MacKenzye Leroy
zuf9mc@virginia.edu
May 2, 2022

Jules Verne and H.G. Wells: A Textual Analytical Comparison

Jules Verne and H. G. Wells are often referred to as the "Godfathers of Science Fiction" for their massive impact on the genre in its early years. Although both were born in the 19th Century and most of their works are well over one hundred years old, they "remain, arguably, the two most famous writers of science fiction. Their names are conventionally linked ... although they never met, and in fact come from different generations (Verne was 38 when Wells was born) (Roberts, 2016)" As such, they make for an interesting comparison case. Some of the foundational methods of exploratory text analytics were employed to compare the top 10 works (as measured by number of downloads on Project Gutenberg) of each author (Figure 1). Methods employed include hierarchical clustering and principal component analysis (PCA) of the entire corpus, topic modelling across the corpus as well as on each individual author, and sentiment analysis on the entire corpus and key documents within the corpus.

Title	Author	PG Number	PG Downloads
The Time Machine	H.G. Wells	35	7598
The War of the Worlds	H.G. Wells	36	6743
The Island of Doctor Moreau	H.G. Wells	159	3027
The Invisible Man	H.G. Wells	5230	2285
The World Set Free	H.G. Wells	1059	557
The Sleeper Awakens	H.G. Wells	12163	339
A Modern Utopia	H.G. Wells	6424	296
The First Men in the Moon	H.G. Wells	1013	244
The War in the air	H.G. Wells	780	240
Tono Bungay	H.G. Wells	718	199
Around the World in Eighty Days	Jules Verne	103	4685
Twenty Thousand Leagues under the Sea	Jules Verne	164	2691
A Journey to the Centre of the Earth	Jules Verne	18857	1996
The Mysterious Island	Jules Verne	1268	1298
Off on a Comet!	Jules Verne	1353	207
The Underground City	Jules Verne	1355	81
Eight Hundred Leagues on the Amazon	Jules Verne	3091	80
Five Weeks in a Balloon	Jules Verne	3526	433
All Around the Moon	Jules Verne	16457	277
Earth to the Moon	Jules Verne	44278	700

Figure 1: Documents in Corpus. Documents were chosen by number of downloads on Project Gutenberg (PG)

Clustering

One of the first methods employed to explore the similarity between Verne and Wells was simple hierarchical or agglomerative clustering. Using cosine distance and ward linkage, we see a few interesting patterns emerge (Figure 2). The first one is the clear split between Verne's works and Wells' works. Apart from one novel, *A Journey to the Center of the Earth* by Jules Verne, we see a clean divide between the two authors. Also of note is that we see closer clustering amongst Wells' works than Verne's works. Looking just at Wells' works we see one novel really separating from the others-*Tono-Bungay*. This work is actually closer to a work done by Verne than it was to any of the other Wells' novels. This is not particularly surprising though as *Tono-Bunga*y is a semi-autobiographical novel making it quite different from most of Wells' other works.

