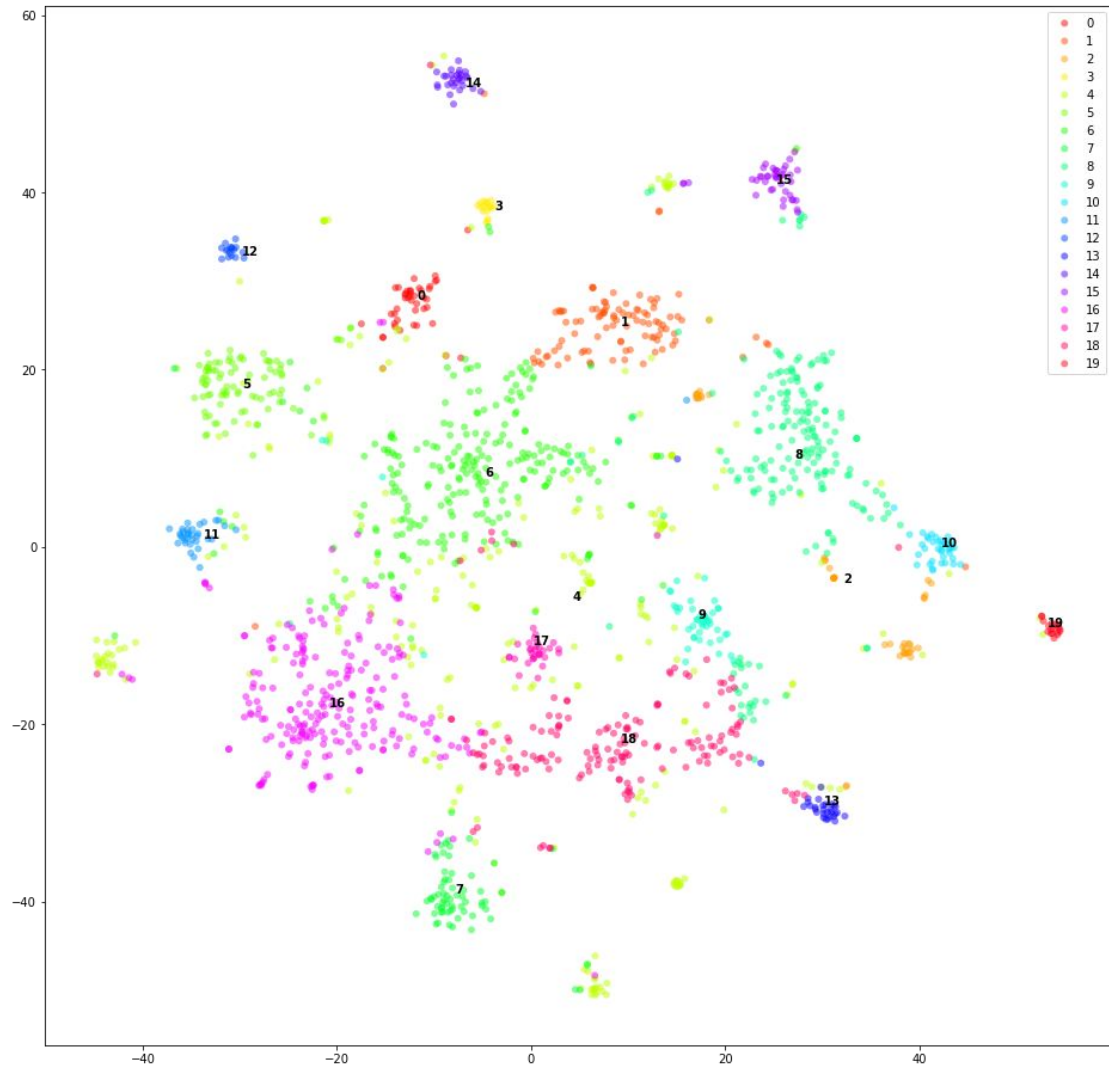
Topic 12: tom astro roger cadet connel _polaris_ corbett strong said walter manning solar deck spaceman wa

Topic 24: ship wa space rocket control fleet screen crew pilot air radar hull planet radio men

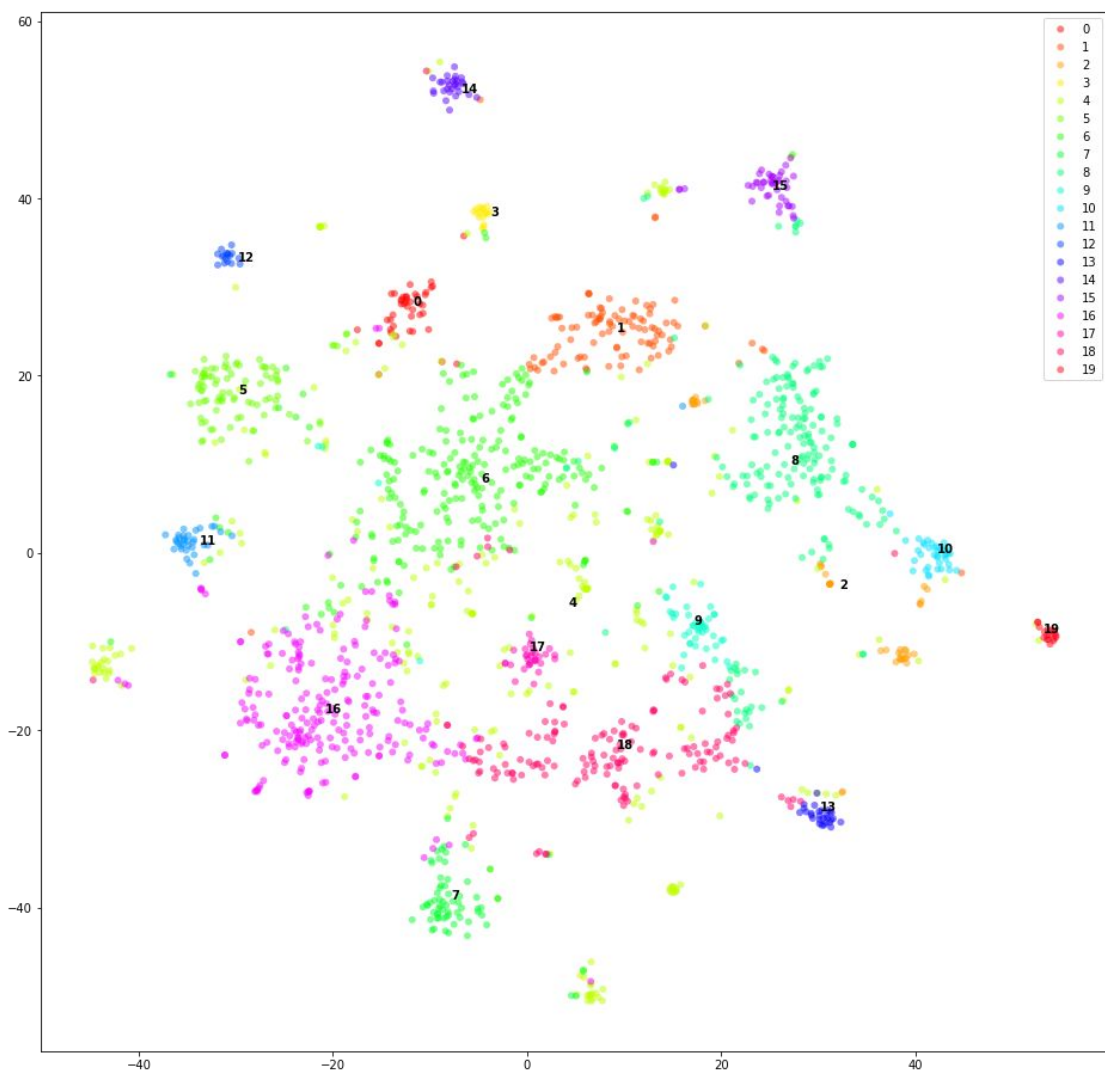
Genres:

