Text Analysis

Past versus Present

Science Fiction Novels

By

Julian Escasa & Matt Schindler

Introduction

Project Description: Text analysis of past versus present science fiction books. The publication timeline of the old science fiction books will be from the late 19th century to early 20th (1850-1915), while present books will come from the 21st century (2000-2015). Resulting argument will be based on textual differences such as, text patterns, frequency of specific words, and grammatical relationships.

Argument Question: "How has the language, syntax, and grammar of science fiction novels changed over time?"

Significance and Themes: This project will give a better understanding about the evolution of science-fiction writing styles and diction. We'll also be familiarized with the use of text analysis software such as KH Coder, and online libraries like project Gutenberg and 'other' resources.

Sample and Procedure: To start, pre-existing coding software was used to analyze all texts (KH Coder). The sample size is fourteen books, seven old and seven recent. Next, looked for books that can be easily converted into .txt or .pdf file. Next, using the KH Coder software, conducted text analysis of all books. After the data compilation, the resulting three graphs generated per book (Multi-Dimensional Scaling, Hierarchical Cluster Analysis, and Co-Occurrence Network) was then qualitatively analyzed to determine any relevant differences between past and present books.

Books: The selection is fourteen popular science fiction books (seven each), based on reviews, popularity, and reader ratings. The following is a list of the books used, then subsequently, the text analysis results of each book.

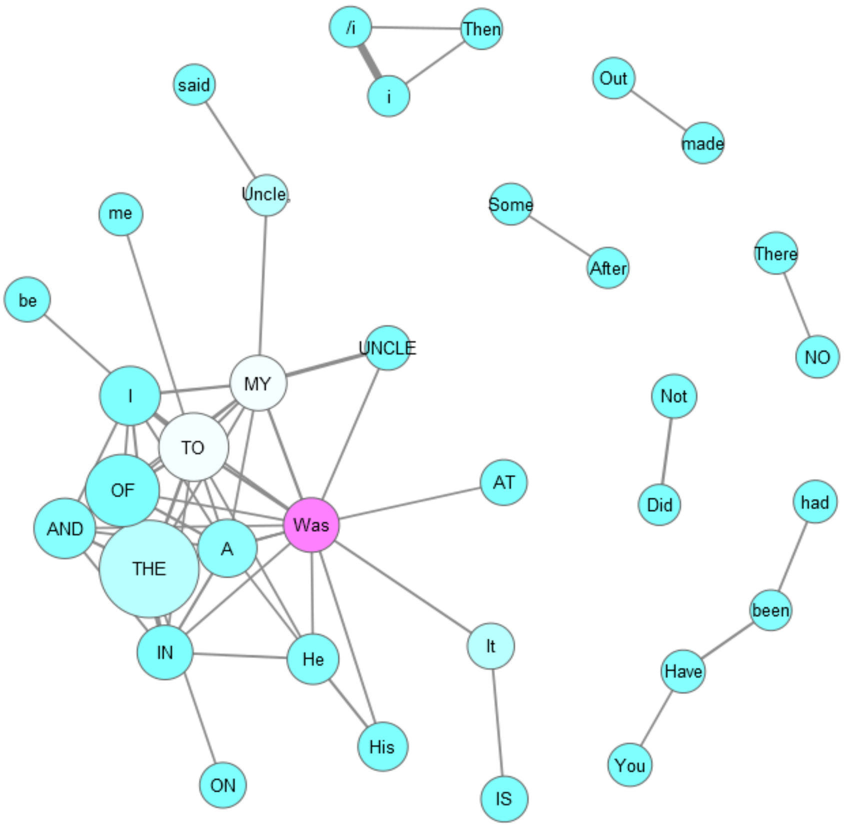
Past

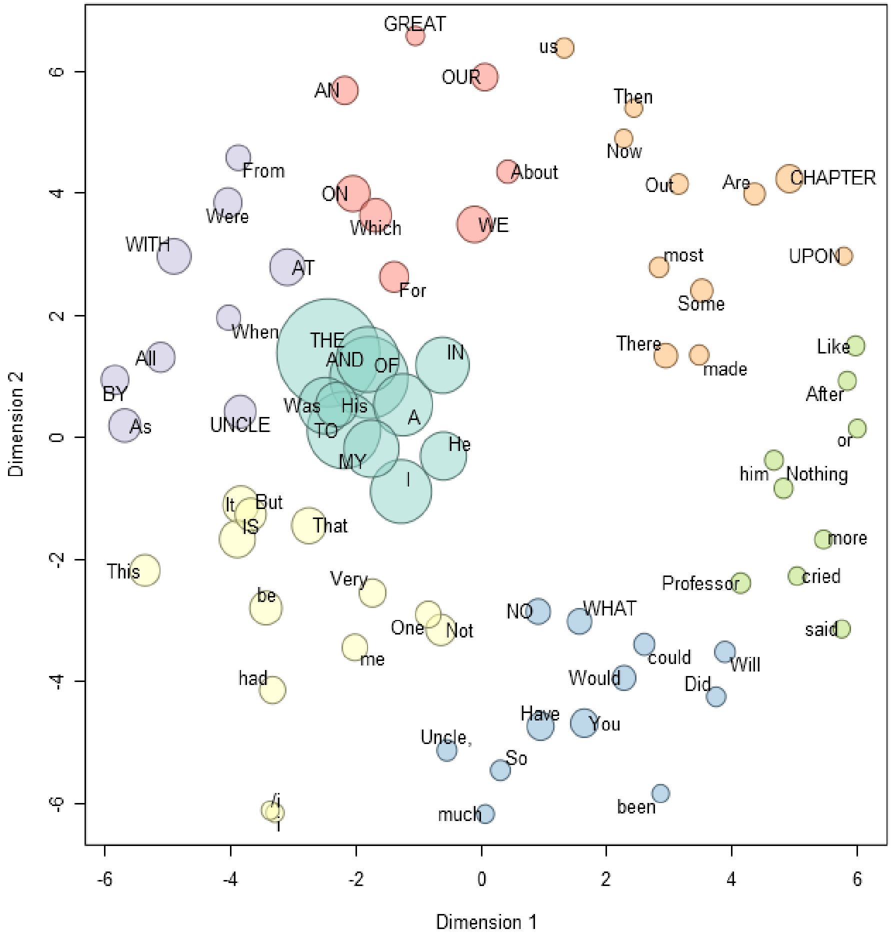
Journey to the Center of the Earth (1894) - Jules Verne
Frankenstein (1818) - Mary Shelley
Strange Case of Dr. Jekyll and Mr. Hyde (1886) - Robert Stevenson
The Coming Race (1871) - Edward Bulwer-Lytton
The Invisible Man (1897) - HG Wells
The Time Machine (1895) - HG Wells
The War of the Worlds (1897) - HG Wells

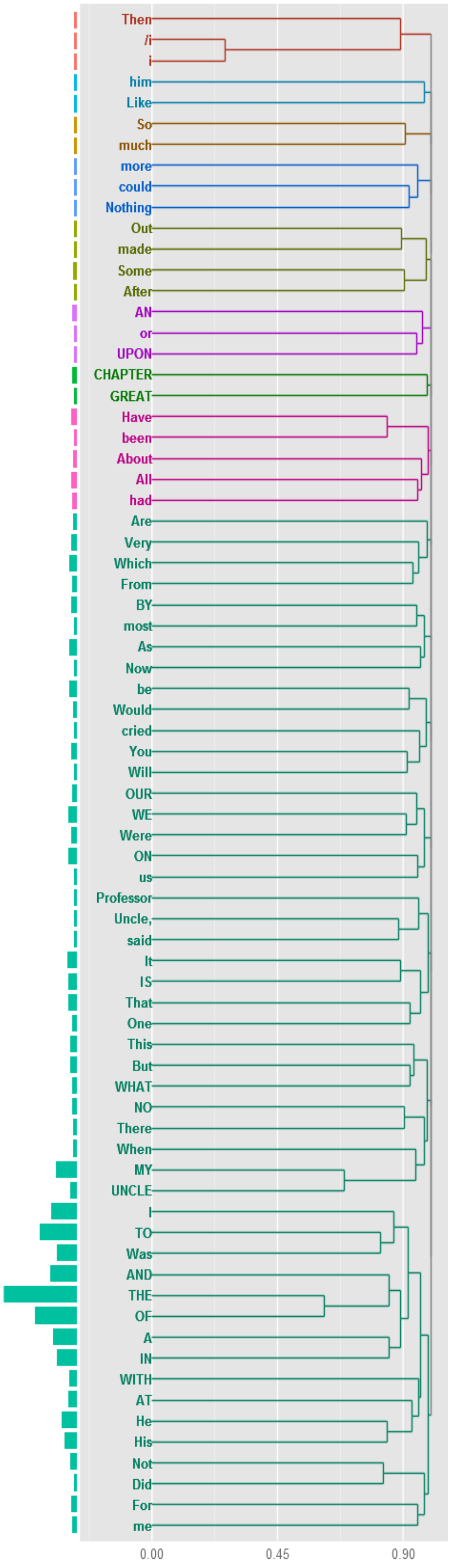
Present

11/22/63 (2011) - Stephen King
House of Suns (2008) - Alastair Reynolds
Ready Player One (2011) - Ernest Cline
Snow Crash (1995) - Neal Stephenson
The Martian (2011) - Andy Weir
The Maze Runner (2009) - James Dashner
Under the Dome (2009) - Stephen King

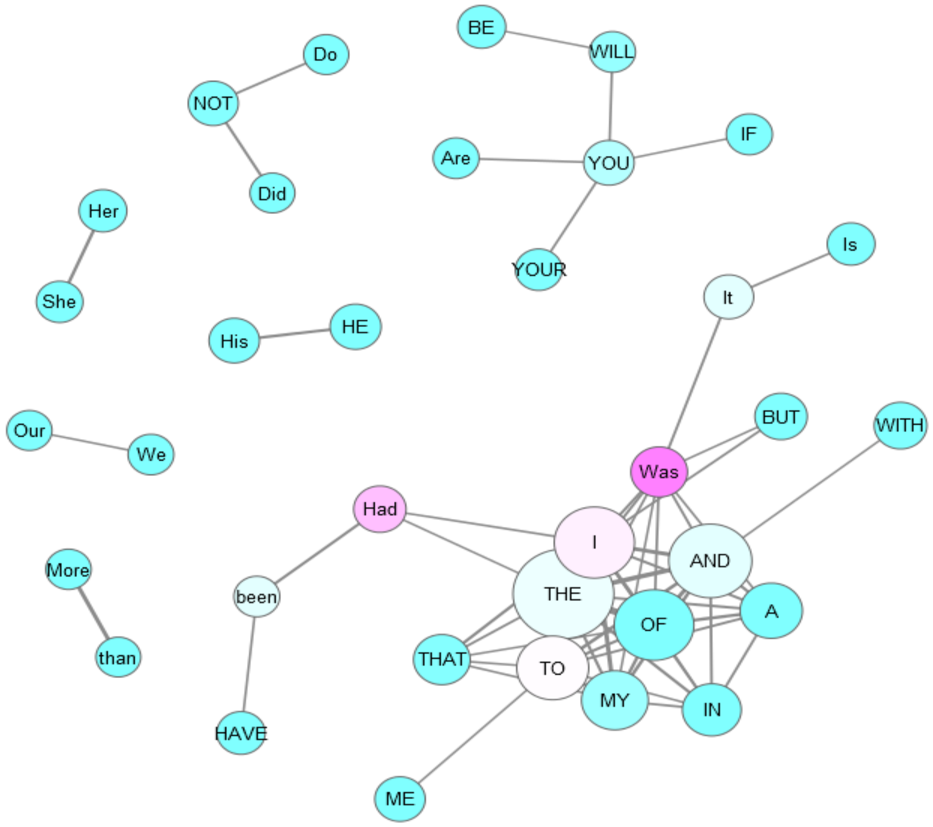
Past: A Journey to the Center of the Earth (1864)
By Jules Verne

