**MacKenzye Leroy**

[zuf9mc@virginia.edu](mailto: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.

**Chart, box and whisker chart

Description automatically generated**

**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.

**PCA**

Chart, scatter chart

Description automatically generated

**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.

Chart, bar chart

Description automatically generated

**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).

Table

Description automatically generatedTable

Description automatically generated

**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.

A picture containing chart

Description automatically generated A picture containing chart

Description automatically generatedA picture containing chart

Description automatically generatedA picture containing chart

Description automatically generated

**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*.

**Chart

Description automatically generated**

**Chart, bar chart

Description automatically generated**

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

**Chart

Description automatically generated**

**Chart

Description automatically generated**

**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**

Chart, bar chart

Description automatically generated

**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).

Chart

Description automatically generated

**Figure 9:**  Level of “Trust” by Novel, high scoring Verne novels highlighted

Chart

Description automatically generated

**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.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

**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.*

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

**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

Graphical user interface, application, table

Description automatically generated