t-SNE visualization

Color coded by KMeans
clustering of topics

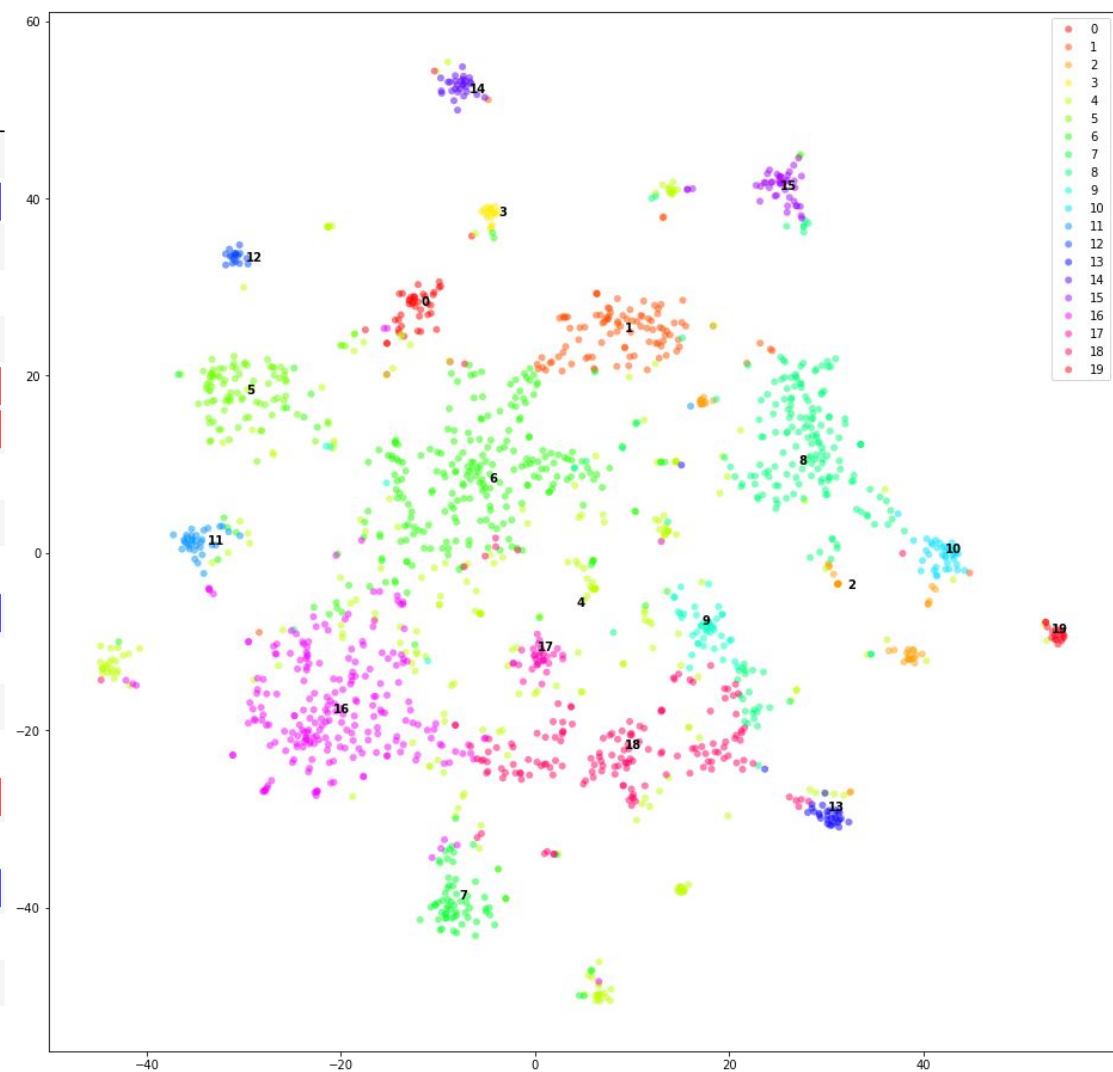


clusters		max_topic_cluster_words	max_topic_num
0	0	[mother, girl, dear, father, aunt, love, lady,...	3
1	1	[sir, knight, king, launcelot, arthur, ye, tri...	5
2	2	[joe, mike, haney, farrell, benson, brett, ken...	14
3	3	[dr, car, machine, doctor, colonel, maybe, pau...	9
4	4	[sea, rock, mile, tree, moon, round, shall, sh...	0
5	5	[martian, mar, earthman, planet, sand, dome, c...	17
6	6	[robot, jon, alan, robert, rankin, clark, ned,...	26
7	7	[god, king, city, yann, sea, song, sword, prie...	27
8	8	[professor, jack, andy, henderson, mark, washi...	25
9	9	[thou, thee, thy, shall, ye, spake, hast, wilt...	6
10	10	[dorothy, ozma, oz, scarecrow, wizard, nome, w...	13
11	11	[peter, reddy, unc, billy, jimmy, fox, rabbit,...	7
12	12	[mr, miss, sir, lady, mary, captain, paper, an...	16
13	13	[äù, äôs, äôt, äô, äúi, äöll, äôd, äí, äúit, ä...	2
14	14	[ship, planet, captain, alien, control, rocket...	4
15	15	[johnny, chuck, anne, baba, harry, mccord, der...	18
16	16	[king, prince, princess, queen, palace, daught...	1
17	17	[dr, car, machine, doctor, colonel, maybe, pau...	9
18	18	[george, gloria, fred, charlie, helen, uncle, ...	29
19	19	[state, war, social, general, german, class, b...	15