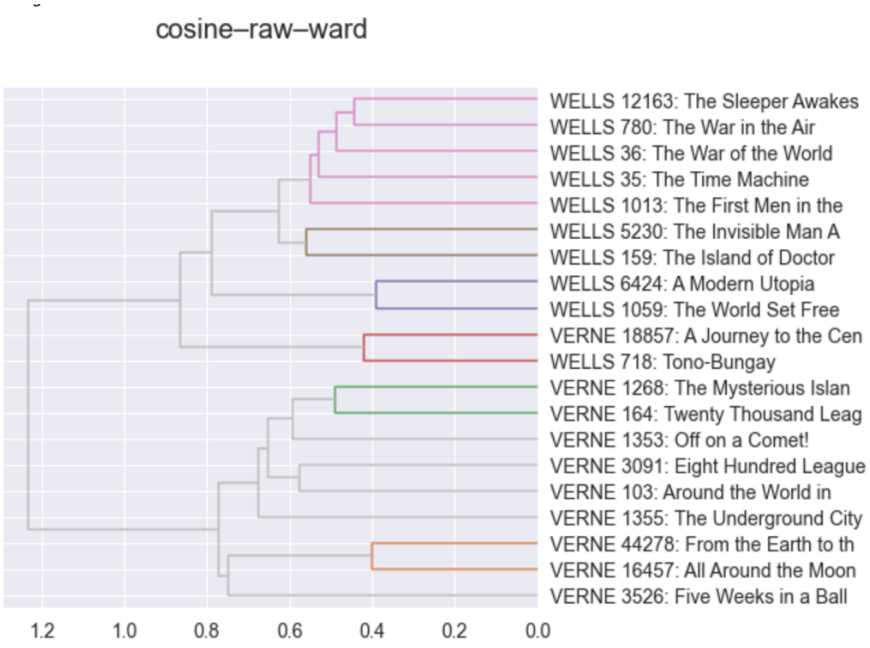

Figure 2: Dendrogram produced using Cosine Distance and Ward Linkage

Looking specifically at Verne's works we see a few different interesting patterns emerge. First, works with sequels are closely clustered to those sequels. *From the Earth to the Moon* is closely linked to its sequel *All around the Moon* and *Twenty Thousand Leagues Under the Sea* is closely linked to its sequel *The Mysterious Island*. Finally, just like Wells, with Verne we see one outlier novel-*Five Weeks in a Balloon*. *Five Weeks in a Balloon* was Jules Verne's first published novel and was really the product of over a decade of writing and tinkering, which may be why it's so different than the other books in this corpus. I believe this to be the case, but I'd be remiss if I failed to acknowledge that another potential reason this book may stand out so much may have nothing to do with the original writing, but rather an artifact of it being a translation from French. Not all translations are created equally and a poor one may lead to unexplained separation.

Principal Component Analysis (PCA)

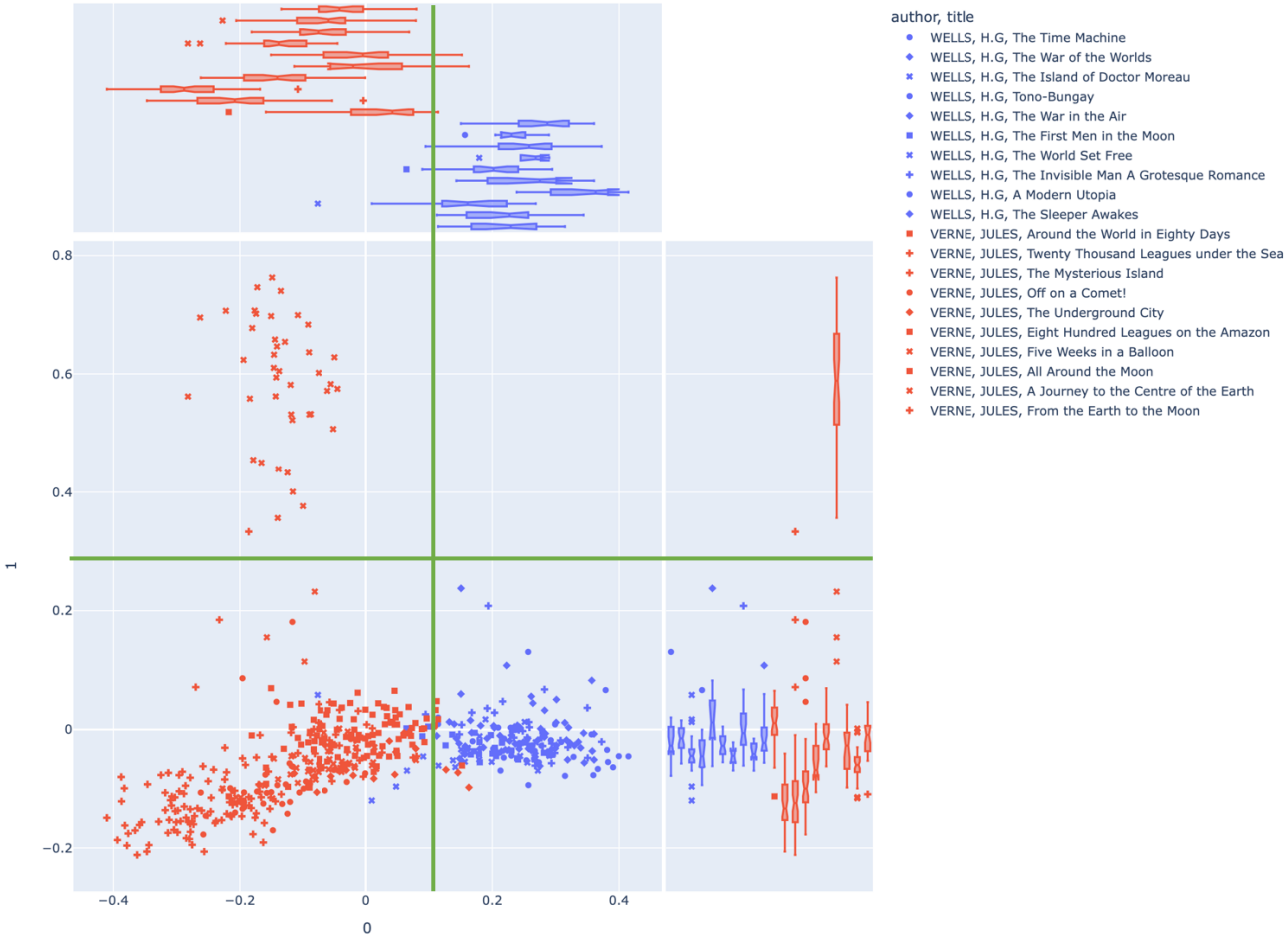

Figure 3: First and Second Principal Components. X-Axis = 1st Principal Component, Y-Axis = 2nd Principal Component