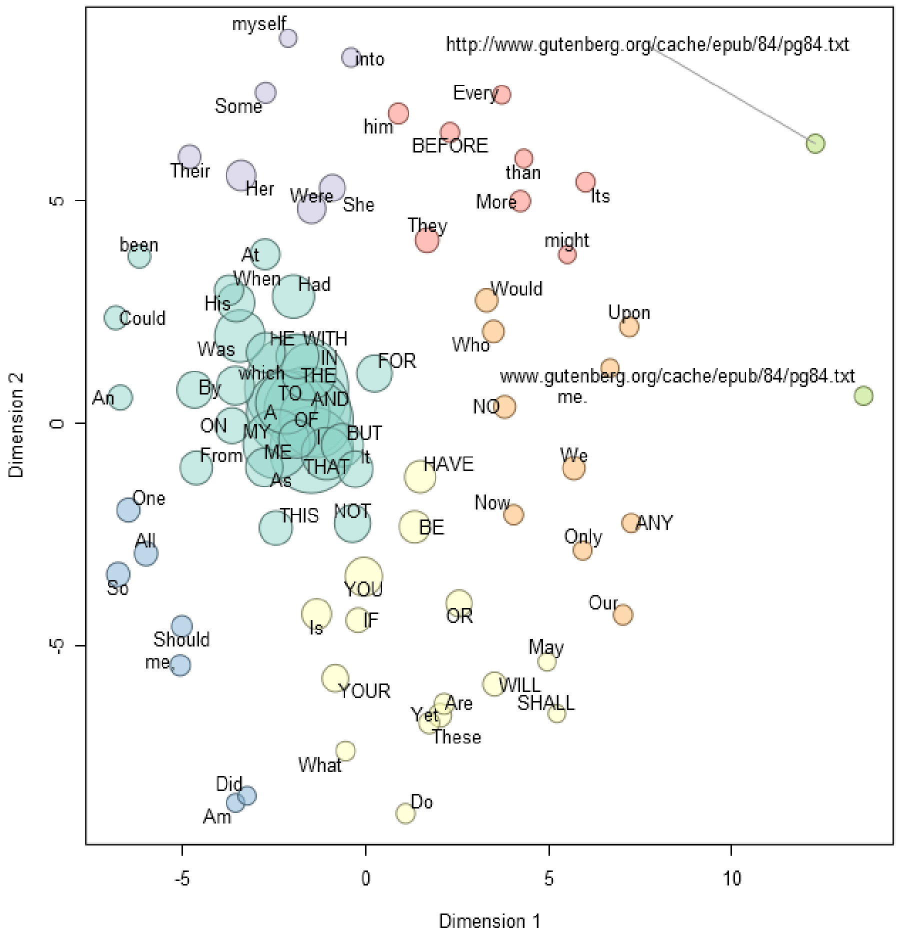


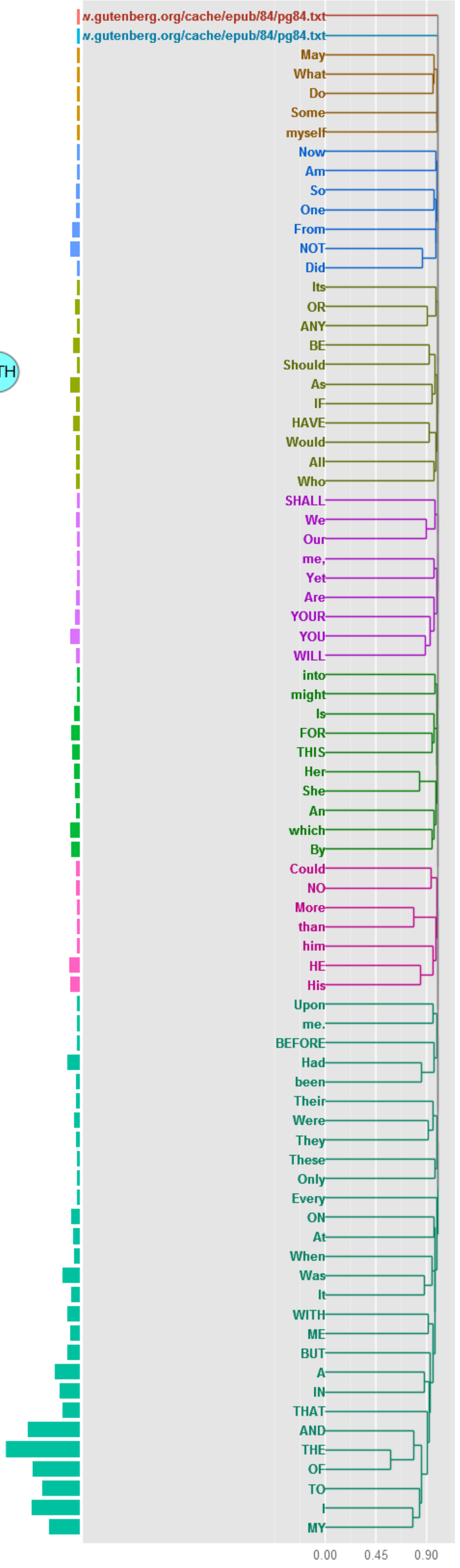




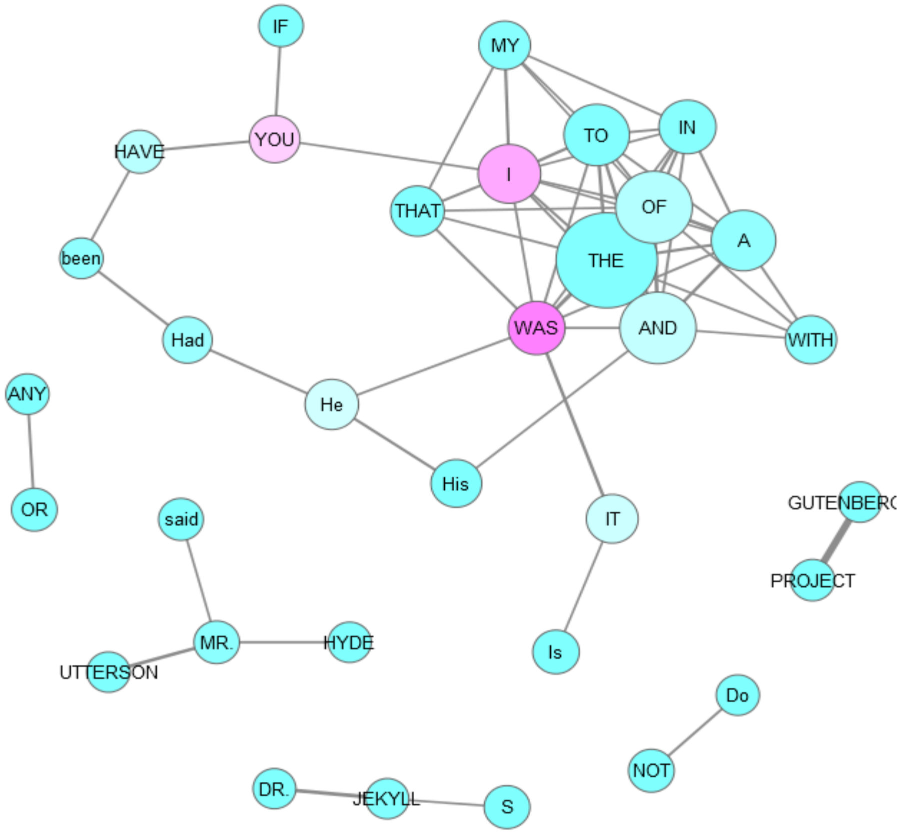
Past: Frankenstein (1818)
By Mary Shelley

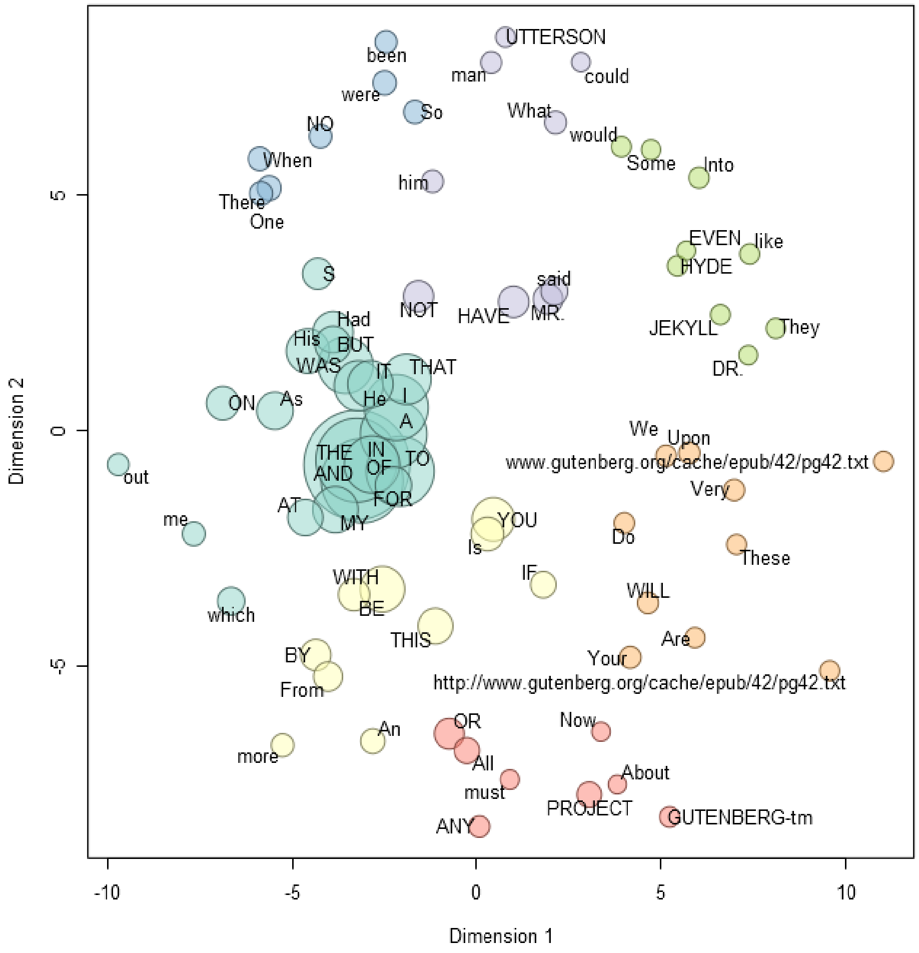


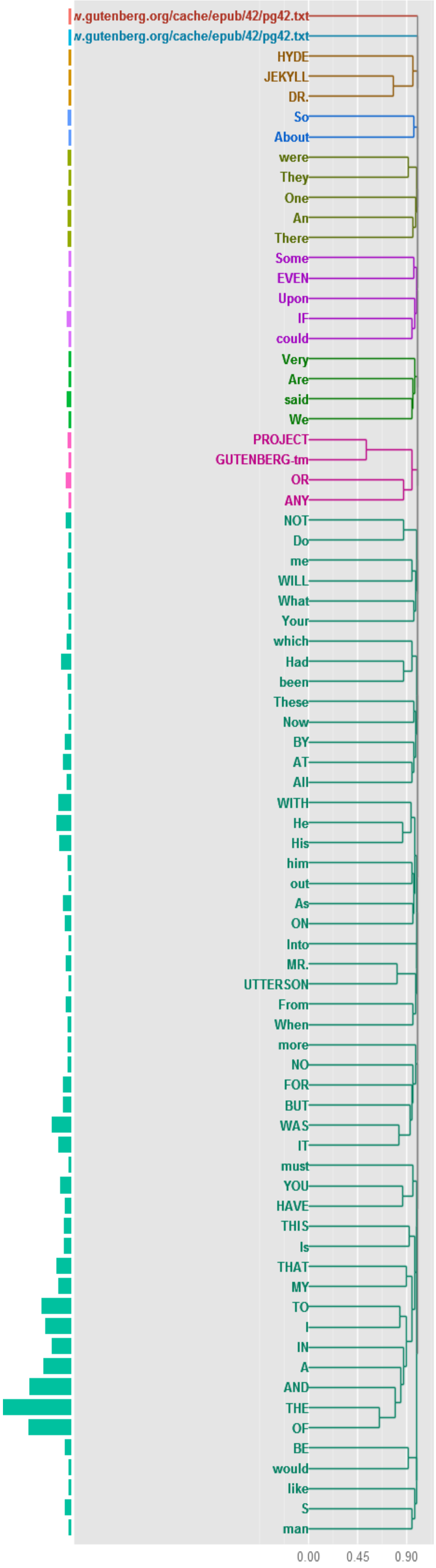




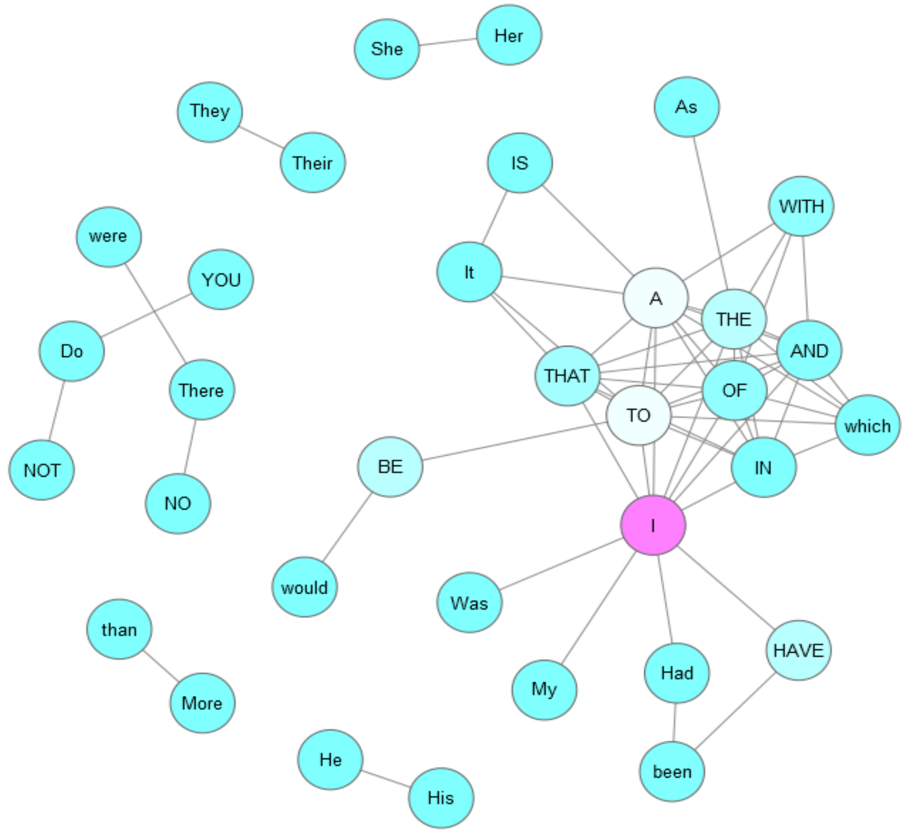
Past: Strange Case of Dr. Jekyll and Mr. Hyde (1886)
By Robert Stevenson