Fairy tales // Sci-fi

clusters		max_topic_cluster_words	max_topic_num
0	0	[mother, girl, dear, father, aunt, love, lady,...	3
1	1	[sir, knight, king, launcelot, arthur, ye, tri...	5
2	2	[joe, mike, haney, farrell, benson, brett, ken...	14
3	3	[dr, car, machine, doctor, colonel, maybe, pau...	9
4	4	[sea, rock, mile, tree, moon, round, shall, sh...	0
5	5	[martian, mar, earthman, planet, sand, dome, c...	17
6	6	[robot, jon, alan, robert, rankin, clark, ned,...	26
7	7	[god, king, city, yann, sea, song, sword, prie...	27
8	8	[professor, jack, andy, henderson, mark, washi...	25
9	9	[thou, thee, thy, shall, ye, spake, hast, wilt...	6
10	10	[dorothy, ozma, oz, scarecrow, wizard, nome, w...	13
11	11	[peter, reddy, unc, billy, jimmy, fox, rabbit,...	7
12	12	[mr, miss, sir, lady, mary, captain, paper, an...	16
13	13	[äù, äôs, äôt, äô, äúi, äöll, äôd, äí, äúit, ä...	2
14	14	[ship, planet, captain, alien, control, rocket...	4
15	15	[johnny, chuck, anne, baba, harry, mccord, der...	18
16	16	[king, prince, princess, queen, palace, daught...	1
17	17	[dr, car, machine, doctor, colonel, maybe, pau...	9
18	18	[george, gloria, fred, charlie, helen, uncle, ...	29
19	19	[state, war, social, general, german, class, b...	15



clusters		max_topic_cluster_words	max_topic_num
0	0	[mother, girl, dear, father, aunt, love, lady,...	3
1	1	[sir, knight, king, launcelot, arthur, ye, tri...	5
2	2	[joe, mike, haney, farrell, benson, brett, ken...	14
3	3	[dr, car, machine, doctor, colonel, maybe, pau...	9
4	4	[sea, rock, mile, tree, moon, round, shall, sh...	0
5	5	[martian, mar, earthman, planet, sand, dome, c...	17
6	6	[robot, jon, alan, robert, rankin, clark, ned,...	26
7	7	[god, king, city, yann, sea, song, sword, prie...	27
8	8	[professor, jack, andy, henderson, mark, washi...	25
9	9	[thou, thee, thy, shall, ye, spake, hast, wilt...	6
10	10	[dorothy, ozma, oz, scarecrow, wizard, nome, w...	13
11	11	[peter, reddy, unc, billy, jimmy, fox, rabbit,...	7
12	12	[mr, miss, sir, lady, mary, captain, paper, an...	16
13	13	[äù, äôs, äôt, äô, äúi, äöll, äöd, äí, äúit, ä...	2
14	14	[ship, planet, captain, alien, control, rocket...	4
15	15	[johnny, chuck, anne, baba, harry, mccord, der...	18
16	16	[king, prince, princess, queen, palace, daught...	1
17	17	[dr, car, machine, doctor, colonel, maybe, pau...	9
18	18	[george, gloria, fred, charlie, helen, uncle, ...	29
19	19	[state, war, social, general, german, class, b...	15

Genres:

- Defining a genre (sub-genre) by the cluster for the purposes of recommendations
- Using the most common topic in the cluster to ballpark what the cluster is about

Aims:

ML-generated

Genre

Summary

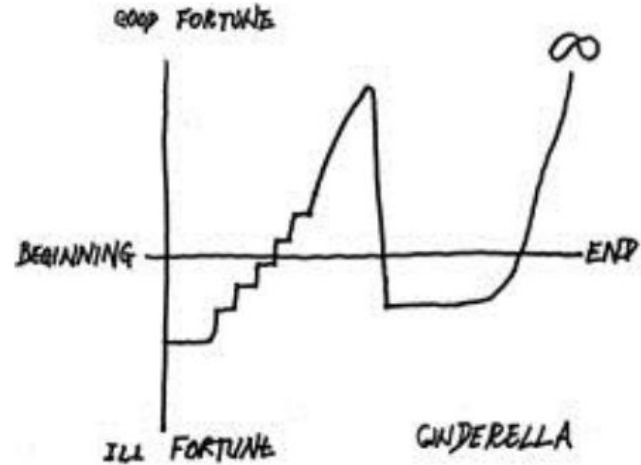
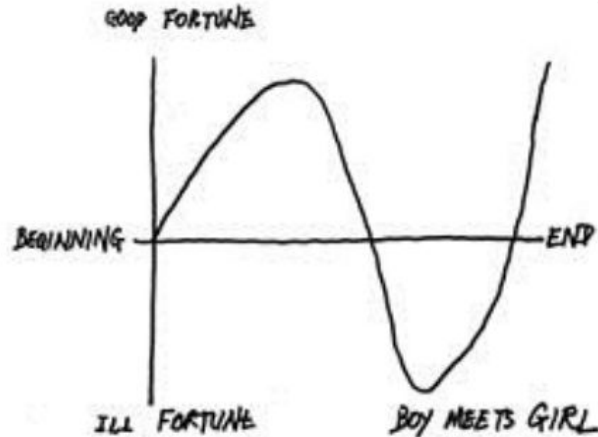
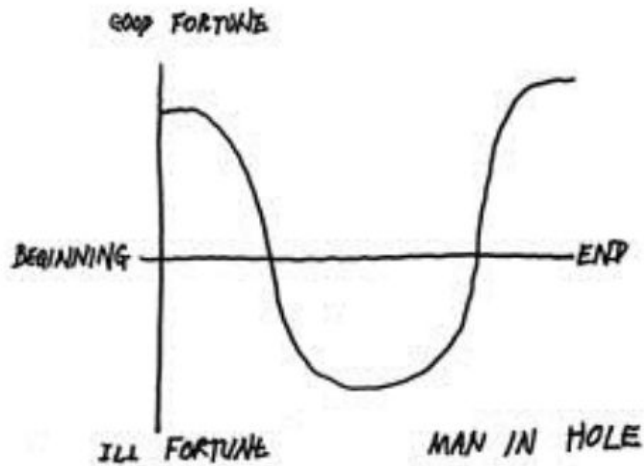
This is difficult if not impossible with extractive summarizers

Abstractive summarizers take FOREVER

What CAN we show a potential reader instead that's interesting?

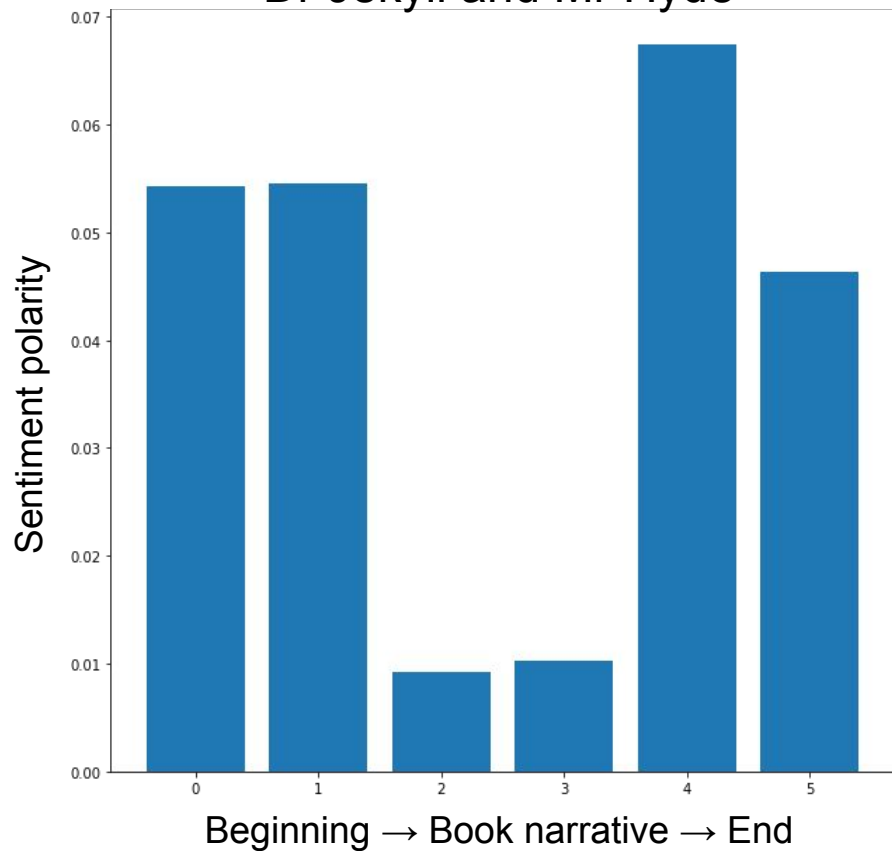
Archetypal characters

Vonnegut's story arc shapes:



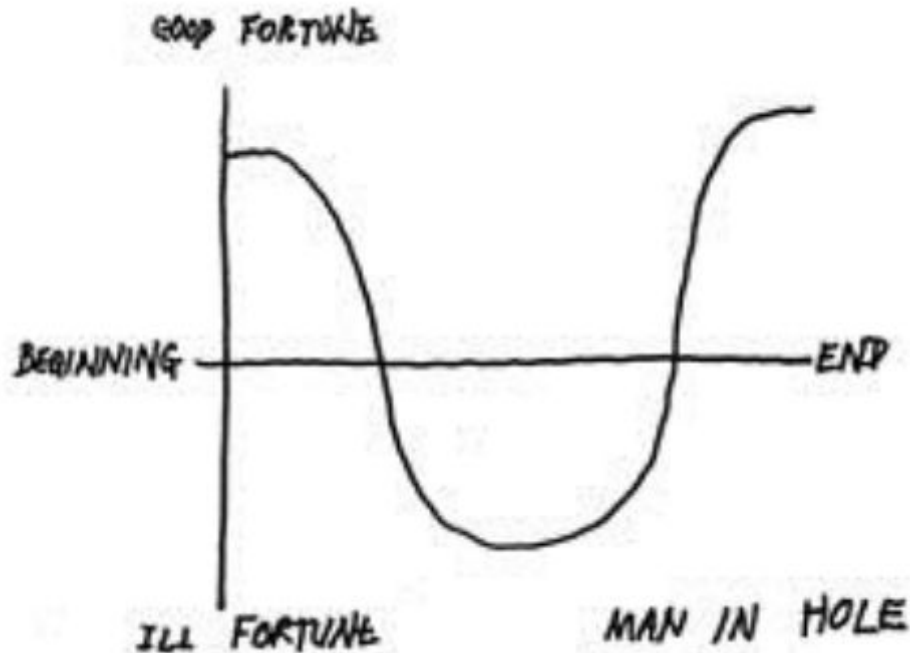
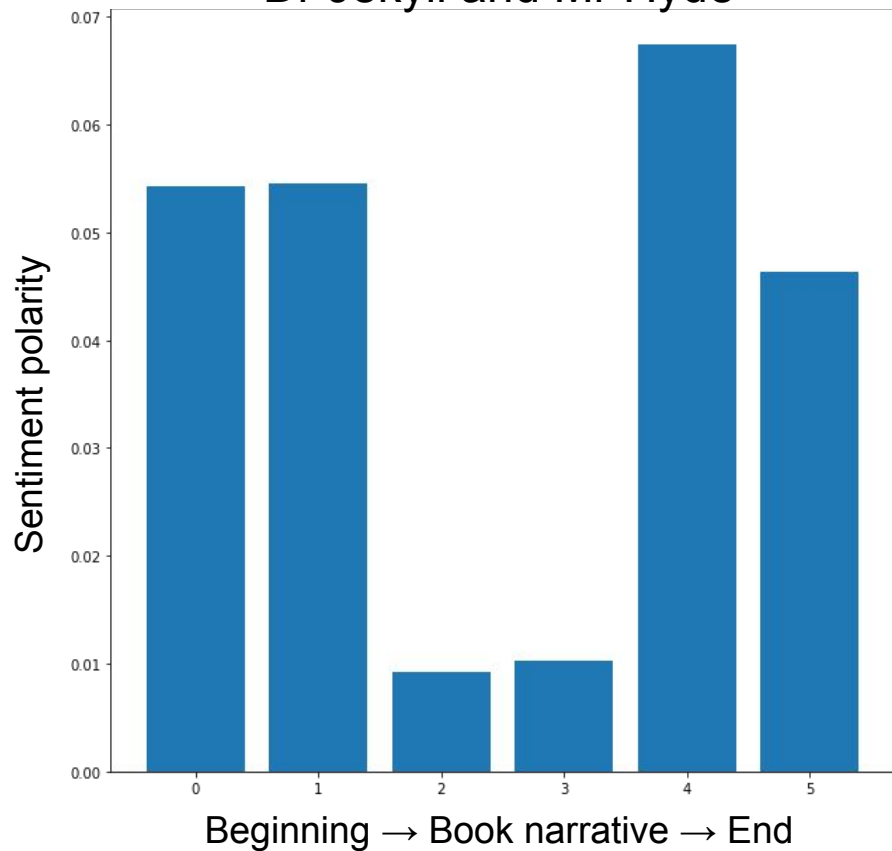
Sentiment trajectory (binned sentiment polarity):