Moving on from simple hierarchical clustering, Principal Component Analysis (PCA) was used to further examine similarities and differences between the two authors. This of course requires finding the TFIDF (max TF, chapter as bag) for the entire corpus before employing PCA. Much like we see with hierarchical clustering, the first principal component cleanly splits the data between Verne and Wells (Figure 3, X-Axis). It's clear that in our corpus, whether a book was written by Verne or Wells was the primary way to separate out documents. More interestingly though, is that the second principal component essentially just splits out the data on whether the document is *Five Weeks in a Balloon* or not (Figure 3, Y-Axis). We saw with hierarchical clustering that this book was a bit of an outlier, but here it's even further separated from the other 19 works essentially comprising the entire variance of the second principal component. As noted above, this book was Verne's first published and a product of many years of tinkering giving it good reason to be a bit different than his later works which were often written in the span of a few months (Roberts, 2016).

Moving beyond the first couple principal components, we no longer see clear separation between the two authors like we saw in the first principal component, or any individual documents showing extreme separation like we saw in the second. Rather, we see different groups of documents beginning to emerge. For example, across principal component 3, we see mostly Verne's works making up the far ends of the component, with most of Wells' works hovering around the center. This, paired with Wells' works closer clustering in both hierarchical clustering and across the first two principal components shows a potentially interesting pattern beginning to emerge-Wells' works seem to be more closely related to each other than Verne's. If we examine the far ends of this principal component, we see the following works by Verne- *All Around The Moon, From Earth to the Moon, Off on a Comet*!, and *A Journey to the Center of the Earth* on one end. Except for the final novel in this group, these are all of course about space travel. Further, the one Wells book that ended up in this group was none other than *The First Men in the Moon* – another work about space travel. On the other end of this component, we find *Around the World in Eighty Days, The Mysterious Island, The Underground City, Eight Hundred Leagues on the Amazon*, and The *Invisible Man* by Wells. While *The Invisible Man* is a bit of an outlier in this group, the other four works, all written by Verne, are definitively earthbound adventure novels showing that this principal component may be differentiating between earthbound and space travel adventures.