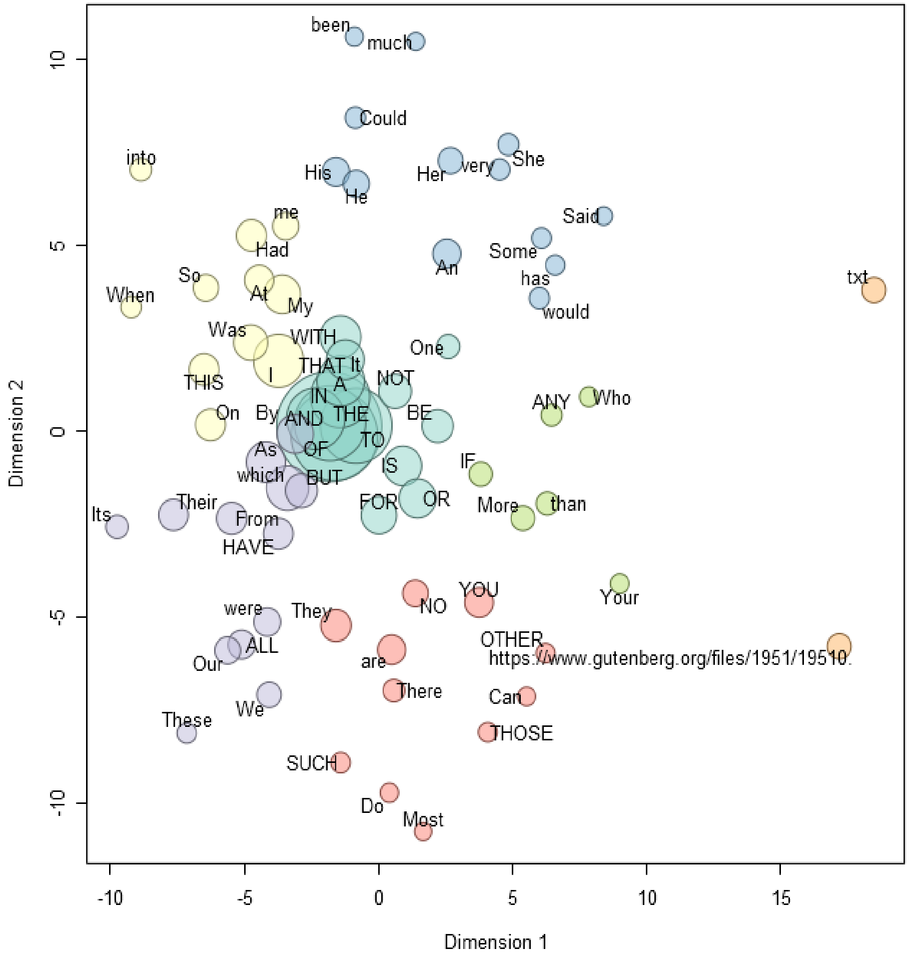


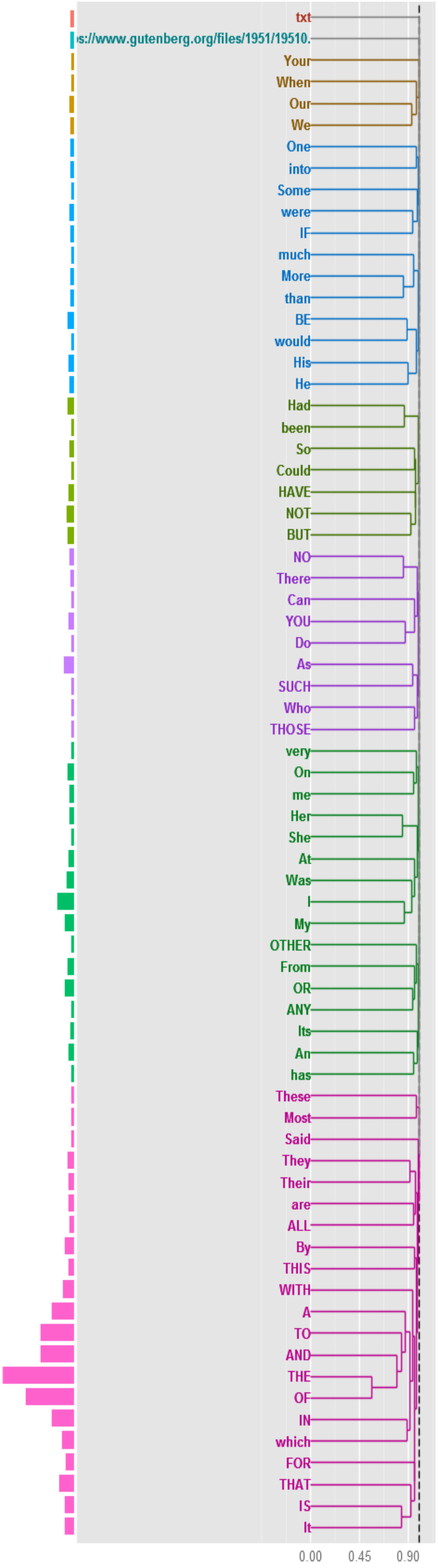




Past: The Coming Race (1871)
By Edward Bulwer-Lytton

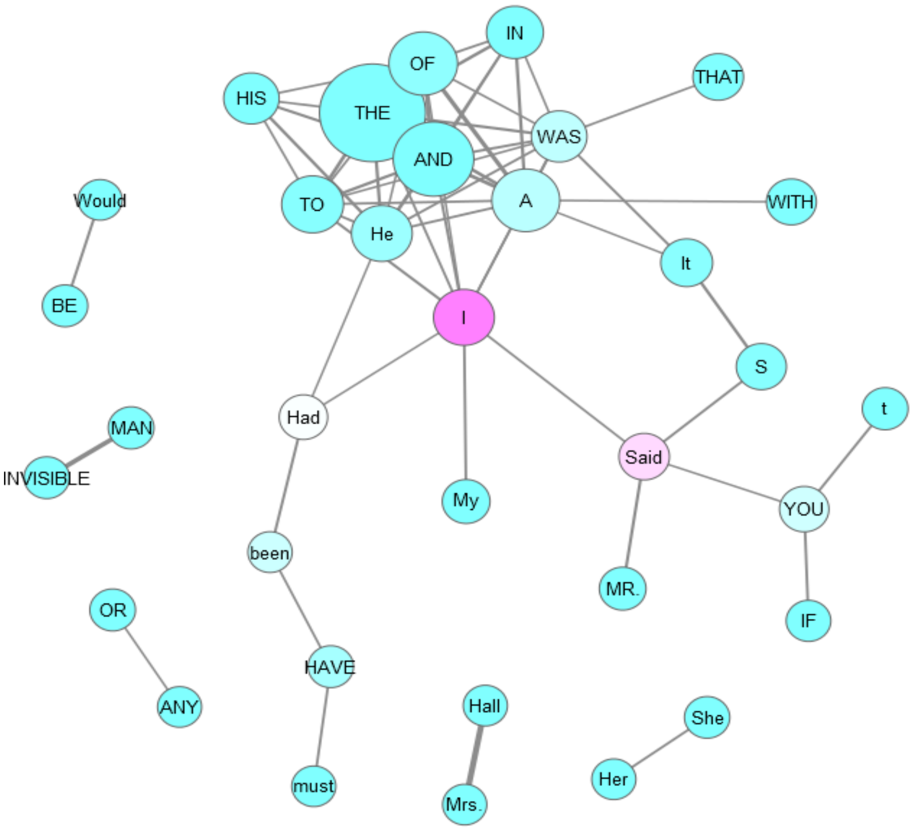


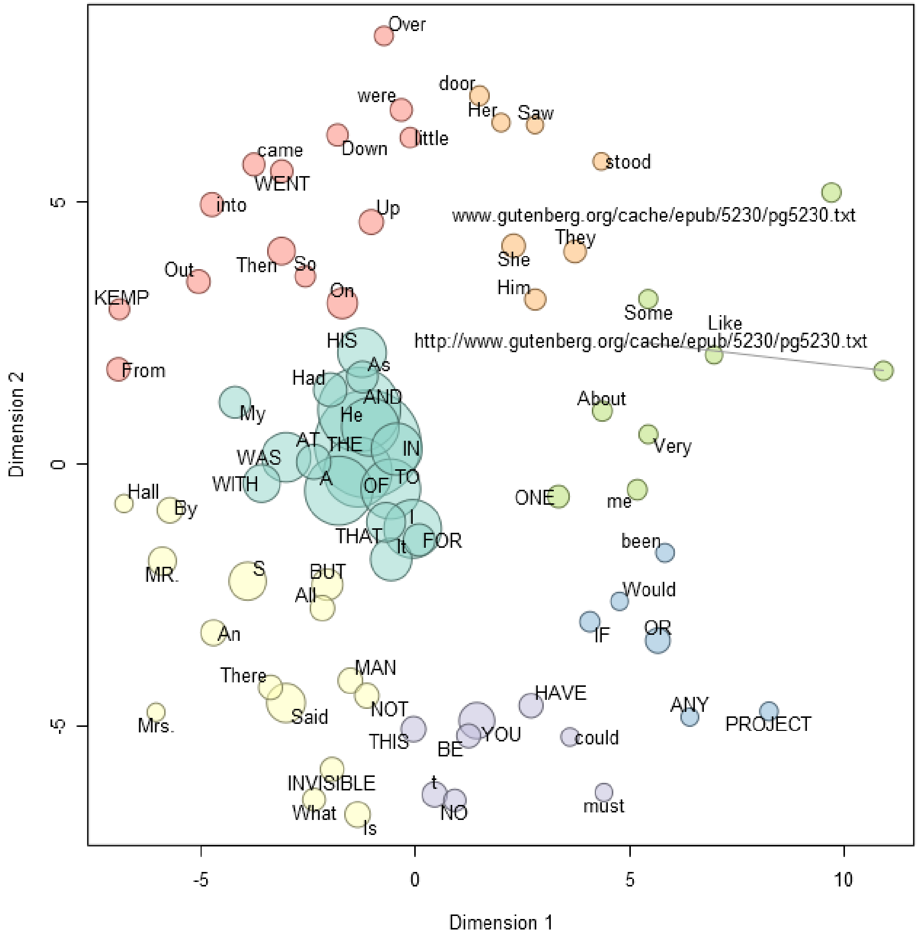


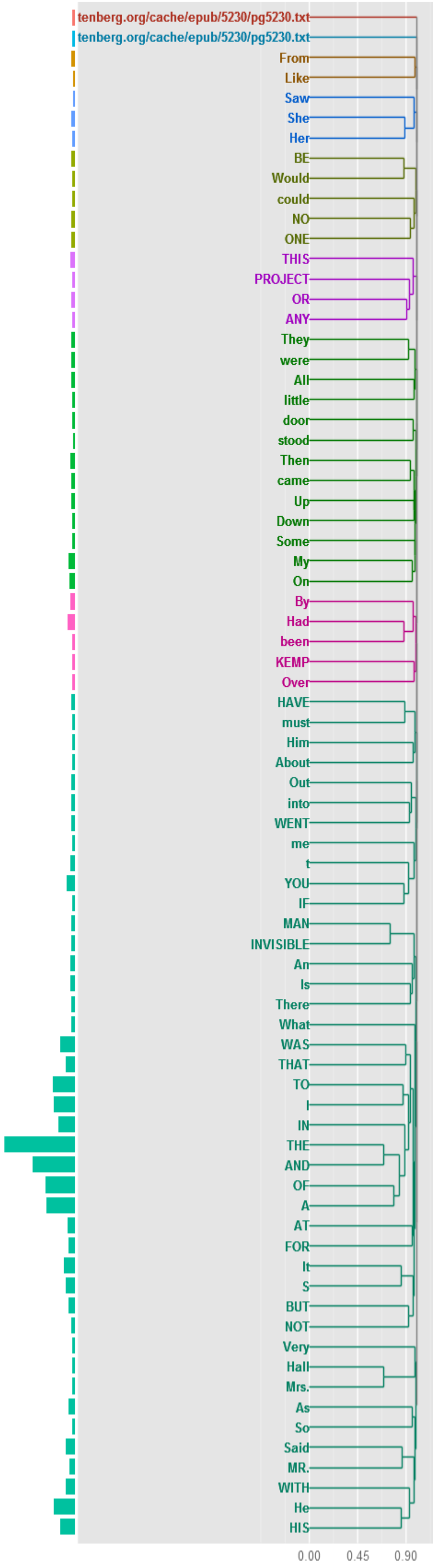


Past: The Invisible Man (1897)

By HG Wells

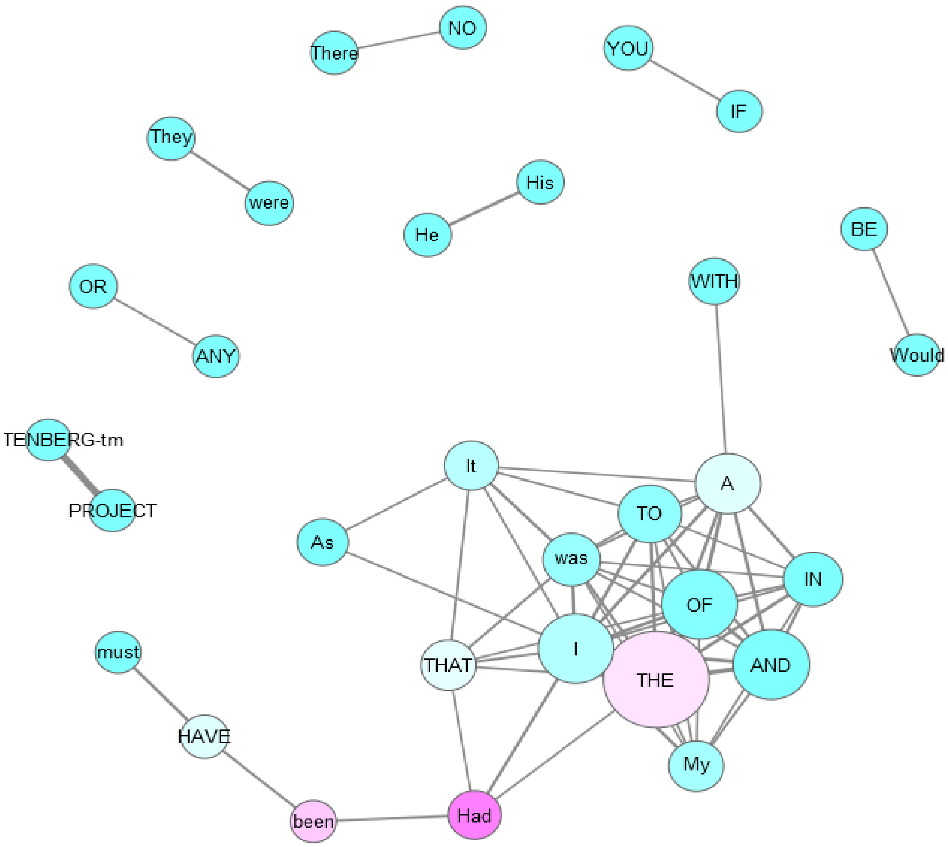


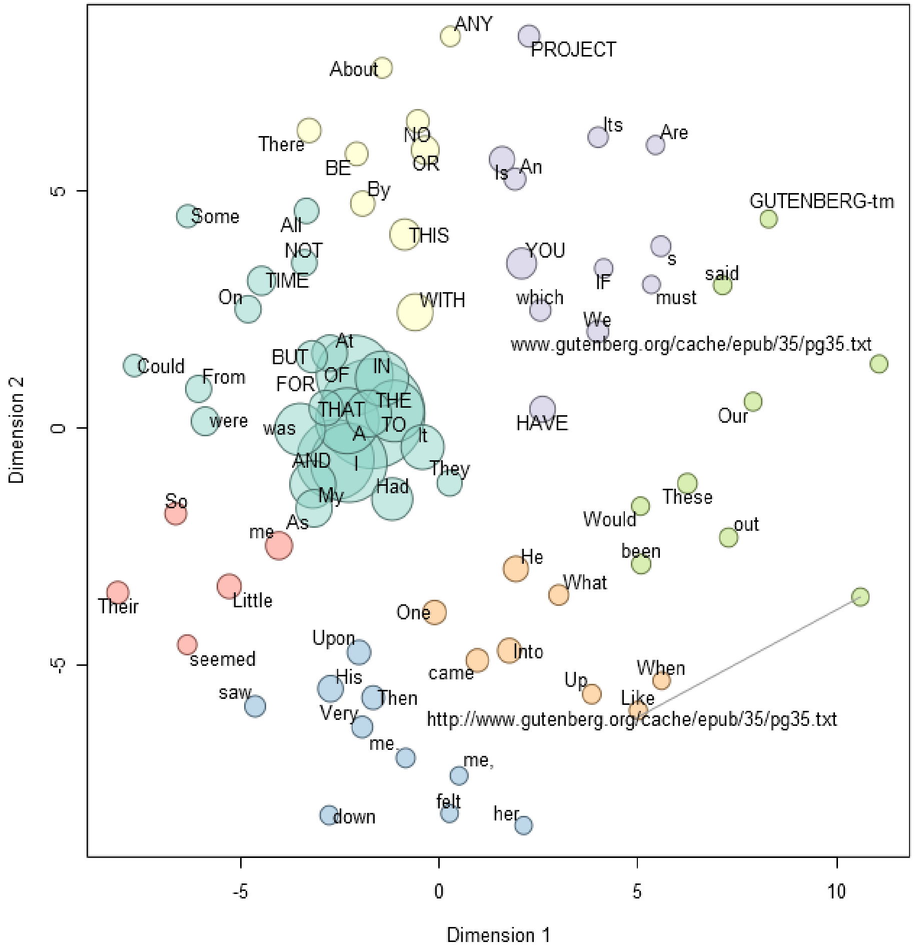


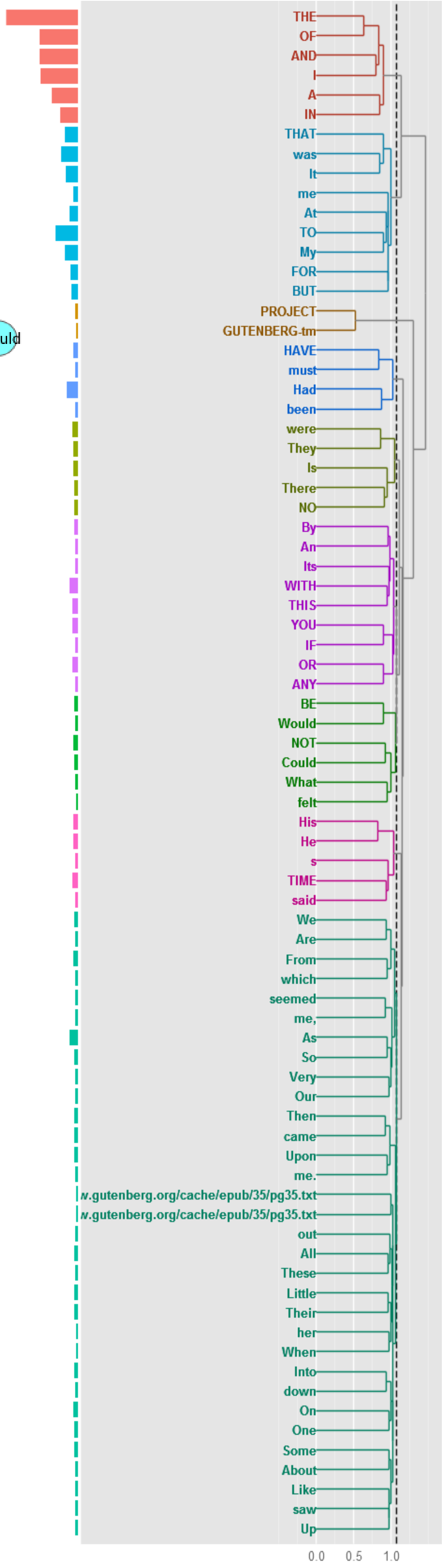


Past: The Time Machine (1895)

By HG Wells

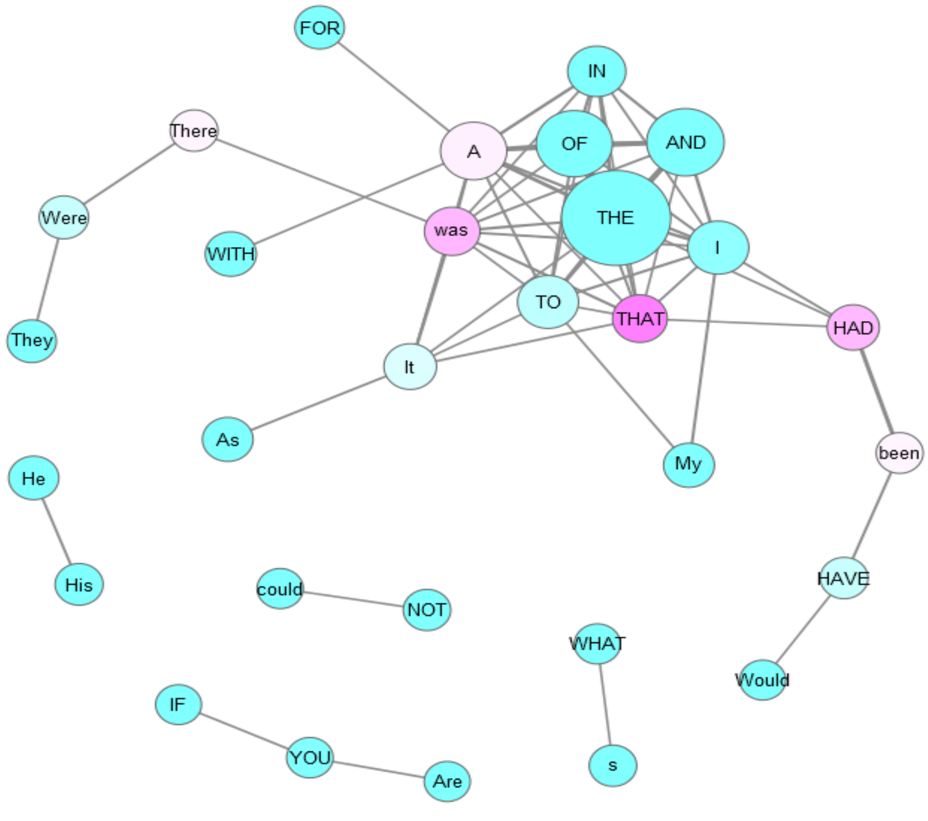


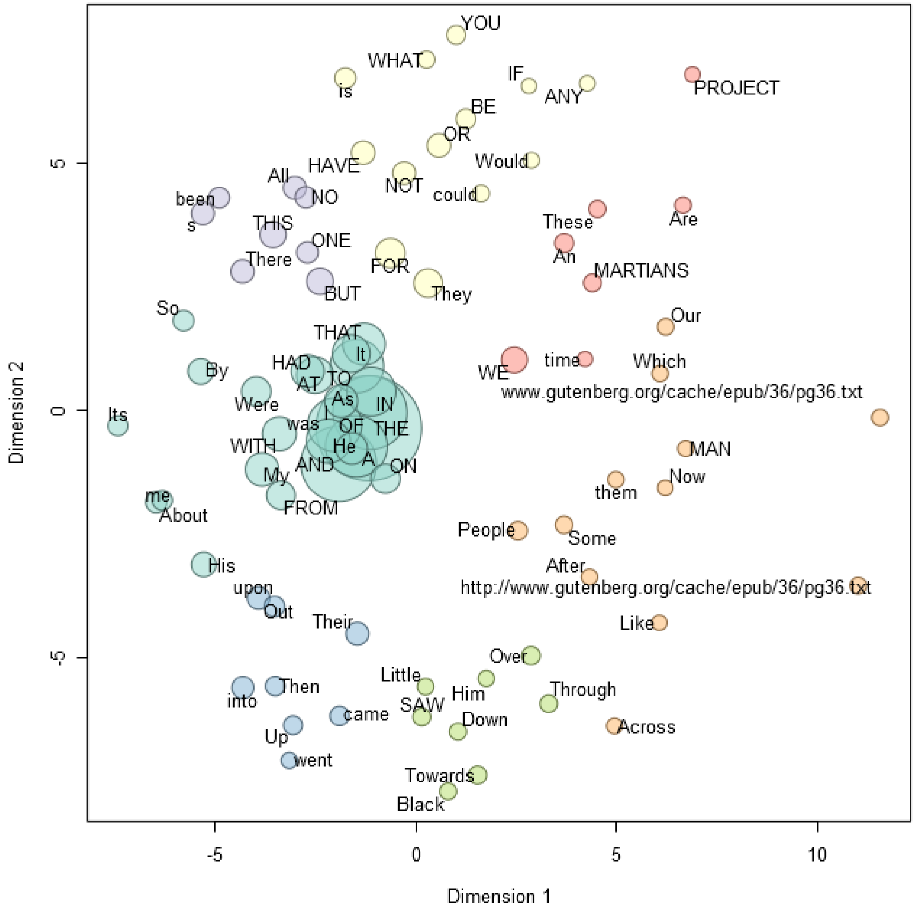


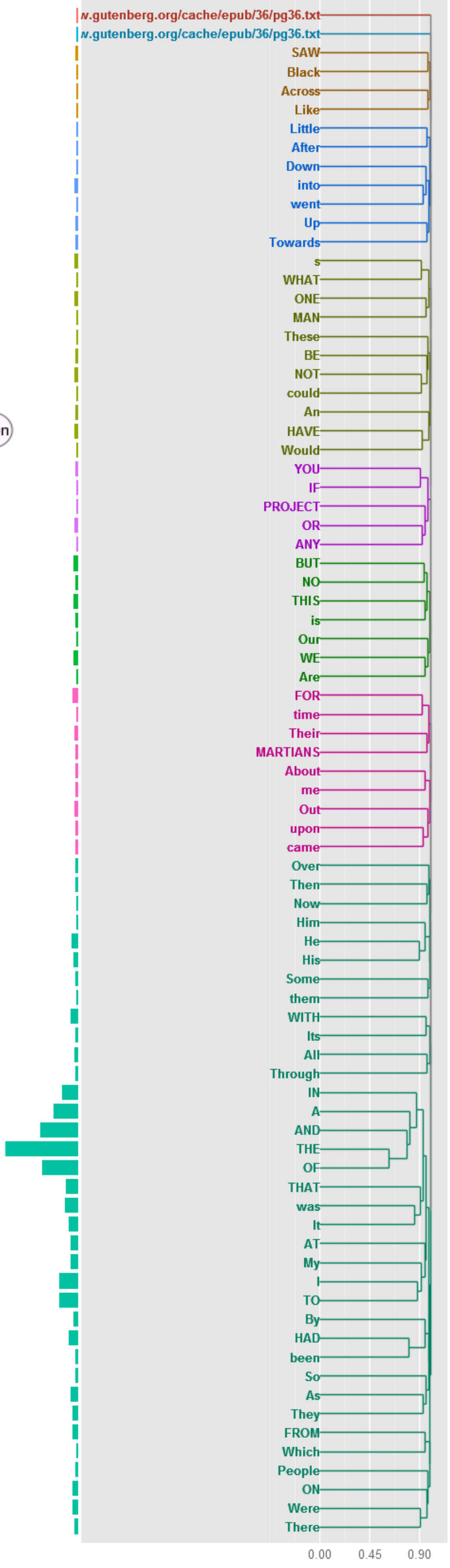


Past: The War of the Worlds (1897)

By HG Wells

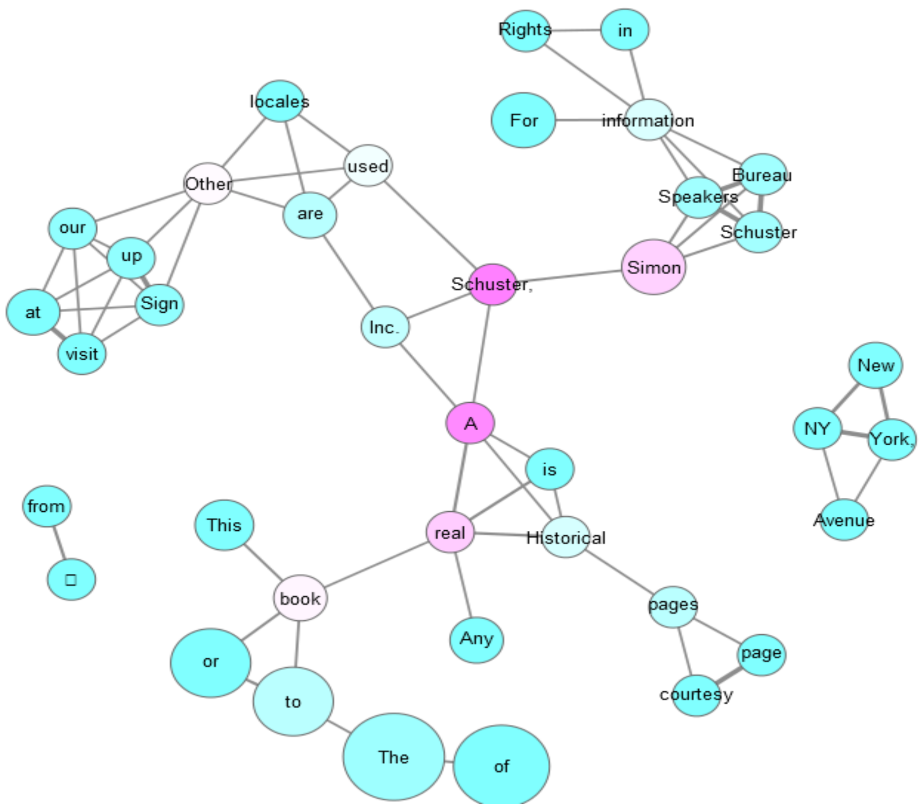


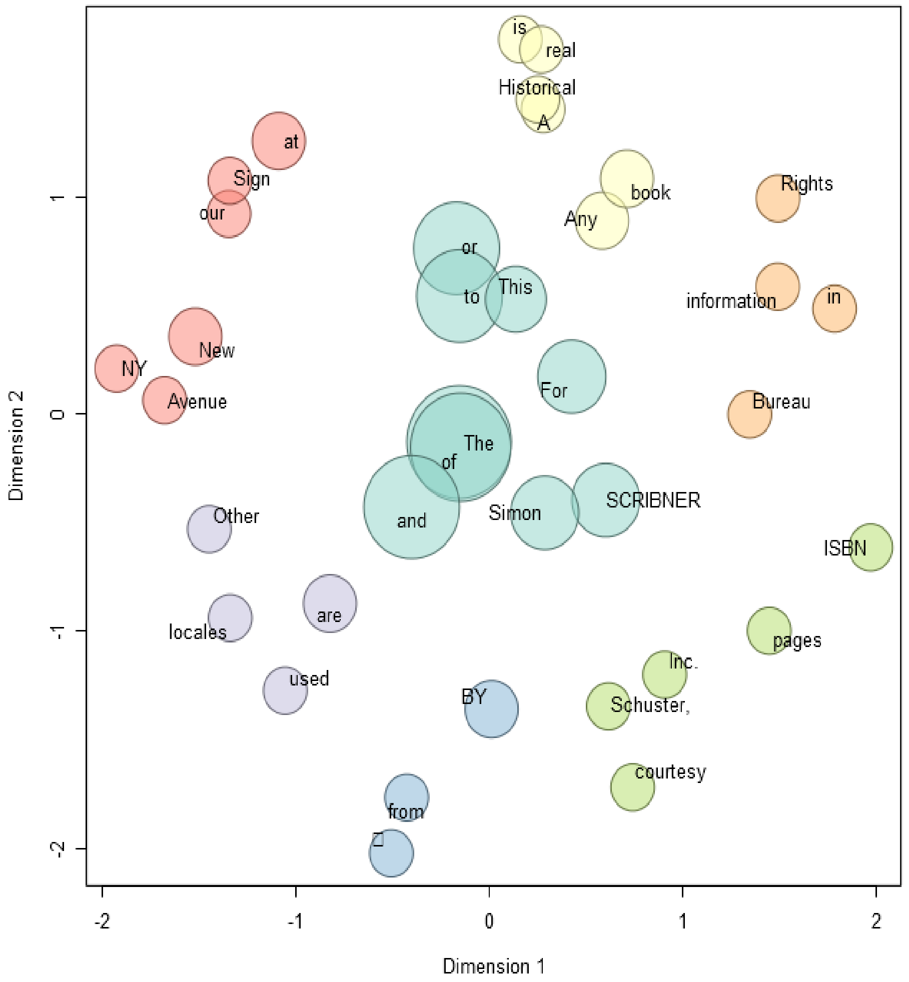


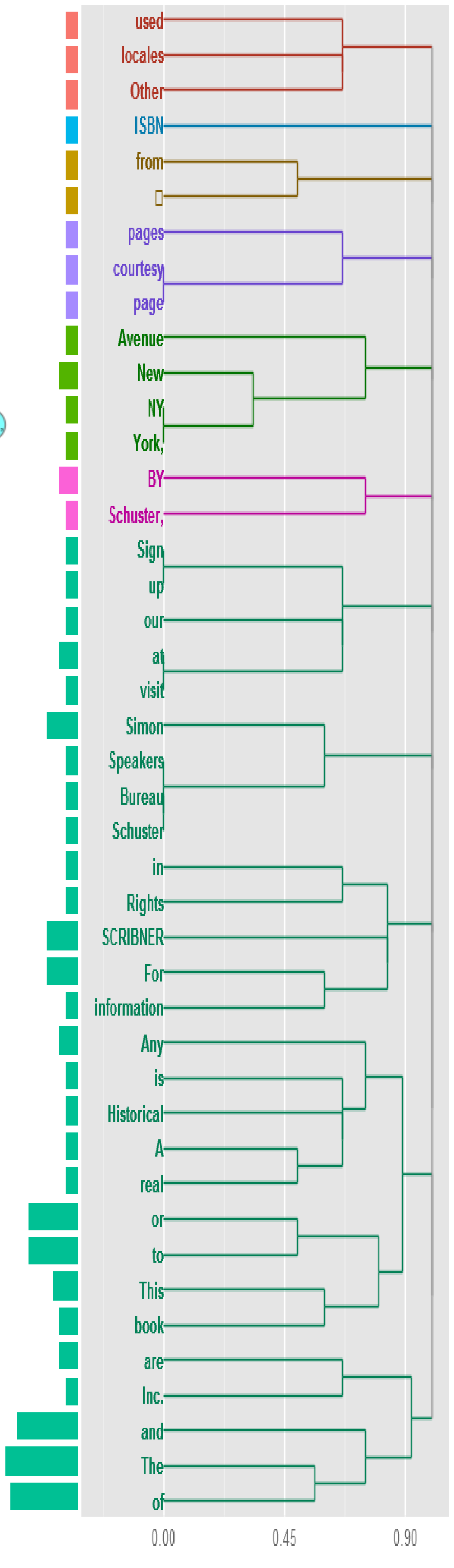


Present: 11/22/63 (2011)

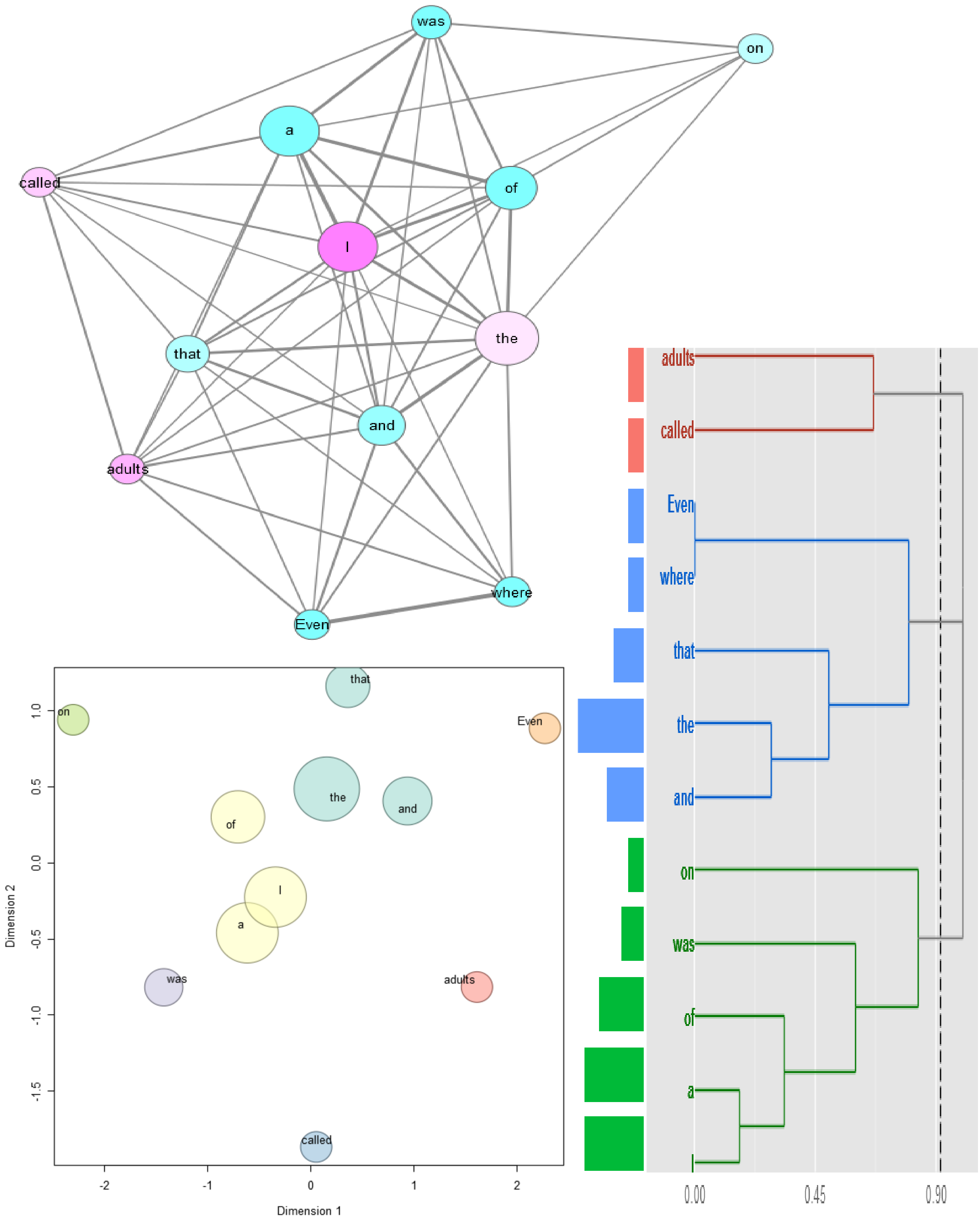
By Stephen King



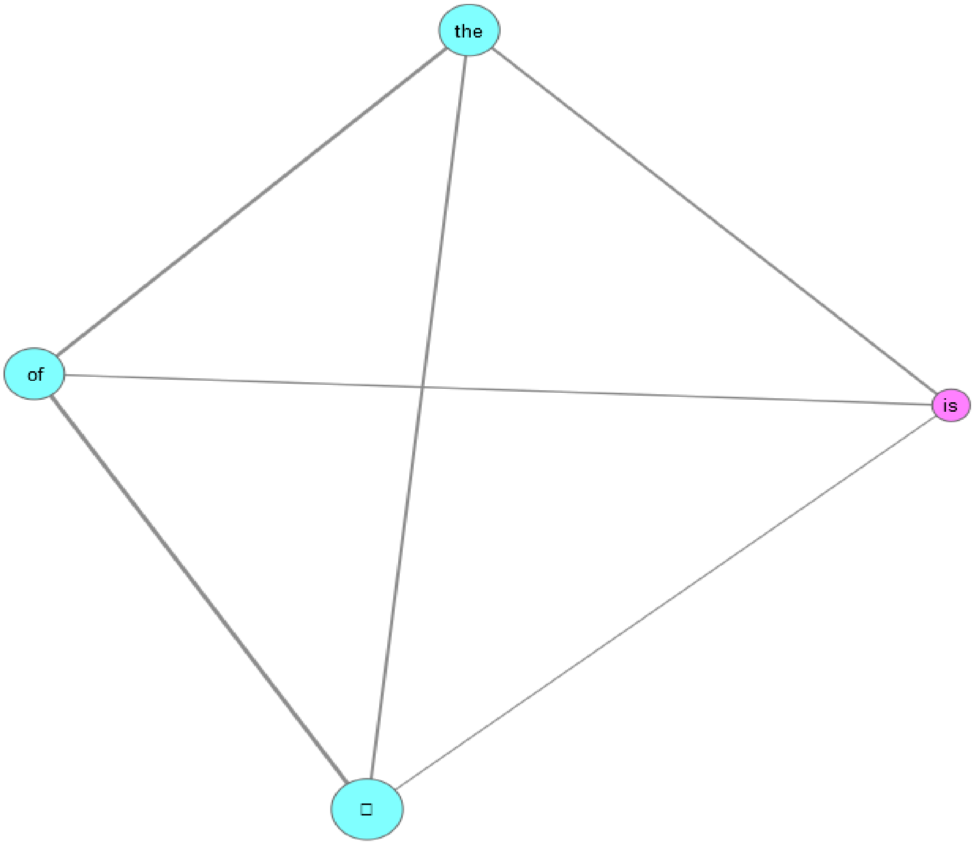


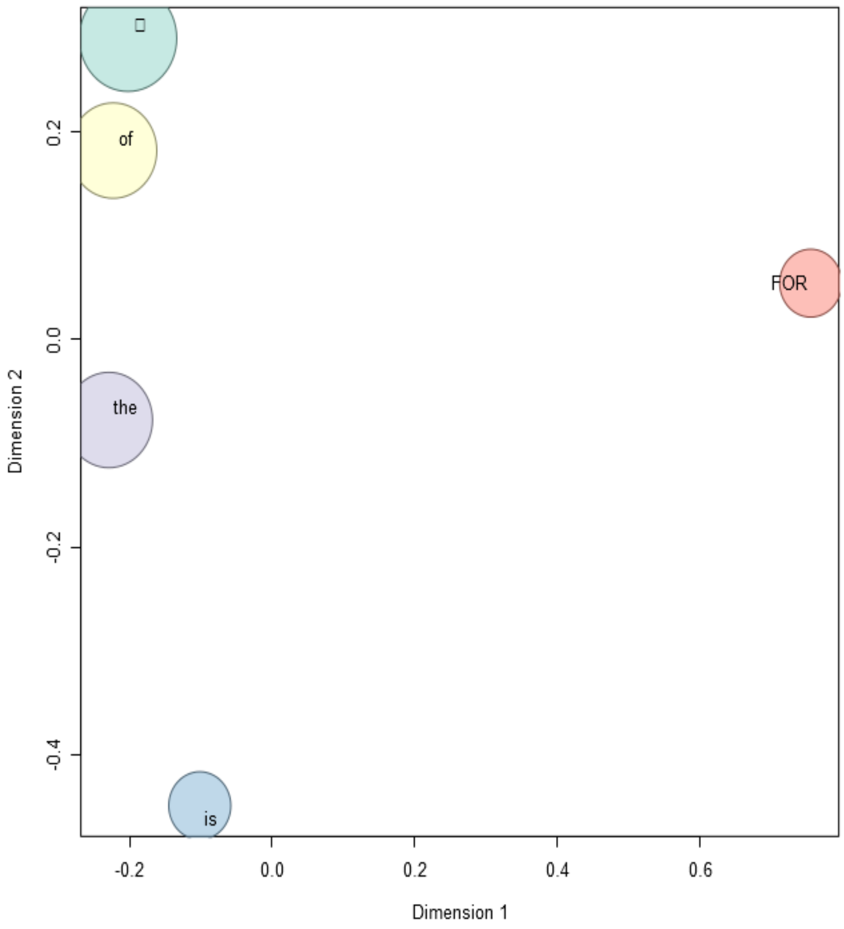


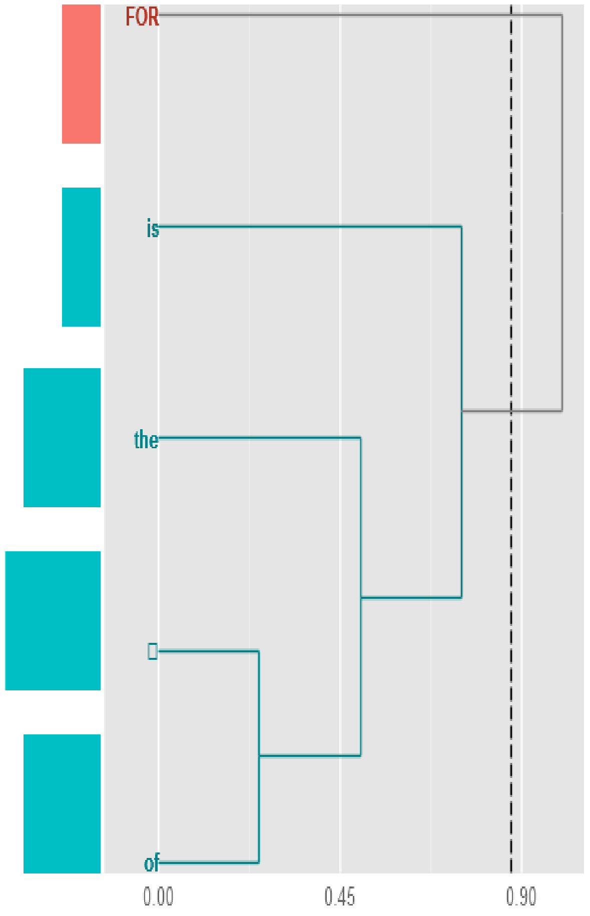
Present: House of Suns (2008)
By Alastair Reynolds



Present: Ready Player One (2011)
By Ernest Cline

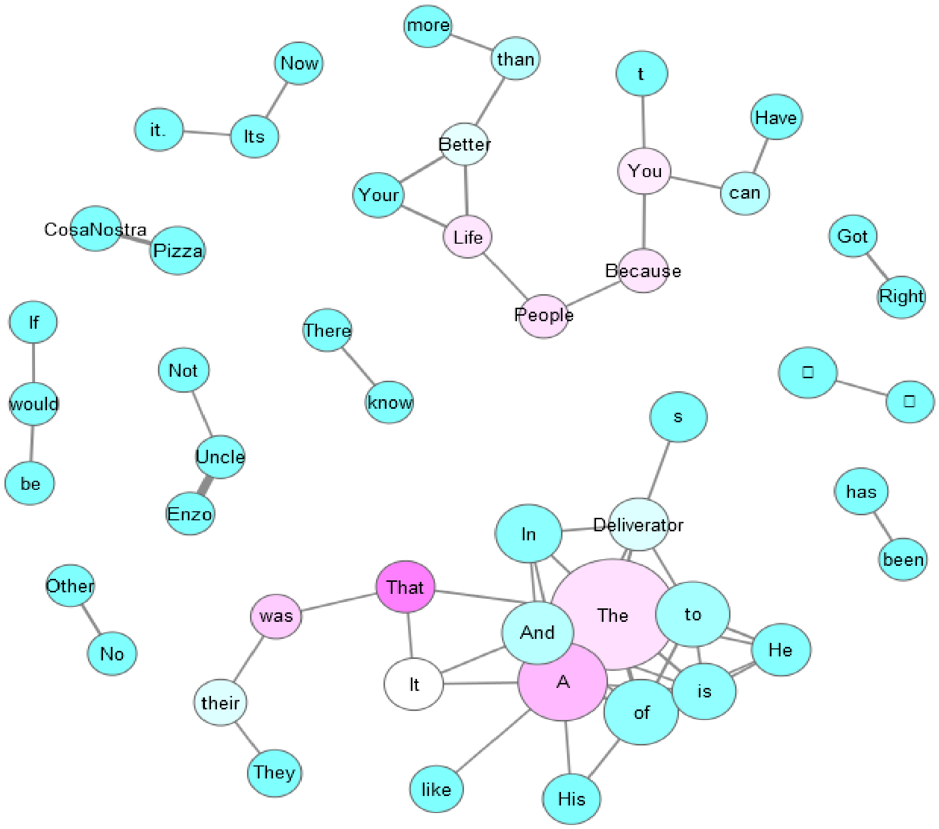


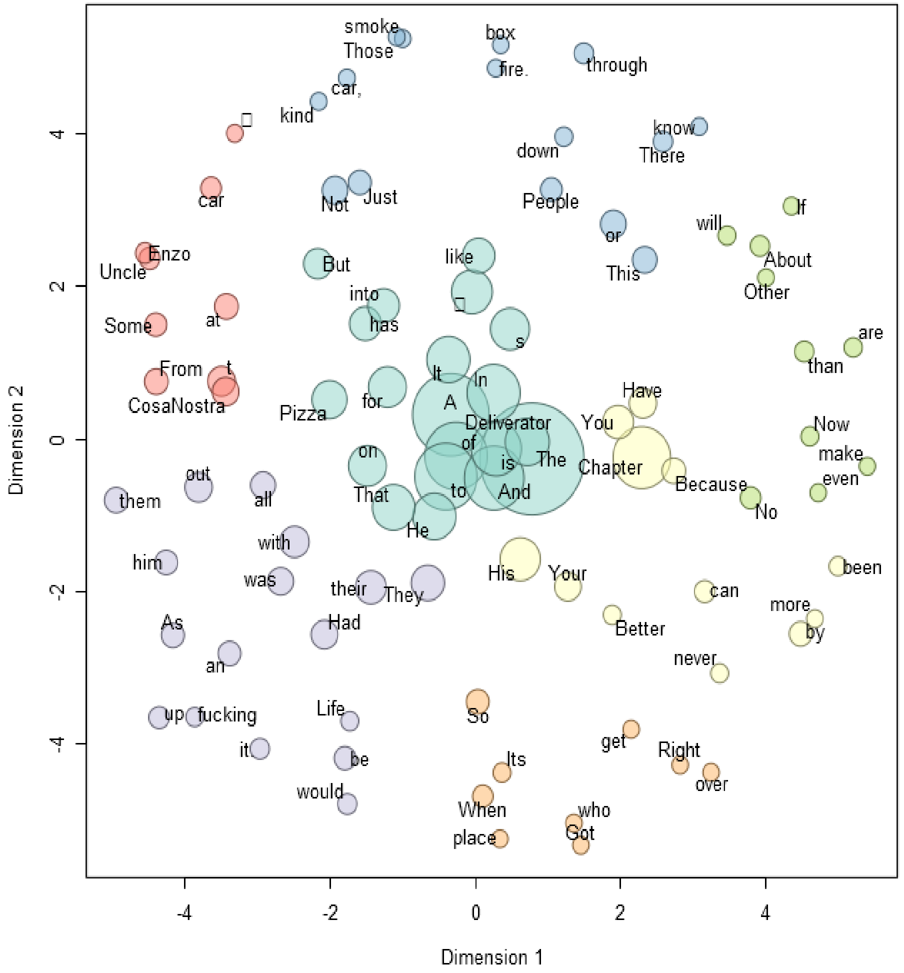


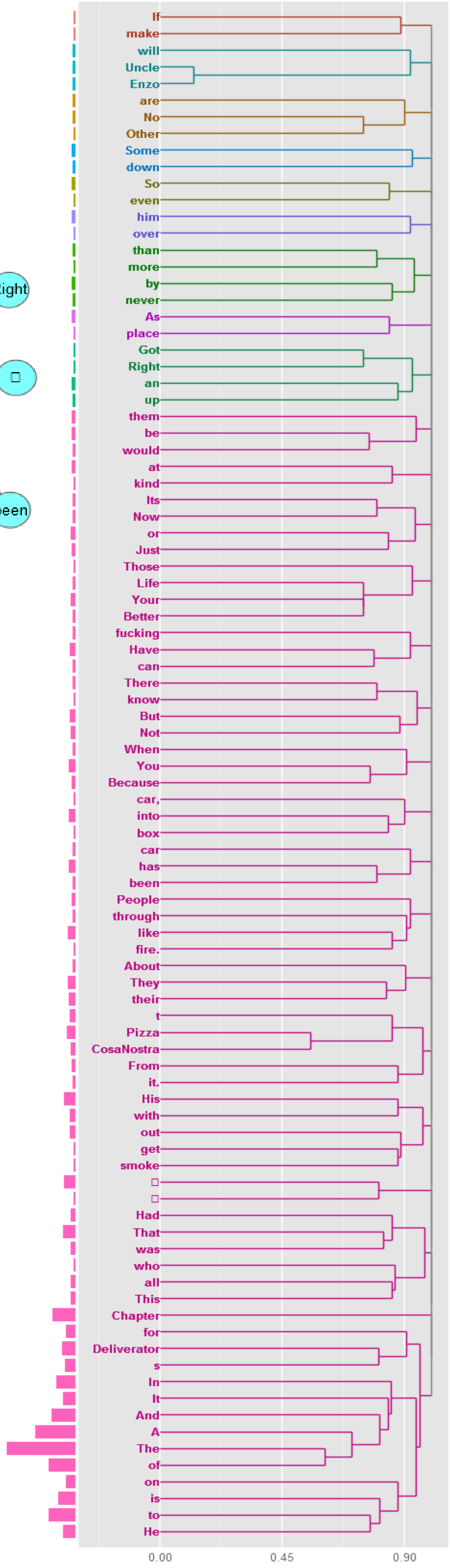


Present: Snow Crash (1992)

By Neal Stephenson

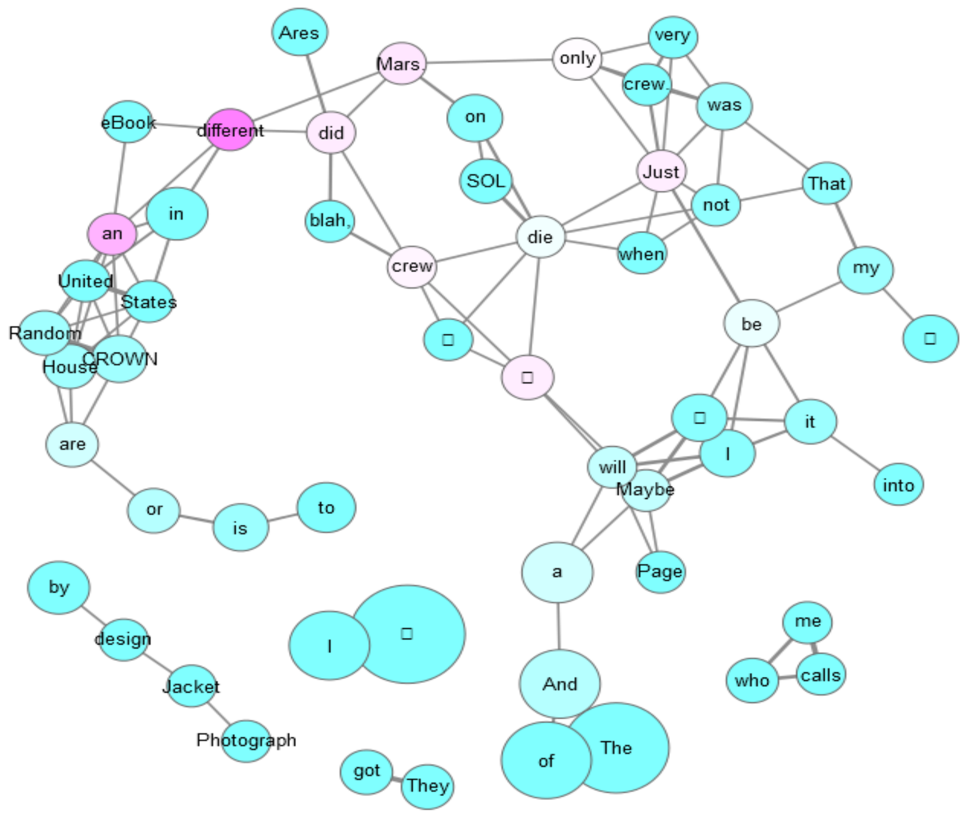


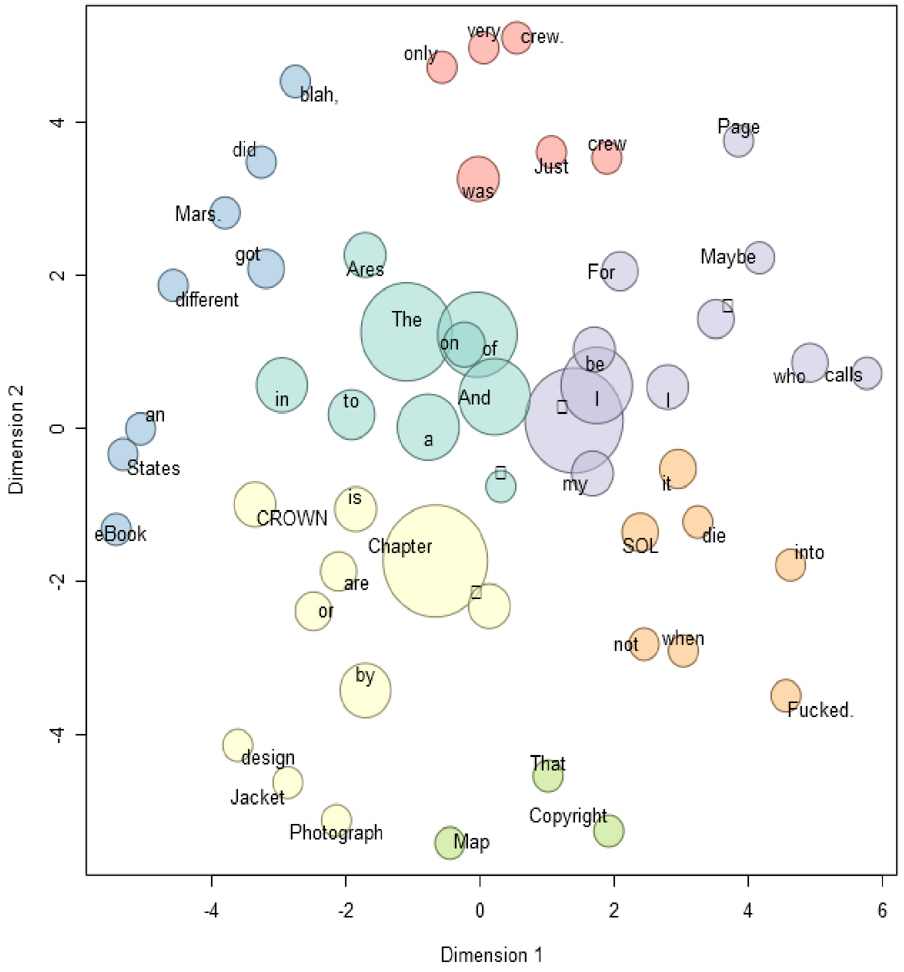


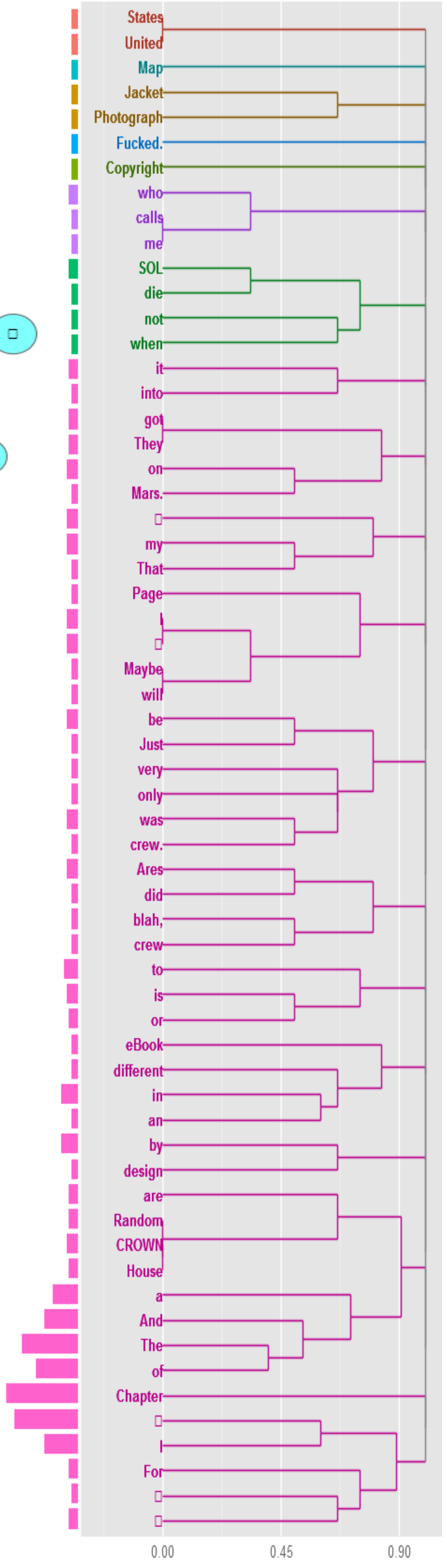


Present: The Martian (2011)

By Andy Weir

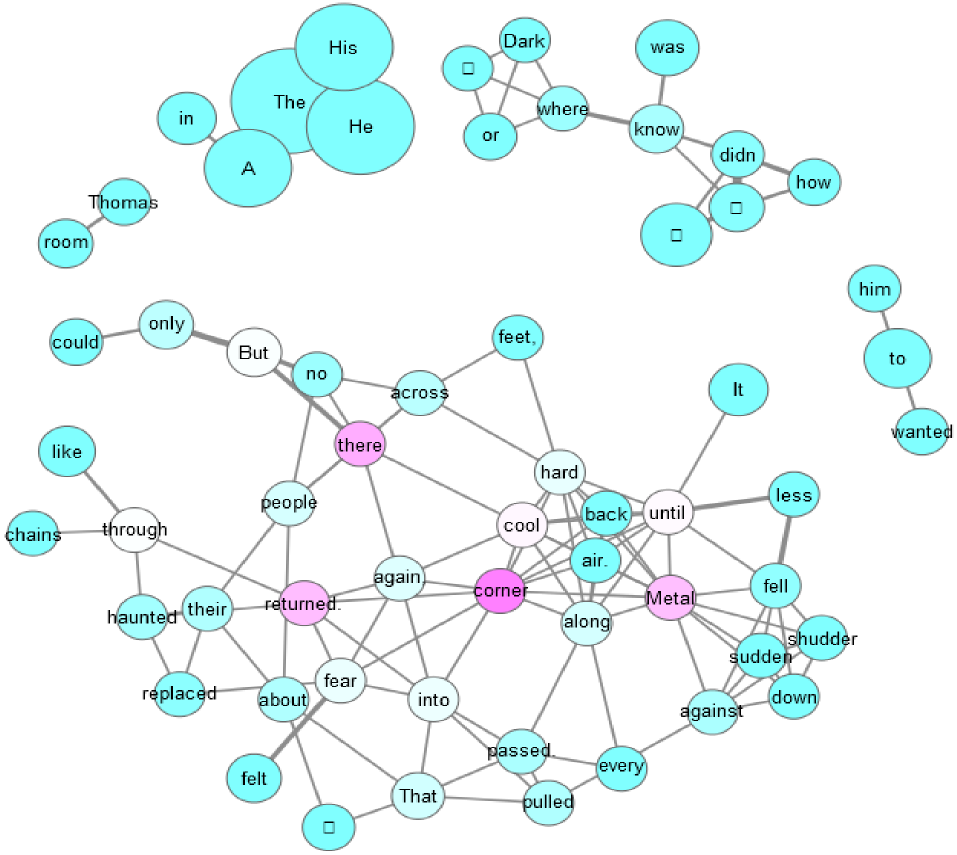


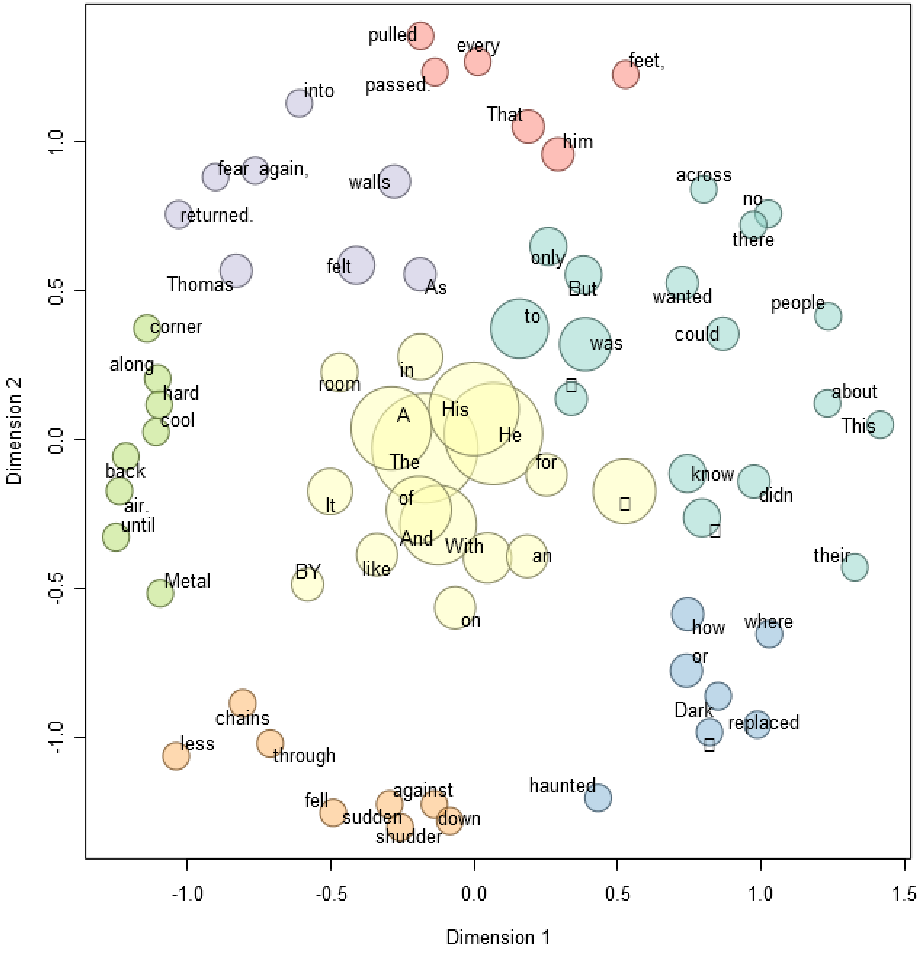


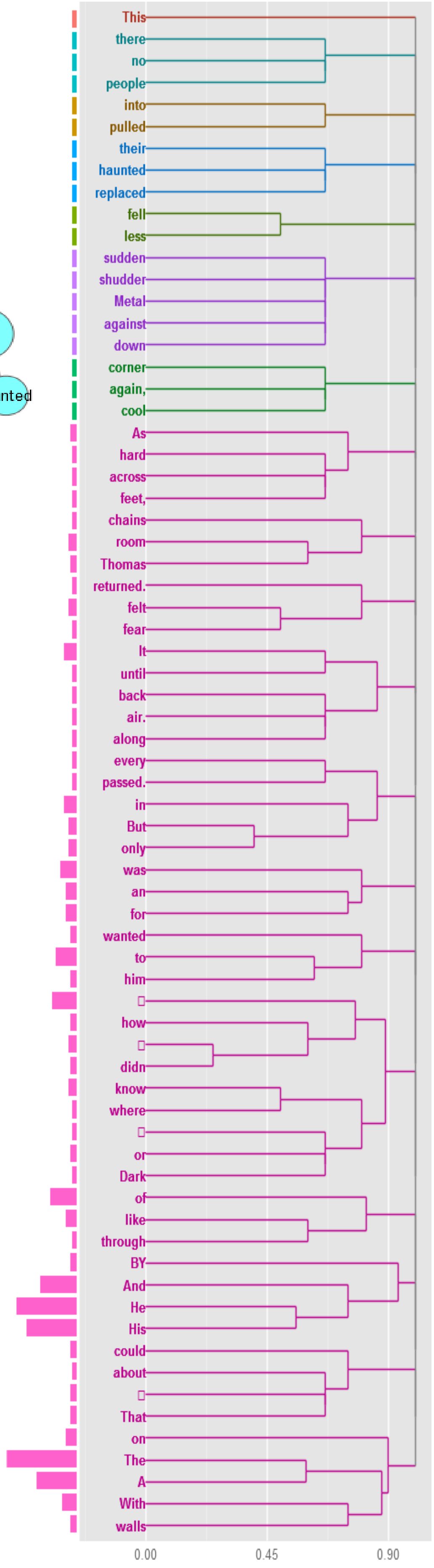


Present: The Maze Runner (2009)

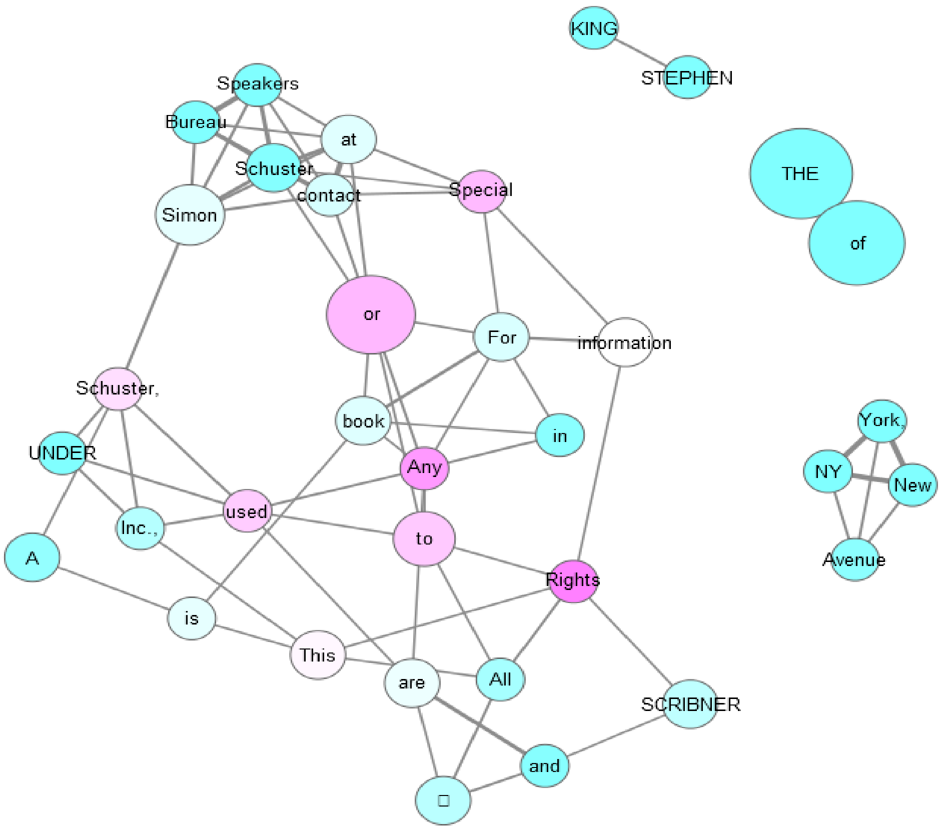
By James Dashner

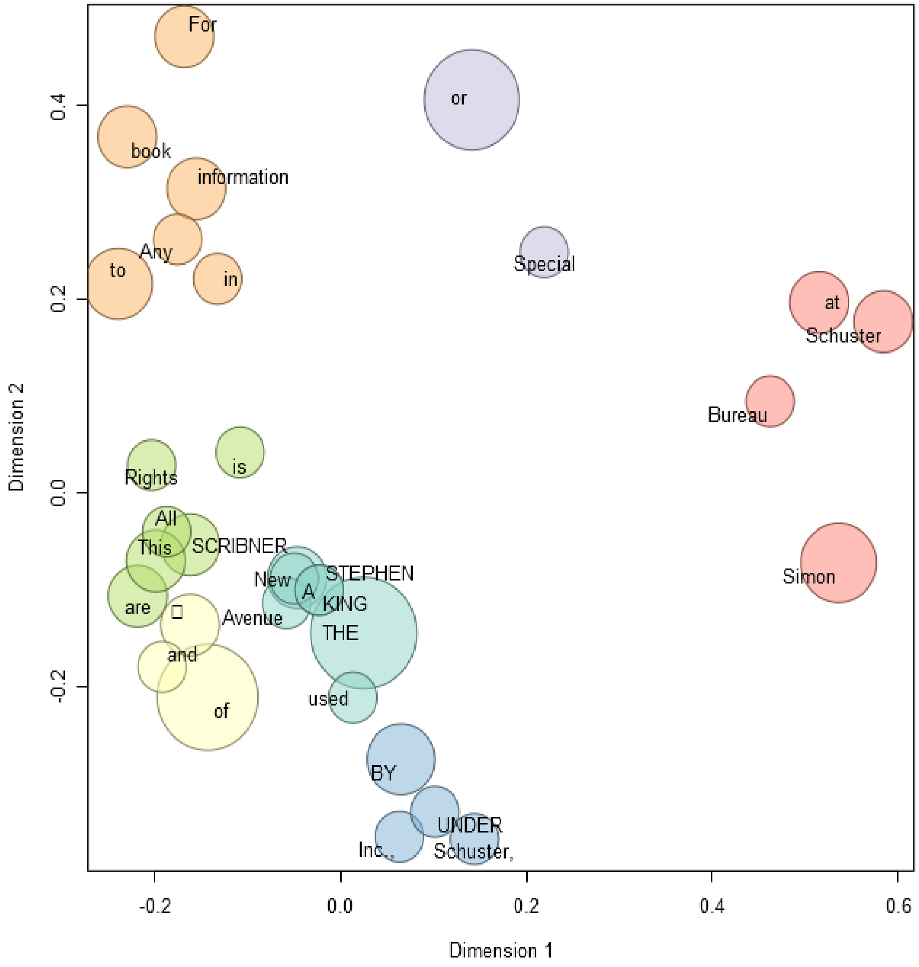


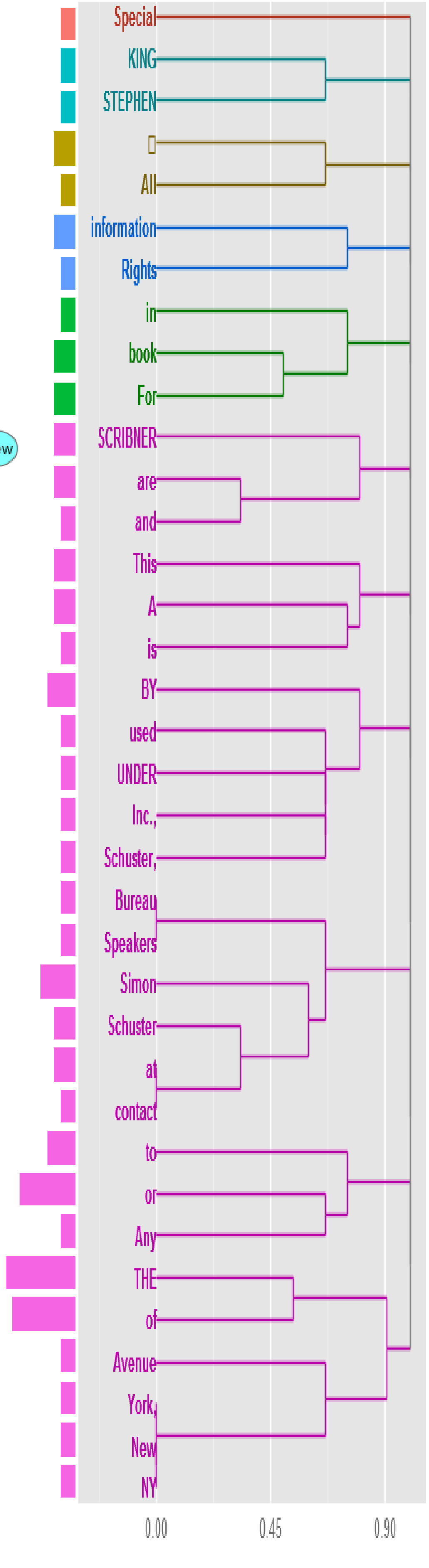




Present: Under the Dome (2009)
By Stephen King







Understanding the Results: The previous pages showed the use of qualitative and quantitative data through three different visual representations. Further information about each visual are detailed below:

- **Hierarchical Cluster Analysis** – Words are grouped together based on their contextual presence with each other. Words in the same cluster generally have similar appearance patterns. Moreover, an entire cluster is differentiated by colors, while the bars on the left side of each word shows the frequency it's used in the book.

- **Multi-Dimensional Scaling** – This scatter plot finds combinations or groups of words that have similar appearance patterns, co-occurrence. Words that are on the zero scale are ones that co-occur within the text, while words with a higher dimensional scale will have weaker appearance patterns. The size of the word bubbles illustrates the word frequency.

- **Co-Occurrence Network** – A network illustration that connects words with its co-occurring words, typically within a sentence. The color of the word circles differentiate the clusters of words, while the circle size shows the word frequency.

Disclaimers (Errors): The execution of this project proved to be challenging. From the lack of coding experience, to the limited online resources worked with, and the poor file compatibility with some of the text analysis software, this project was far from perfect. The following are hurdles to surpass to properly conduct and execute this project:

- Lack of coding experience limited led to using simple pre-exiting text analysis software.
- Preparing and testing several text analysis software.
- The seemingly reliable text analysis softwares tested often had file compatibility problems or did not offer the text analysis tools needed (unless paid for).
- During the analysis, three of the present books was entirely incompatible due to unfamiliar text formatting within the books themselves such as; pictures, unrecognized symbols, or poor text structure.
- As a result of the above, it proved necessary to decrease the sample size of books to fourteen (7 to 7), in an effort to even out the samples tested.
- The past books downloaded from Gutenberg had several instances where the link was imbedded within the text, which negatively impacted the production of the resulting text analysis visuals.
- After the data compilation and the generation of the visual representations, two of the present books – House of Suns & Ready Player One – produced subpar data compared to the rest of the books. The text analysis software did not analyze as many words anticipated (Ready Player One only interpreted five words! C'mon!) making the visual representations of the two books under-saturated.

Observations of the Analysis:

- Due to a poor analysis, Ready Player One and House of Suns are often not reflected in any of the statements made below.
- In terms of hierarchical cluster diagrams, the color diversity of present books is very poor - in four books, purple takes up over half the diagram- while in past books, the color diversity is rich.
- Co-occurrence networks are more spread out in present books.
- The preposition “to” is a high frequency word in ten books (7 past, 3 present).
- The co-occurrences of the pronouns he-his, she-her, are present in eight books (6 past, 2 present).
- According to the multidimensional scale, the word “the” one of the most occurring word in every book. Although, this is normal since it’s one of the most frequently used words in the English language.
- The noun “a” is a high frequency word in all books.
- The visual representations show that the word “I” is one of the most occurring in all seven past books, but only occurs in two present books.
- The hierarchical cluster diagram shows that there’s, less occurrence of the word “was” in present books.
- For all seven past books, the words “I, in, and, the, of” are all in the same hierarchical cluster and grouped together in the multidimensional scale. All five words are within close co-occurrence proximity within each other. Although surprisingly, this is false for all seven present books.

Argument and Conclusion:

The analysis and observations show that there are in fact a number of differences between the writing styles of past versus present science-fiction books. First, the resulting visual representations between past and present are subtly different, as co-occurrence networks model a more scattered connection and hierarchy clusters show less color diversity in present books. Second, since there is evidence that the preposition “to” and the pronouns, “I, he, his, she, her” co-occur more often in past books, it can be assumed that past authors utilize a greater use of pronouns and prepositions. Third, in terms of word frequency, the words “the” and “a” share this commonality among all books, because one often can’t write a paragraph about anything without one of those words. Lastly, One of the most interesting findings was the volume and cluster co-occurrence of the same five words- I, in, and, the, of- in all seven past books, but none within present books. That’s text analysis at its finest!

In conclusion, the argument question “How has the language, syntax, and grammar of science fiction novels changed over time?” was indeed answered, as prior analysis and observations show that the conventionally-written past books has significantly changed over time to the creatively-written present books. These differences are what makes classic past books and ambitious present books unique from each other.