Dr Jekyll and Mr Hyde

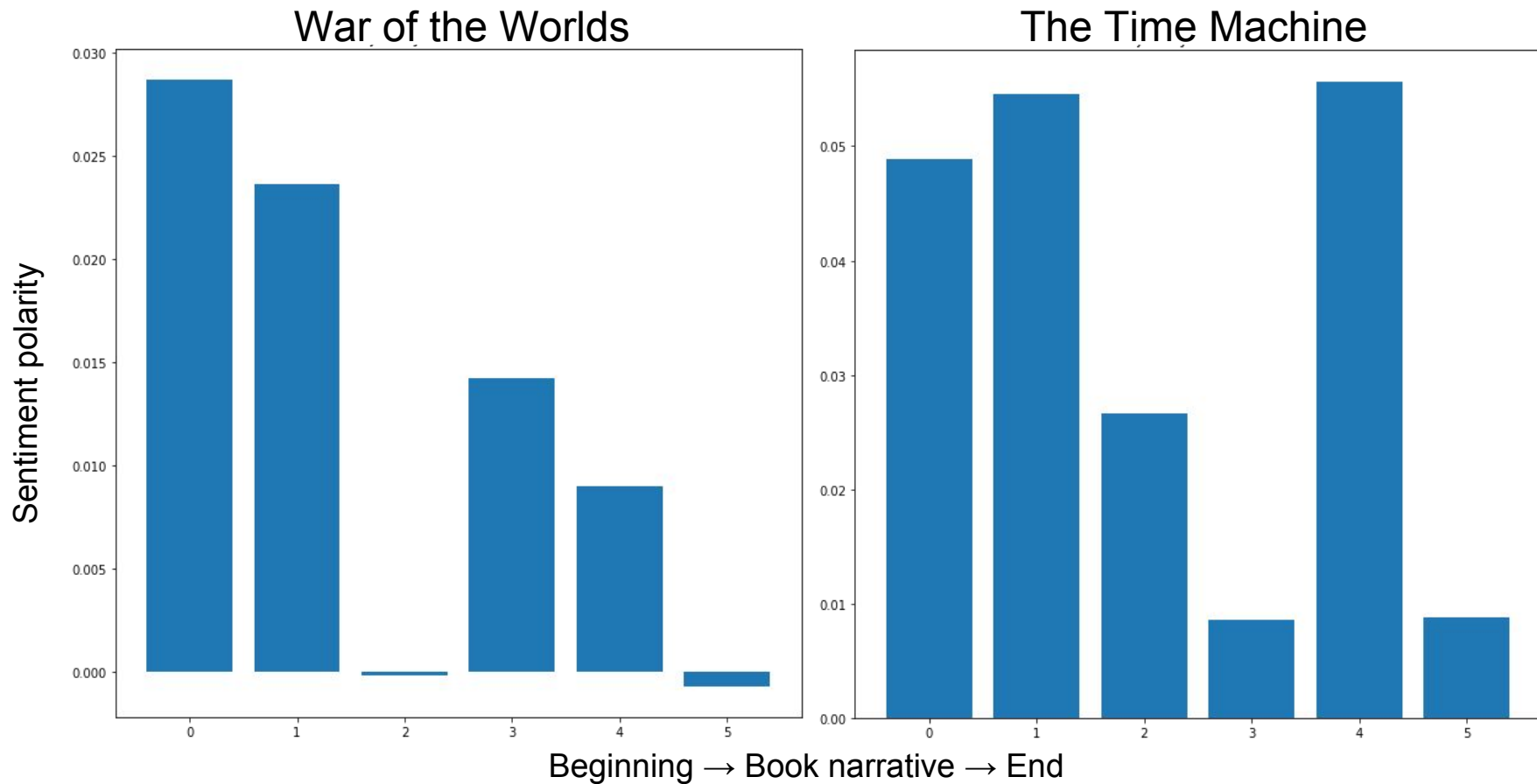


Sentiment trajectory (binned sentiment polarity):

Dr Jekyll and Mr Hyde



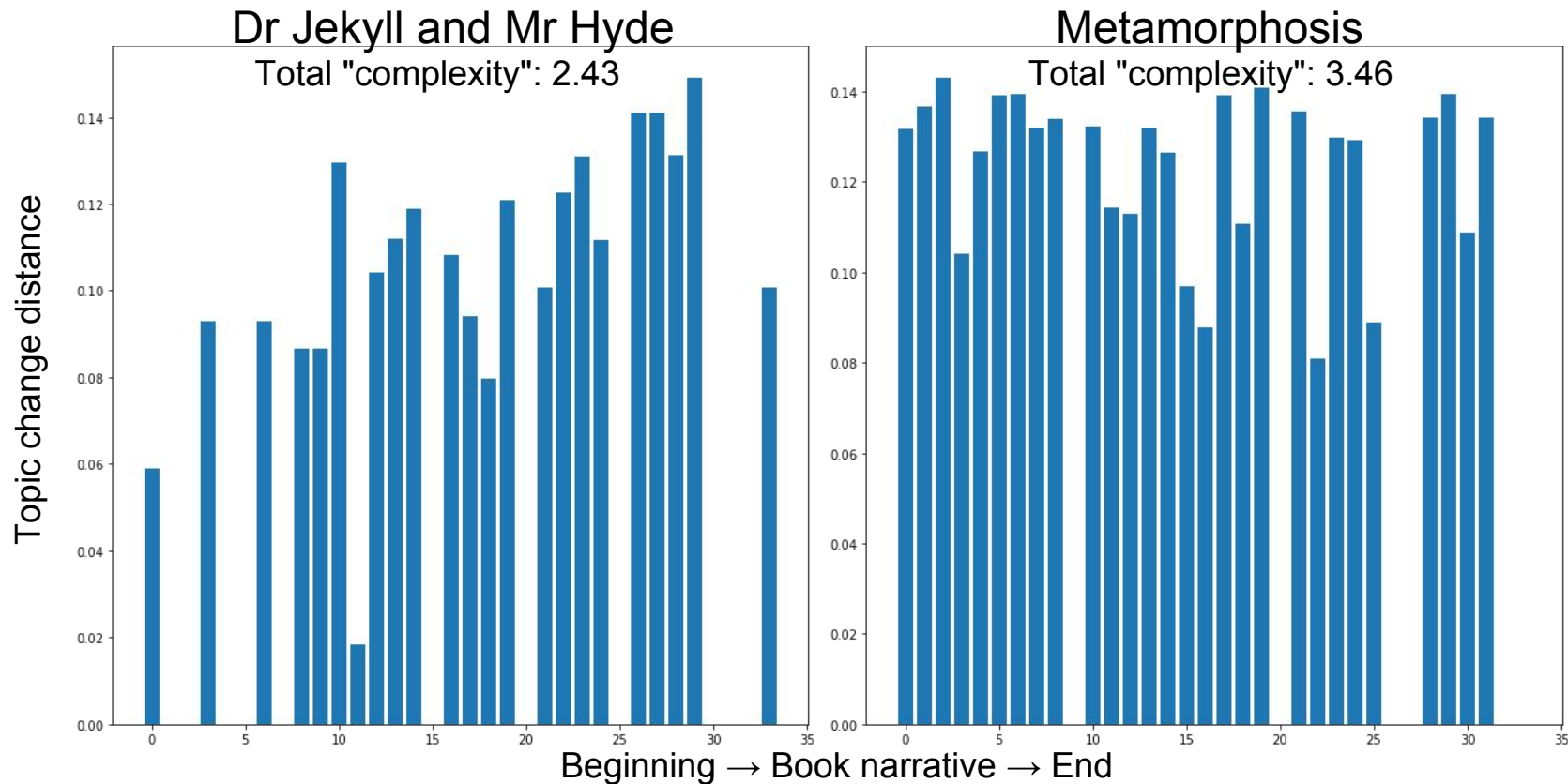
Most are more complicated shapes:



Narrative complexity

- Binned sentences and took the max topic per sentence
- Took the mode of each bin
- Calculated cosine distance between topics per bin (the deltas)