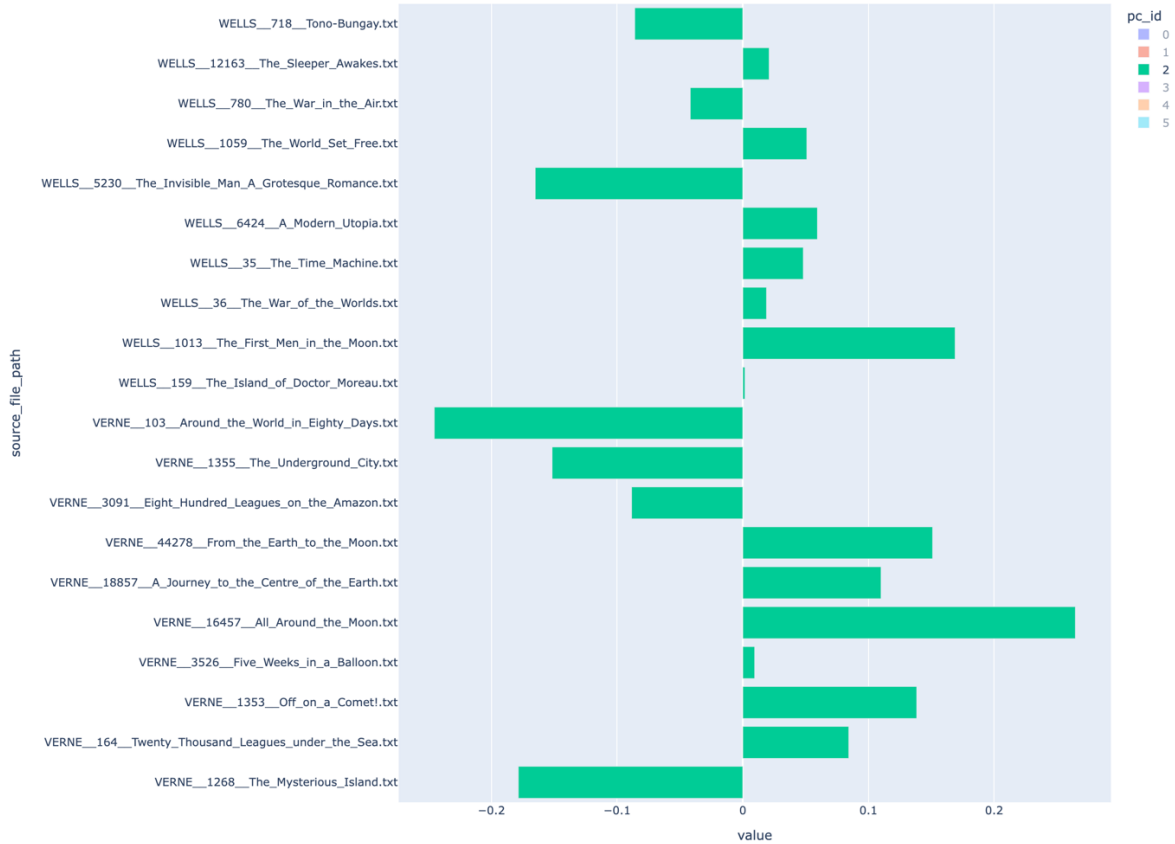

Figure 4: Third Principal Component

Topic Modelling

After using hierarchical clustering and PCA to explore the vocabulary choices of the two authors, topic modelling was employed to better understand the actual content or substance of the writing of each author. First, a topic model for the entire corpus was created, then individual topic models for the collective works of each individual author were created. In the topic model created for the entire corpus, Wells in general shows much less variance in topics than Verne between his novels. In fact, 6 of the 10 novels in the corpus written by Wells had the same topic as the dominant topic and another 2 of the 10 also shared a topic as the dominant topic (Appendix I). Verne's novels on the other hand almost all had a different topic as the most dominant one in the work. Further, in the joint model, one topic (topic_id 5) accounted for nearly 40% of Wells' content, while Verne's highest contributing topic was associated with under 11% of his total content (Figure 5).

topic_id	auth_WELLS	top_terms
5	0.393292	people face stood turned suddenly heard white towards
11	0.153643	things world thing life sort mind something people
16	0.142242	air sky night black towards water became wind
8	0.091934	got mr get london dont old im house
14	0.088490	world state people life new human things women
15	0.023884	moon earth captain cried surface miles point travellers
2	0.018072	old years new mine want yes cried let
9	0.012711	cried professor words let earth water appeared heard
19	0.012192	mr sir replied minutes twenty hours master oclock
13	0.011991	river trees hundred bank large whose forest several
4	0.011298	captain sea land sir water vessel ocean board
10	0.009429	captain professor earth sun days new temperature degrees
17	0.006492	replied judge answered yes life exclaimed words years
3	0.006404	island engineer replied granite house companions top sea
0	0.005855	sea mountain rocks vast soil earth enormous rock
1	0.004526	doctor balloon hundred wind let country air ground
6	0.002690	gun replied shot iron american point weight metal
18	0.002659	water pressure surface iron glass lower level depth
7	0.002054	lake town country mountains miles mountain reached city
12	0.000142	room soil worthy house journey four kind whose

topic_id	auth_VERNE	top_terms
3	0.106846	island engineer replied granite house companions top sea
10	0.104411	captain professor earth sun days new temperature degrees
9	0.095975	cried professor words let earth water appeared heard
4	0.091672	captain sea land sir water vessel ocean board
13	0.087216	river trees hundred bank large whose forest several
17	0.082207	replied judge answered yes life exclaimed words years
1	0.063660	doctor balloon hundred wind let country air ground
15	0.059955	moon earth captain cried surface miles point travellers
19	0.048193	mr sir replied minutes twenty hours master oclock
8	0.047431	got mr get london dont old im house
0	0.039265	sea mountain rocks vast soil earth enormous rock
6	0.035217	gun replied shot iron american point weight metal
2	0.025759	old years new mine want yes cried let
16	0.025298	air sky night black towards water became wind
14	0.021452	world state people life new human things women
7	0.019708	lake town country mountains miles mountain reached city
5	0.018839	people face stood turned suddenly heard white towards
11	0.014930	things world thing life sort mind something people
18	0.009255	water pressure surface iron glass lower level depth
12	0.002712	room soil worthy house journey four kind whose

Figure 5:

After applying topic modelling to the entire corpus, topic modelling was also used on each individual author separately. This was especially useful for Wells, whose works largely only fell into one topic with the joint model. Figure 6 shows a heatmap of topic versus novel for each individual author. Here we can actually see some variance in topics between Wells' works. In fact, Wells' works seem to be drawing more from secondary topics than Verne's in some cases. Still, it's important to note that with the joint model, we once again see far less variation between the works of Wells than those of Verne-a pattern that seems to be solidifying.

Figure 6: Left – Heatmap of Topic vs Novel for Wells' Corpus (Wells-Specific Topic Model), Right – Heatmap of Topic vs Novel for Verne's Corpus (Verne-Specific Topic Model)

Beyond the macrotrends in topics within the corpus, topic modelling can also be useful for comparing individual works within the corpus. Of particular interest was how topics did or didn't carry over into sequels. As covered in the clustering section, Verne had two sequels in the Corpus-*All Around the Moon,* which followed *From Earth to the Moon* and *The Mysterious Island,* which followed *Twenty Thousand Leagues Under the Sea.* Using the Verne-specific topic model, we do in fact see some carryover in topics between the two works. The second most frequent topic in *From Earth to the Moon* became the most frequent topic in *All Around the Moon,* though admittedly that was the only carryover topic in the top 4 topics of each novel (Figure 6). The connection between *Twenty Thousand League Under the Sea* and *The Mysterious Island* however was not as strong. We do see some carryover from *Twenty Thousand Leagues Under the Sea,* with the most two frequent topics from this work appearing in the top eight for *The Mysterious Island,* but by and large the two works seem to be drawing mostly from different topics (Figure 7). This may be because *All Around the Moon* is a true sequel to *From Earth to the Moon* while *The Mysterious Island* is a crossover sequel to both *Twenty thousand Leagues Under the Sea* and *In Search of the Castaways,* a novel not featured in this corpus. An interesting point of future analysis would be to include that novel while fitting the topic model and see if any of the prominent topics from that work do in fact carry over into *The Mysterious Island*.

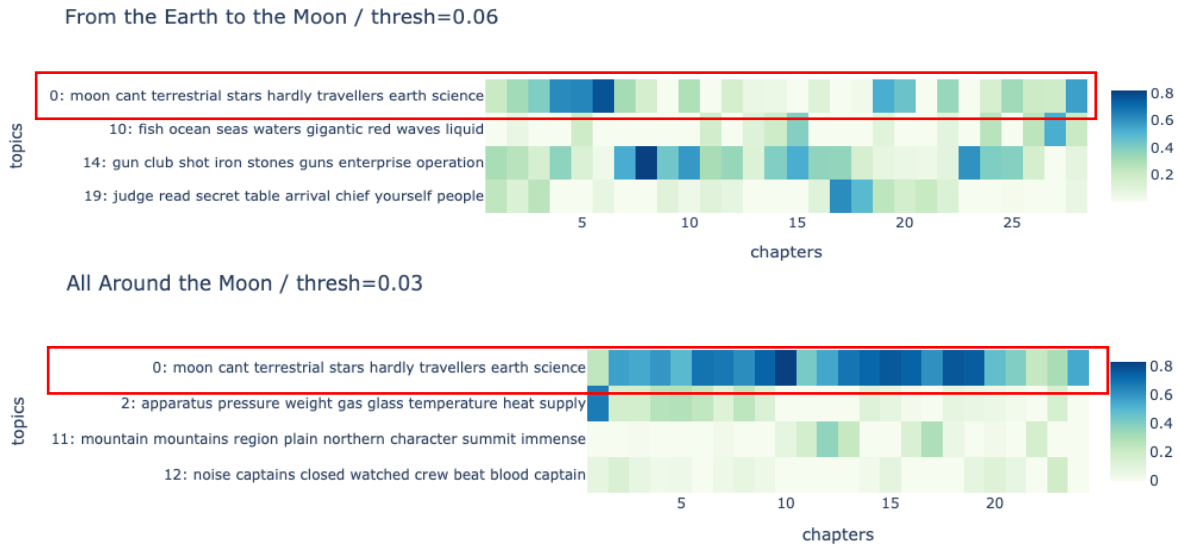

Figure 6: Top Topics by Chapter (Verne-Specific Topic Model) for *From Earth to the Moon & All Around the Moon,* Carryover topics highlighted

Figure 7: Top Topics by Chapter (Verne-Specific Topic Model) for *Twenty Thousand League Under the Sea* & *The Mysterious Island,* Carryover topics highlighted

Sentiment Analysis

Figure 8: Overall Polarity of Novels in Corpus

Finally, the last method used to explore the nuances within this corpus was sentiment analysis. Like with the other methods employed, sentiment analysis was first applied to the entire corpus to identify trends between authors. When it comes to overall polarity, Wells tended to have more negative books than Verne with 8 of his 10 novels having negative polarity (Figure 10). Wells also showed a wider variance in polarity than Verne with the most positive work in the whole corpus (*A Modern Utopia*) but also the most negative work in the whole corpus (*The Island of Doctor Moreau*). This is especially of note since Verne has largely shown more variance between works across all other forms analysis so far. Much of the difference in polarity between the two authors can be attributed to a divide in the use of disgust and trust. Wells' works scored very highly in disgust with 7 of the top 8 novels in terms of disgust (Figure 10), while Verne's works tended towards trust (the opposite emotion) with Verne having 7 of the top 9 novels in terms of trust (Figure 9).

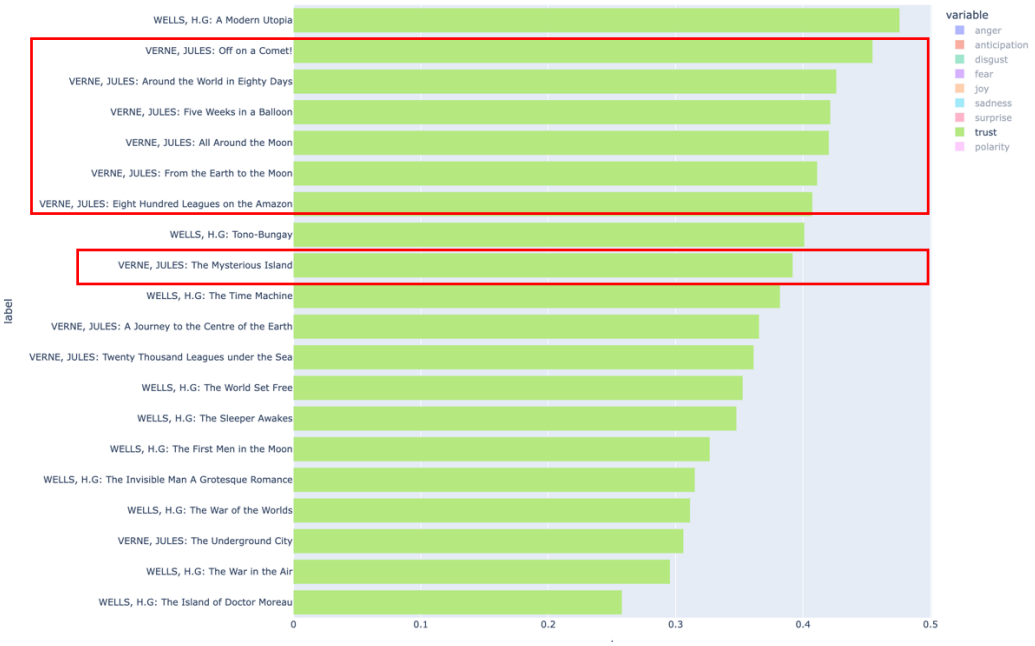

Figure 9: Level of "Trust" by Novel, high scoring Verne novels highlighted

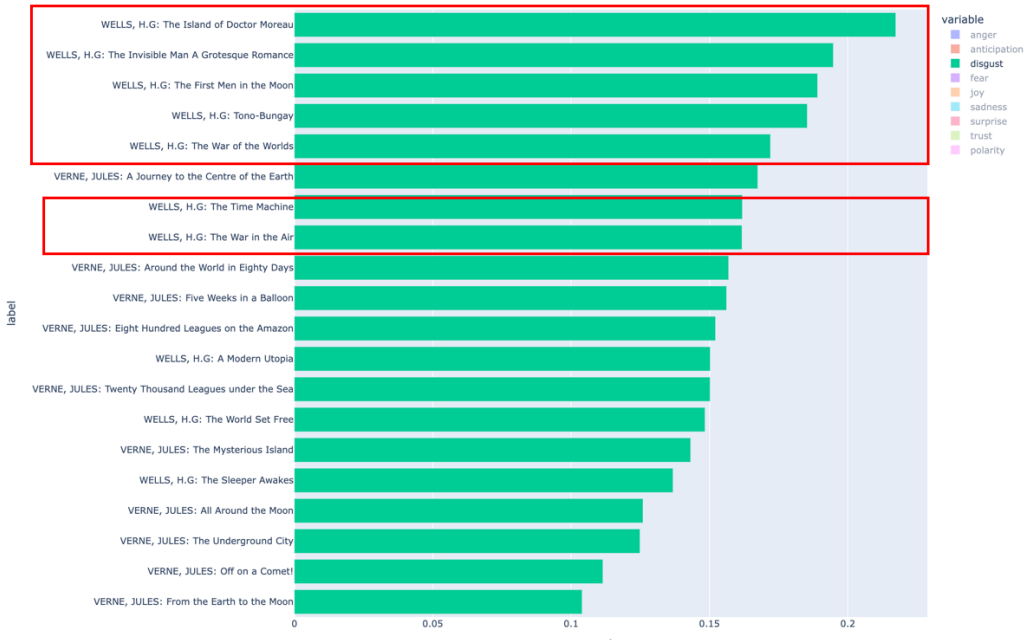

Figure 10: Level of "Disgust" by Novel, high scoring Wells novels highlighted

While many of the other methods (PCA, hierarchical clustering) are mostly useful for exploring the macro trends in the corpus, sentiment analysis (like topic modelling) is also a great tool for identifying how individual works within the corpus compare. Wells' most downloaded book is easily *The Time Machine*, while Verne's is *Around the World in Eighty Days*. When we compare these two novels, we see similar sentiment with trust, fear, and joy all coming in as the top three emotions across both works. This pattern continues throughout with these two books

have the exact same ranking of emotions (Figure 10). That said, *Around the World in Eighty Days* scores much higher than *The Time Machine* in trust giving the book a net positive polarity, while *The Time Machine* had much lower trust value and as a result a net negative polarity.

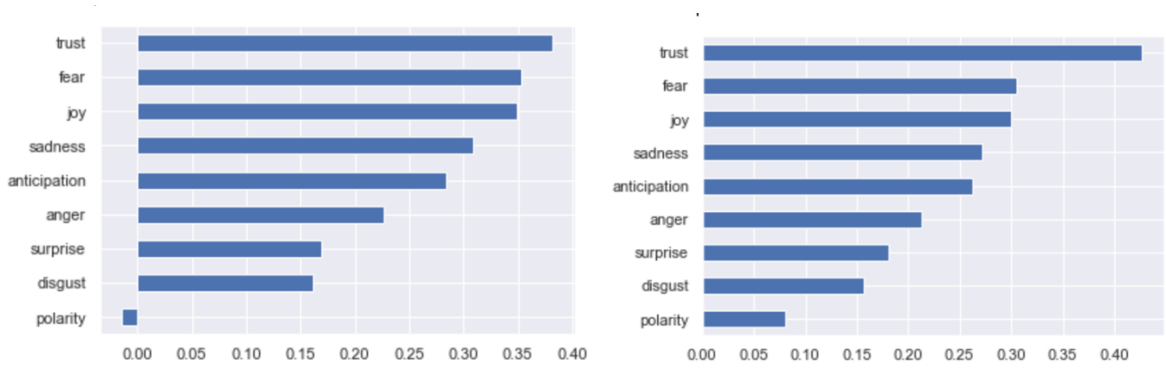

Figure 10: Mean Sentiment for *The Time Machine* (left) & *Around the World in Eighty Days* (right)

The difference between these two works becomes even more interesting when we look at the sentiments by Chapter. We see a largely similar trajectory between the two books until the last few chapters (Figure 11). *The Time Machine* has a largely negative ending in which the Time Traveler never returns, while *Around the World in Eighty Days* has a largely positive one with the protagonist winning the wager the book is centered on and marrying a woman he loves. We see this reflected in the mean sentiment by chapter as trust and overall polarity drop at the end of *The Time Machine* while we see both large increase (along with a sharp drop in fear) in the final chapters of *Around the World in Eighty Days.*

FIGURE 11: Sentiment by Chapter for *The Time Machine* (top) & *Around the World in Eighty Days* (bottom)

Conclusion

Jules Verne and H.G. Wells undoubtedly shaped the genre of Science Fiction in ways that still resonate today. Yet, when we apply a range of exploratory text analytic techniques to a corpus of their most popular works, a few interesting differences emerge. In most measurable ways, Verne's writing varied more from book to book than Wells. While this may be a function of translation, I think it's more likely due to Verne's works coming before Wells. While Science Fiction was still far from a solidified genre while Wells wrote, it had come a long way from giving from the era in which Verne wrote, giving Wells an advantage of some genre-specific structures themes to draw from. The one notable space though where Wells' works showed more range than Verne's was in sentiment analysis. Verne's works largely tended towards negative polarity, but he also had some extremely high positive polarity works. This again may be a function of the times in which these authors wrote, or simply a difference in perspective and style. While this analysis unearthed a few interesting patterns, it also opened a lot more avenues for future research. I think incorporating more works into the corpus as well as a more rigorous comparison of the individual novels within the corpus in terms of both topics and sentiment may yield some very interesting results. Beyond that, incorporating more modern science fiction authors like Isaac Asimov, Arthur C. Clarke, and Anne McCaffrey may help to better understand the genre itself, not just how two particular authors within it compare.

Data: https://virginia.box.com/s/qjumcigrgswjwpiqj7fra78zigsyclfn
GitHub Repo: https://github.com/MacHLeroy/ETAFinalProject.git

References

Roberts, A. C. (2016). Verne and Wells. In *The history of science fiction* (pp. 183–225). essay, Palgrave Macmillan.

Appendix I

title label	The Time Machine	From the Earth to the Moon	The Underground City	The Island of Doctor Moreau	The Invisible Man A Grotesque Romance	The War of the Worlds	Around the World in Eighty Days	The First Men in the Moon	The Sleeper Awakes	A Journey to the Centre of the Earth	Five Weeks in a Balloon	Eight Hundred Leagues on the Amazon	All Around the Moon	The War in the Air	Off on a Comet!	Twenty Thousand Leagues under the Sea	The World Set Free	Tono-Bungay	A Modern Utopia	The Mysterious Island
0: sea mountain rocks vast soil earth enormous rock	0.007145	0.002335	0.000174	0.006583	0.041019	0.000039	0.001897	0.022821	0.006384	0.027874	0.037423	0.043187	0.002718	0.014162	0.000252	0.001152	0.003700	0.029465	0.149081	0.040925
1: doctor balloon hundred wind let country air ground	0.005150	0.004974	0.000174	0.002313	0.000920	0.006892	0.025333	0.001160	0.000013	0.011671	0.011766	0.001089	0.001466	0.519738	0.003185	0.000025	0.003062	0.004659	0.000443	0.009959
2: old years new mine want yes cried let	0.002675	0.002620	0.003330	0.006962	0.009116	0.035185	0.017628	0.004821	0.076588	0.005515	0.045780	0.295229	0.011611	0.005789	0.011893	0.000025	0.063006	0.008813	0.011268	0.002848
3: island engineer replied granite house companions top sea	0.000213	0.000148	0.013126	0.038076	0.023708	0.003168	0.018029	0.000139	0.000349	0.551859	0.048218	0.072374	0.004254	0.011116	0.000252	0.002529	0.001250	0.005326	0.024971	0.007988
4: captain sea land sir water vessel ocean board	0.000213	0.003972	0.063918	0.073501	0.524389	0.007496	0.006887	0.002550	0.000013	0.046159	0.064643	0.005802	0.004971	0.013374	0.000252	0.002531	0.001174	0.032932	0.021831	0.025001
5: people face stood turned suddenly heard white towards	0.428019	0.422486	0.009380	0.575969	0.040702	0.155539	0.181154	0.325888	0.093255	0.014801	0.017629	0.028766	0.015356	0.017084	0.592264	0.013850	0.470778	0.011363	0.018230	0.014578
6: gun replied shot iron american point weight metal	0.000213	0.003560	0.013404	0.000190	0.015934	0.000787	0.026094	0.002593	0.000013	0.010591	0.006215	0.000150	0.009132	0.020551	0.000928	0.000440	0.000104	0.012956	0.003609	0.347650
7: lake town country mountains miles mountain reached city	0.000213	0.001545	0.039408	0.000190	0.005857	0.000039	0.001781	0.000139	0.008142	0.004530	0.015941	0.050089	0.017421	0.039061	0.000252	0.013658	0.004503	0.025148	0.011176	0.016815
8: got mr get london dont old im house	0.052530	0.037550	0.275404	0.017485	0.011587	0.288520	0.247417	0.037539	0.064586	0.010129	0.016959	0.023257	0.021996	0.051809	0.217000	0.003287	0.028431	0.038425	0.027517	0.019333
9: cried professor words let earth water appeared heard	0.018434	0.027163	0.009837	0.009432	0.056223	0.003779	0.000842	0.024469	0.005806	0.034762	0.039581	0.105149	0.044228	0.063086	0.001841	0.002601	0.012229	0.072037	0.481076	0.031994
10: captain professor earth sun days new temperature degrees	0.016106	0.011770	0.027765	0.008351	0.043497	0.000039	0.000033	0.027171	0.002068	0.031010	0.542833	0.053838	0.022005	0.055828	0.003250	0.001301	0.006263	0.023642	0.095515	0.106598
11: things world thing life sort mind something people	0.134765	0.044084	0.004842	0.157951	0.012363	0.411177	0.068706	0.230119	0.103110	0.010676	0.009675	0.031968	0.031557	0.024111	0.038949	0.448840	0.103137	0.010086	0.013950	0.005919
12: room soil worthy house journey four kind whose	0.000213	0.000148	0.000174	0.000190	0.000126	0.000039	0.000033	0.000139	0.000013	0.000081	0.000150	0.000150	0.000134	0.000137	0.000252	0.000025	0.000104	0.000065	0.022804	0.000240
13: river trees hundred bank large whose forest several	0.040327	0.025903	0.072984	0.021597	0.076715	0.004865	0.000488	0.005385	0.001640	0.128643	0.014923	0.045595	0.298740	0.068791	0.002176	0.006861	0.000104	0.018544	0.023520	0.056534
14: world state people life new human things women	0.108372	0.072853	0.016723	0.018752	0.017080	0.022145	0.096656	0.023994	0.400614	0.005193	0.015992	0.030725	0.031826	0.020821	0.015510	0.485933	0.101089	0.022955	0.030302	0.043952
15: moon earth captain cried surface miles point travellers	0.034235	0.008687	0.005482	0.003812	0.007932	0.000039	0.000343	0.128621	0.000013	0.002596	0.029675	0.040292	0.000134	0.012611	0.006488	0.001219	0.000736	0.648383	0.008198	0.141078
16: air sky night black towards water became wind	0.142631	0.304523	0.021063	0.049788	0.031284	0.060134	0.302341	0.150939	0.231809	0.014725	0.051720	0.022803	0.009150	0.029066	0.032334	0.007474	0.182743	0.009261	0.022170	0.041404
17: replied judge answered yes life exclaimed words years	0.000216	0.009616	0.076780	0.004608	0.041843	0.000039	0.000033	0.000817	0.005555	0.063628	0.028065	0.047371	0.446221	0.019857	0.011498	0.004882	0.016769	0.005717	0.021436	0.042634
18: water pressure surface iron glass lower level depth	0.001970	0.003264	0.001137	0.003349	0.028915	0.000039	0.000033	0.004270	0.000013	0.007830	0.002662	0.005923	0.008801	0.007217	0.005445	0.001475	0.000714	0.006697	0.008953	0.009747
19: mr sir replied minutes twenty hours master oclock	0.006362	0.012799	0.344895	0.000902	0.010790	0.000039	0.004272	0.006425	0.000013	0.017729	0.000150	0.096244	0.018281	0.005790	0.055977	0.001892	0.000104	0.013525	0.003950	0.034803