Narrative complexity



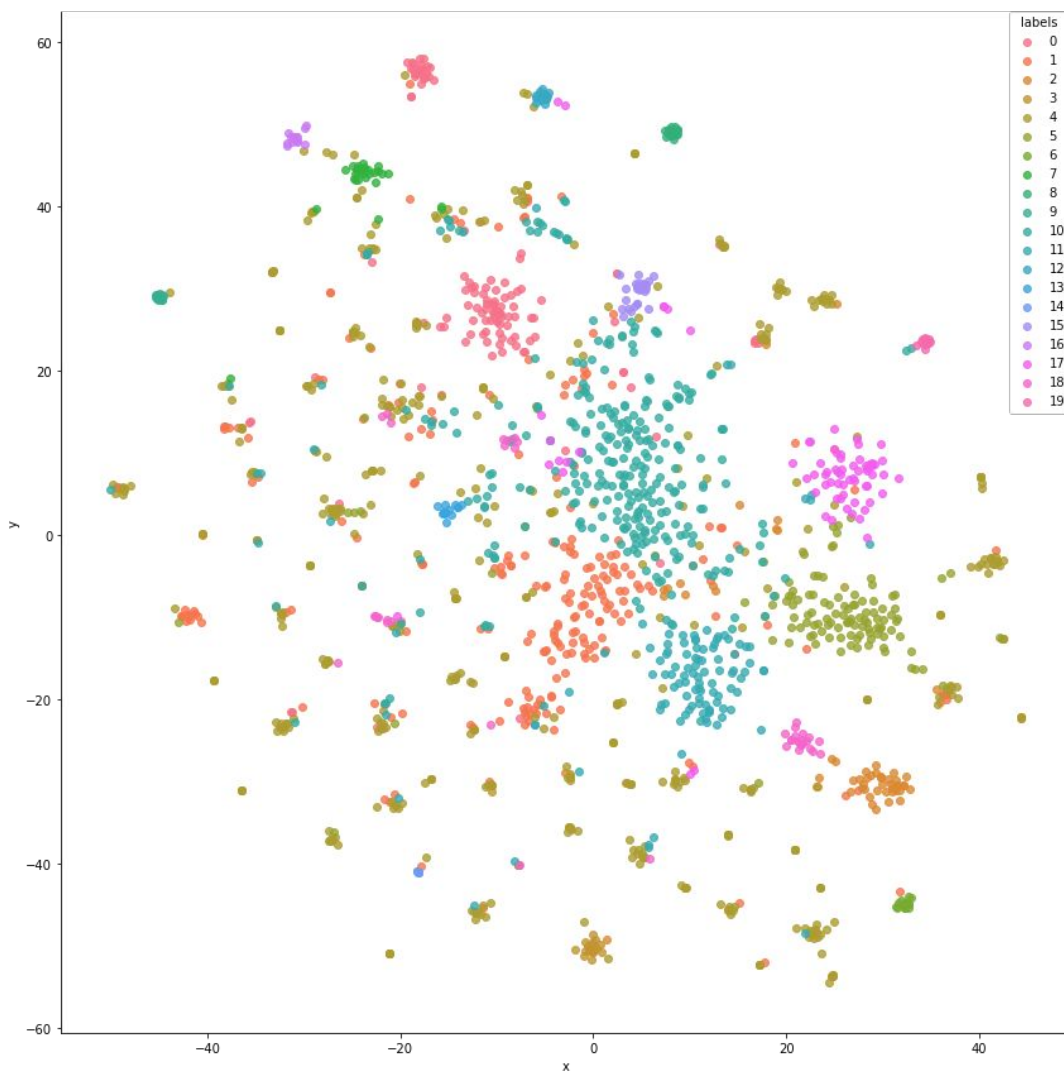
Show app

Future directions:

- Deploy to elastic beanstalk
- Compare authors
- ML-generated archetypal characters
- Add the remaining books

Sources

<http://msmcclure.com/pdf/Five%20Story%20Arcs.pdf>

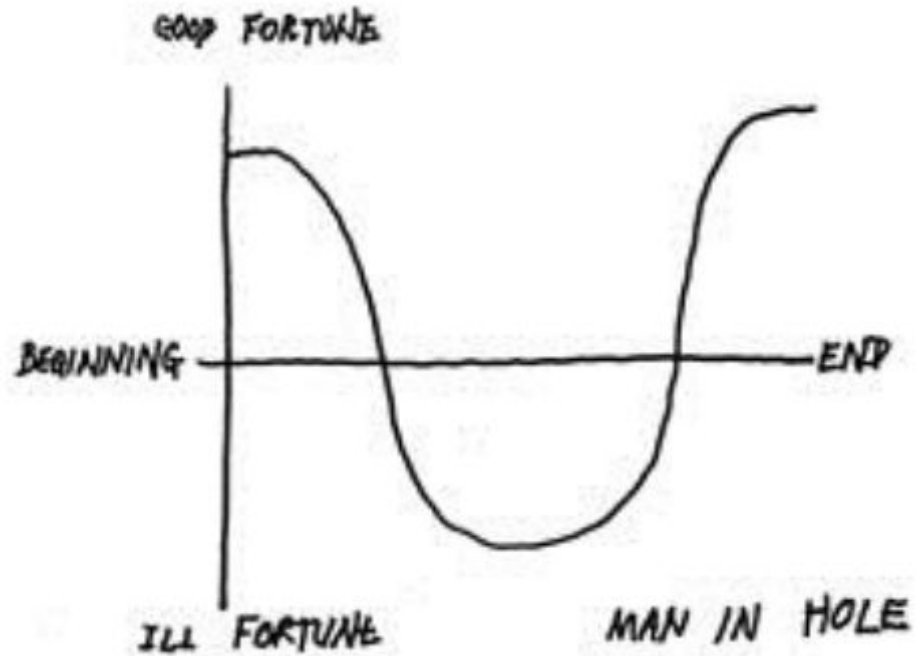


Genres:

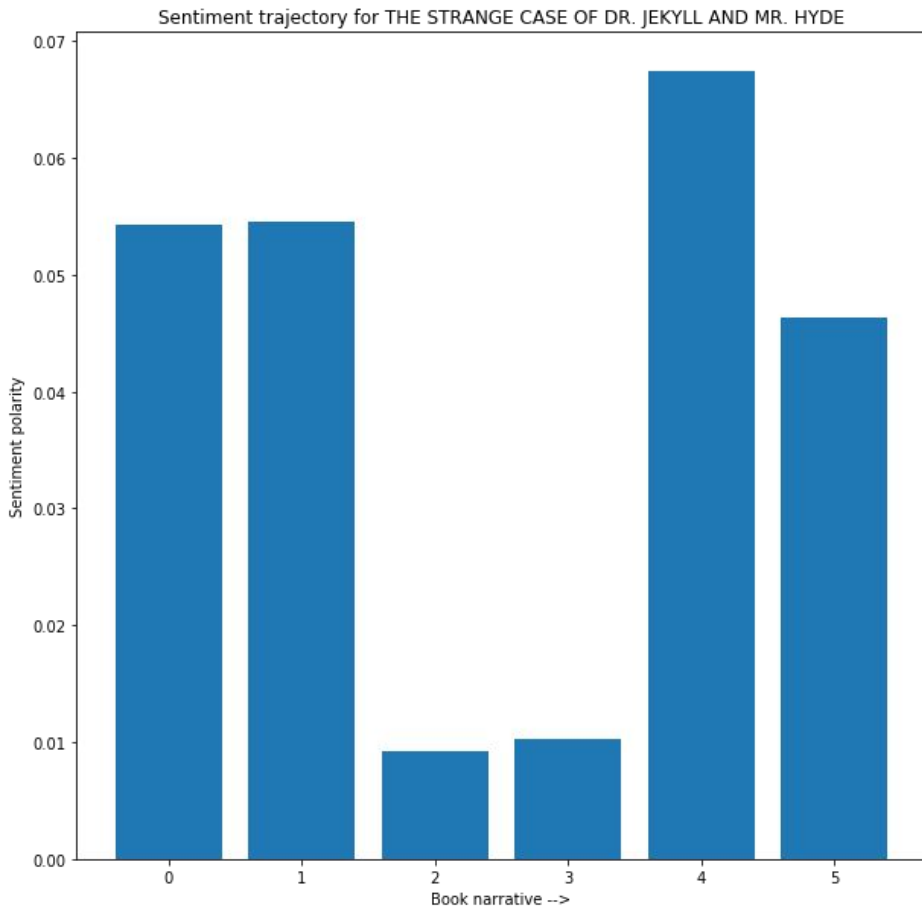
t-SNE visualization

Color coded by KMeans
clustering of topics

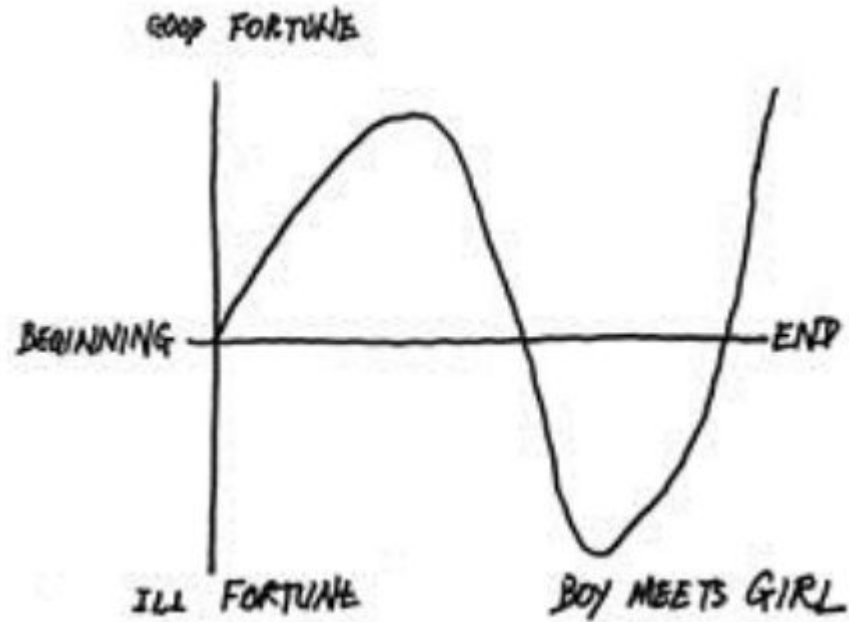
Man in a hole:



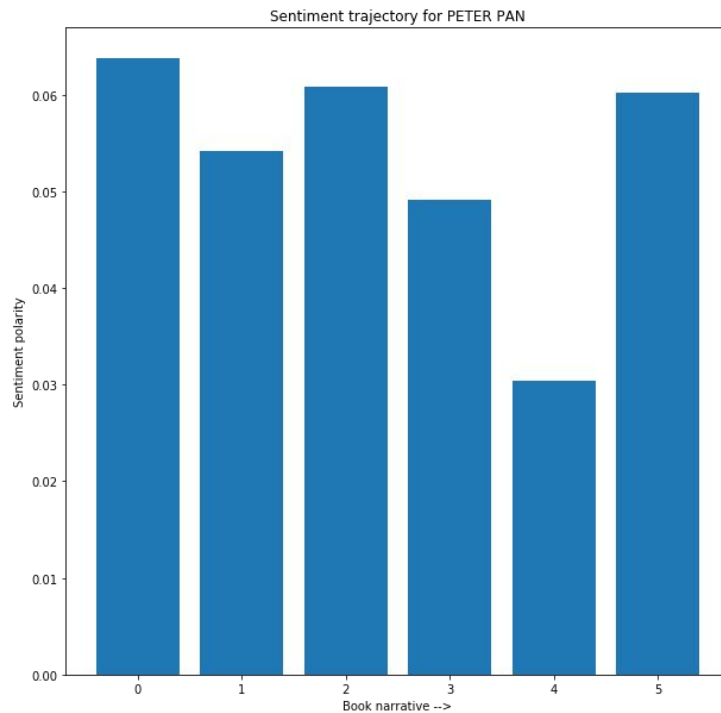
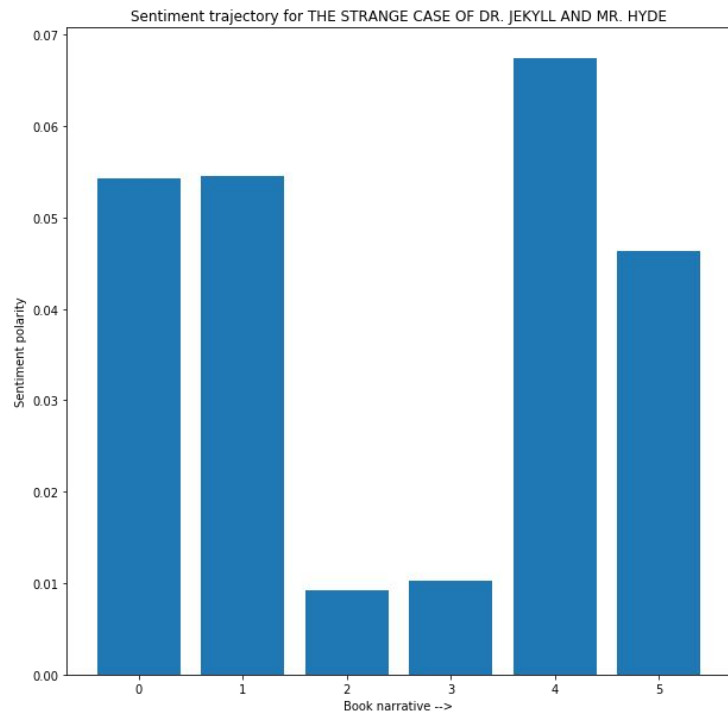
Sentiment trajectory (binned sentiment polarity):



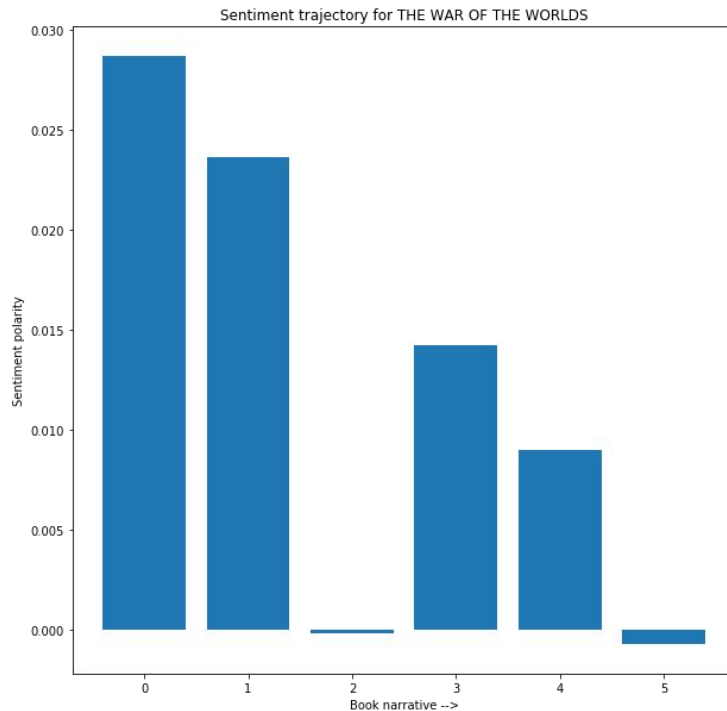
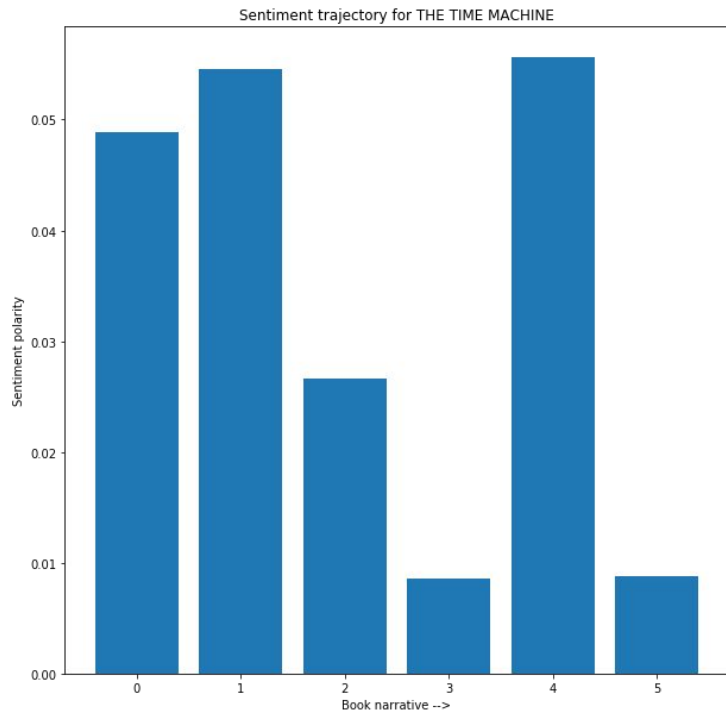
Boy meets girl:



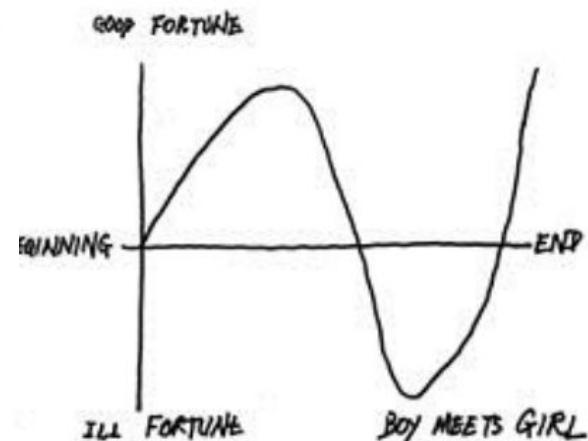
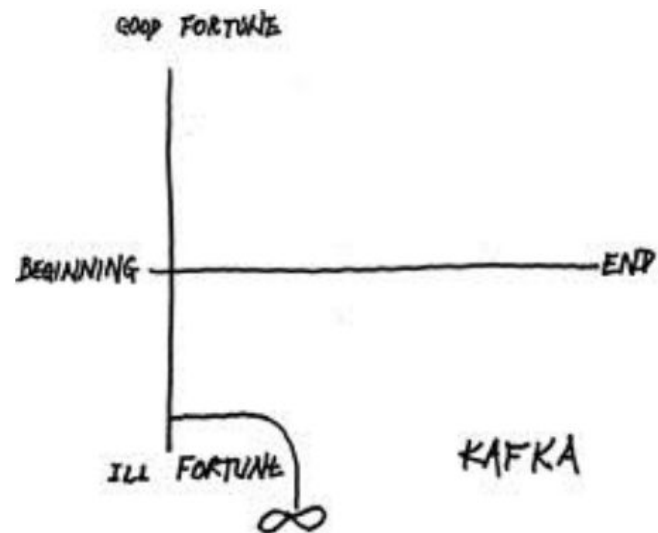
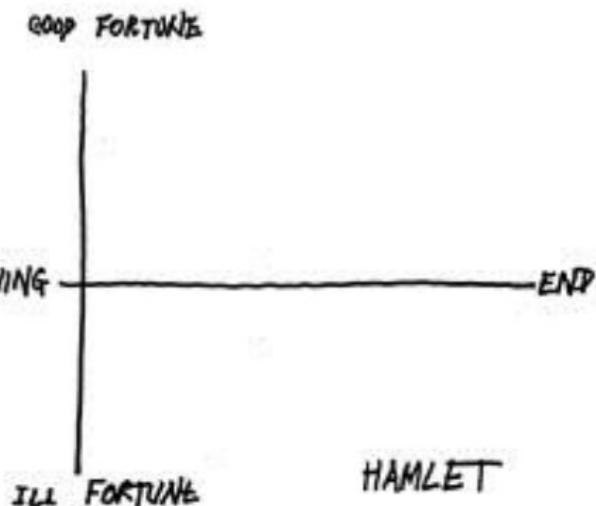
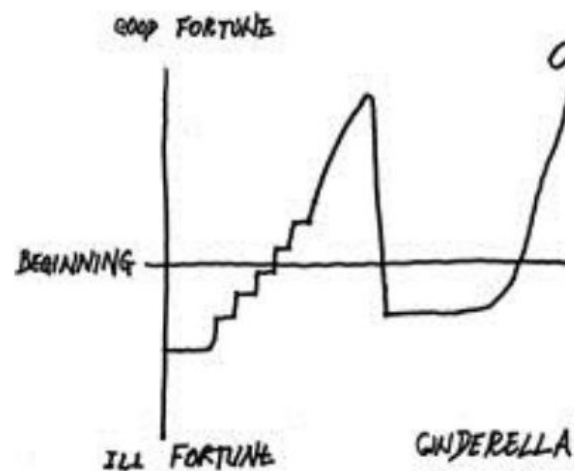
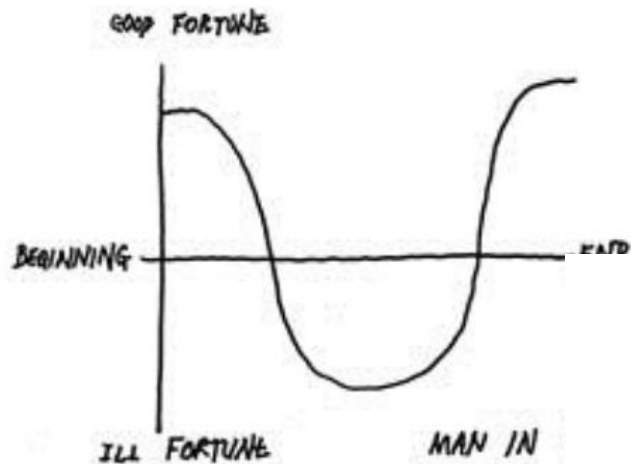
Sentiment trajectory



Most are more complicated shapes:



arc shapes:



arc shapes:

