A Spatial Text Analysis of J. R. R. Tolkien's Middle-Earth

Michael H. Beasley
University of South Carolina - Columbia, mhbeasley116@gmail.com

Follow this and additional works at: https://scholarcommons.sc.edu/senior_theses
Part of the Digital Humanities Commons

Recommended Citation
https://scholarcommons.sc.edu/senior_theses/297
A SPATIAL TEXT ANALYSIS OF J. R. R. TOLKIEN’S MIDDLE-EARTH

By

Michael H. Beasley

Submitted in Partial Fulfillment
of the Requirements for
Graduation with Honors from the
South Carolina Honors College

May, 2019

Approved:

Dr. Michael Gavin
Director of Thesis

Dr. W. Matthew J. Simmons, Sr.
Second Reader

Steve Lynn, Dean
For South Carolina Honors College
Table of Contents

Introduction 3
Previous Scholarship 5
Data and Methods 16
Analysis 25
Conclusion 50
Works Cited 52
Appendices 54

Thesis Summary

Using a technique called “distant reading”, which aggregates and analyzes large amounts of data with computational models that can be used to understand literature and find insights or patterns that may otherwise be undetectable, this paper seeks to understand the system of J. R. R. Tolkien’s fictional world of Middle-earth through a statistical approach. In an attempt to provide an entirely different view on his work, this paper examines semantic aspects of Tolkien’s corpus and how they complement the narrative structure, with particular emphasis on the topographic landscape and the journeys of specific characters through the space of Middle-earth itself.
Introduction

J. R. R. Tolkien created the entire world of Middle-earth, populating it with people, languages, landmarks, and history. It has inspired generations of writers, artists, and musicians with its richness and depth, and nearly every aspect of his masterful corpus has been analyzed for decades. Less work, however, has been done on analyzing the narrative and geography of Tolkien’s work with a computational approach. This paper does exactly that: using statistical models created in the programming language R, both the geospatial framework of Middle-earth and the narrative structure of *The Lord of the Rings* are analyzed, with emphasis on how they complement one another, demonstrating that meaningful semantic analyses can be performed on fictional worlds. The vocabulary that Tolkien uses in his work is shown to correlate with specific locations in Middle-earth, and thematic connections can be drawn between words that coincide with characters and the journeys undertaken by those characters, particularly with regards to where they are spatially in Middle-earth at a given time. This offers a new depth to Tolkien’s work, and offers some fascinating insight and interpretation of how words are used spatially within it.

In order to analyze a fictional world like Middle-earth in such a way that statistical models would be relevant and useful, it is necessary to employ a technique called “distant reading”, a term coined by Franco Moretti. Modern application of this technique aggregates and analyzes large amounts of data, based on the assumption that computational models can be used to understand literature and find patterns that may otherwise be undetectable. In Moretti’s own words: “‘Distant reading’, I have once called this type of approach; where distance is however not an obstacle, but a specific form of knowledge; fewer elements, hence a sharper sense of their
overall interconnection” (Moretti 2005, 1). Discussing the narrative structure and overarching themes of fiction in a literary sense is one thing, but understanding the system of a fictional world as a whole, through the lens of computational models, provides an entirely different view on a work in which the aspects of the geographical landscape and how it complements the narrative structure may otherwise be lost in analysis. Distant reading allows a broader look at the connections between elements of a work – words, places, concepts – to better define those connections and examine how they influence one another.

Applying this to Tolkien is the purpose of this paper. Simply reading through The Hobbit and The Lord of the Rings makes several elements of the narrative extremely apparent – the characters, their journeys, the overarching themes, and other aspects of the world itself. Distant reading as an approach, however, is able to show mathematically which of these elements are interconnected, relying on the semantics used by Tolkien himself rather than external interpretation. Where characters are in Middle-earth, how those locations are described, and the impact that certain places have on the narrative are all more easily analyzed with semantic data, and it shows that Tolkien was able to construct a fictional world that has meaningful connections between its individual locations and other elements of the narrative.
Previous Scholarship

There are many discussions on various aspects to Tolkien’s life and work; many focus on him as an academic, the process of his writing, the construction of *The Silmarillion*, and the creation of his mythology. Others write of themes in his novels, analyze his characters, and (despite Tolkien’s insistence to the contrary) examine the work as allegory. There are many writings that focus on his invented languages and writing systems, and the real-life languages, cultures, and mythologies that inspired Middle-earth. His literary contributions and legacy are frequent topics, as well, in addition to specific themes that are found in his work (such as good and evil, war, music and beauty, and environmentalism). Among the plethora of writings on Tolkien, there are some topics in particular that will be important to this paper going forward; these all focus on major themes of his work, its narrative structure, and how language and writing style add to its depth.

Perhaps most important to this paper is discussion on the concept of the Hero’s Journey, the geography of Middle-earth, and how Tolkien uses all of these together to enhance his narrative. Scholars write about the interlaced narrative of *The Lord of the Rings* and how its characters’ journeys are separate, yet connected. The scholarly consensus is that the novel forms a variety of interconnected quest narratives, each of which explore different themes and follow different characters through different locations in Middle-earth. This is all crucial to understanding the spatial aspects of language that Tolkien uses, and how specific locations are tied to both overarching themes of the novel and important moments of development for the characters.

John R. Holmes begins “The Lord of the Rings” with a brief history and synopsis of the eponymous novel before discussing its narrative structure. He describes it as a quest narrative
with the appropriate elements, and notes Tolkien’s practice of establishing safe havens for the characters between times of peril. Because these havens are spatial locations, “… Tolkien’s imagination is not just mythopoeic, but also geographic” (Holmes 2014, 136). *The Lord of the Rings* has an interlace narrative structure, where characters are separated and their individual tales told separately, before reuniting near the end of the story. “Tolkien does not leave the reader helpless in lacing the plot strands: he offers key narrative ‘landmarks’ to help readers ‘synchronize’ events in separate chapters, even when we read them out of sequence” (Holmes 2014, 137). Though he denied *The Lord of the Rings* to be an allegory, Tolkien did consider the three major concerns of his work to be “Fall, Mortality, and The Machine”; a common theme throughout the three books is the world and all the good things therein dwindling over time. Holmes also discusses the role that Tolkien’s linguistics play in the novels, especially in the fictional languages that he invented; the names that Tolkien gave to objects, people, and locations are all “phonaesthetically” satisfying, having an uncanny fitness to them.

Discussion on the importance of languages and nomenclature in Tolkien’s work is continued in Arden R. Smith’s “Invented Languages and Writing Systems”. He examines the aesthetic and artistic appeal of creating languages, Tolkien’s method of constructing them, and how the development of *Quenya* led to the creation of the entire Elvish mythology. Tolkien himself describes *The Lord of the Rings* as “fundamentally linguistic in inspiration,” claiming the novel to have been written around the languages he made; for him, no matter the subject, the name preceded the story. Smith presents an overview of Tolkien’s various constructed languages, their origins, their writing styles, and where they appear in the novel.
Aside from the linguistics used by Tolkien, the very structure of *The Lord of the Rings*’ narrative deserves special attention. In fact, as Diana Wynne Jones writes in “The Shape of the Narrative in *The Lord of the Rings*”, “Tolkien was far more concerned with the matter of his narrative than the manner, and he only exercised his undoubted gift for language in inventing names, particularly names of places” (Wynne Jones 1983, 89). The novel is organized in a series of “movements”, each one with an extension (referred to as a “coda”) that reflects on the previous movement, as well as the movement to come. Understanding these movements (a term that even in itself implies there to be a spatial aspect to the narrative) is key to understanding the narrative itself, and some time must be devoted to smaller details that have significant meaning.

The first movement is described as ending with the flight from Bree, its coda being the journey to and incident at Weathertop. The second movement, a more peaceful one, concerns the debate at Rivendell, where the narrative changes – Frodo has achieved what he initially set out to do, and the debate extends the journey with the movement’s coda of the Fellowship setting out. The third movement, in two parts, covers the journey through Moria and the sojourn in Lothlórien; it ends with the coda of Frodo donning the Ring and the Breaking of the Fellowship. Following this, the narrative divides into two concurrent stories, both of equal importance. The positive section is told first – of the Ents, Rohan, and the victory over Isengard. The negative section is told second – of the Dead Marshes, Ithilien, and the capture of Frodo after the defeat of Shelob. Though the tremendous Battle of the Pelennor Fields, in which Rohan and Aragorn both come to the aid of Minas Tirith, ends in victory, Sauron and the power of Mordor are still at large, and thus the large movement has a coda, in which the decision is made to march on the Black Gate to serve as a distraction to give Frodo a chance to destroy the Ring. This leads to the
final movement, where the quest fails save for actions by both hobbits in the past – sparing Gollum’s life and turning him against the pair – leading to an accidental success. The coda to this movement takes up nearly half of the sixth and final volume, ending the narrative with Frodo departing into the Sea, which was signaled from the very beginning of the narrative.

The division of the fourth movement into two concurrent stories has been briefly mentioned already – Richard West writes of this literary technique in “The Interlace Structure of The Lord of the Rings”. West uses earlier literature to explain the novel’s form, which he describes as a technique similar to medieval interlace. This contrasts with the more modern style of organic unity, which uses clear and simple patterns to create a manageable narrative with a single major theme. Interlace, on the other hand, provides multiple concurrent narrative lines, dividing the reader’s attention and mirroring the feel of a chaotic reality in a real world. West echoes Wynne Jones’ concept of the coda when describing its effectiveness; “… the apparently casual form of the interlace is deceptive; it actually has a very subtle kind of cohesion. No part of the narrative can be removed without damage to the whole, for within any given section there are echoes of previous parts and anticipations of later ones” (West 1975, 79). The chain of events that make up the narrative are even outlined in Tolkien’s initial title suggestions for the six books that comprise The Lord of the Rings: The First Journey, The Journey of the Nine Companions, The Treason of Isengard, The Journey of the Ring-bearers, The War of the Ring, and The End of the Third Age. Each of these books are centered around events that would not be possible without preceding events. These mixed narratives ebb and flow, with the novel’s themes and motifs interwoven throughout and between them.
The narrative structure and story of *The Lord of the Rings* has long been the subject of debate in regards to being a quest narrative. Traditionally, in this sort of narrative, a hero sets out on a journey to acquire an object; however, as Anna Caughey writes in “The Hero’s Journey”, this is a rather narrow definition of a “quest”. The purpose of Frodo’s undertaking is to destroy (rather than acquire) an important object in the Ring, though the responsibility that he takes upon himself at the Council of Elrond supports the notion of him undergoing what is referred to as the Hero’s Journey. Even setting Frodo aside, the other figure on a traditional quest in the novel is Aragorn, who seeks to reclaim the throne of Gondor – there is no doubt that *The Lord of the Rings* is a quest narrative. Again, the importance of having multiple narratives is clear; “[t]he text works successfully in both the adventure-story and elegaic modes because *The Lord of the Rings* offers the reader not one quest-narrative or Hero’s Journey but several, which run simultaneously in a number of registers and at a number of levels” (Caughey 2014, 404).

The Hero’s Journey is broadly defined by Caughey as a three-stage pattern: separation, initiation, and return. She also uses Northrop Frye’s three stages that are found within *The Lord of the Rings*:

1. The first stage is identified as the *agon* or conflict, the stage of “preliminary minor adventures” (Frye 1957, 187), roughly corresponding in Tolkien’s work to the events of Books One and Two (*FR*).
2. The second is the *pathos*, “the crucial struggle, usually some kind of battle in which either the hero or his foe, or both, must die” (Frye 1957, 187), most obviously apparent in Frodo’s questionably successful struggle on Mount Doom … but also present in other variations within the text.
3. The third and final stage is “the *anagnorisis* or discovery, the recognition of the hero” (Frye 1957, 187), which is perhaps the least conventional aspect of Tolkien’s text, as different levels of recognition – or lack of recognition – are meted out to different characters in the final chapters of Book Six… (Caughey 2014, 405)
Much of the structure and style of the novel is inspired by medieval literature, with two techniques in particular: doubling and entrelacement. Doubling not only echoes previous events (such as Faramir’s success in resisting the Ring after his brother Boromir fails), but also gives a certain depth to otherwise flat characters; “while conceivably a rather flat hero when read alone, [Frodo] becomes far more convincing when read in conjunction with Gollum as his double, not merely an antagonist but an indication of Frodo’s own potential to be enslaved by the Ring” (Caughey 2014, 408). Entrelacement (or interlacement) has been discussed already, though an important aspect to which this paper is building is that of seeing the narrative device through a geographical lens; not only do the members of the broken Fellowship have their own stories that happen simultaneously, they all occur in different locations, and this has an effect on things. Interlacement “allows for connections between key plot events to be made by the reader on a second journey through the text and contributes to the variation in reading experience from first to subsequent readings” (Caughey 2014, 407).

Caughey argues that there are three distinct Hero’s Journeys in The Lord of the Rings, and involve two types of heroes: the high mimetic and the low mimetic (as identified by Northrop Frye). On a spectrum, Aragorn is easily identified as a high mimetic hero – the leader, the mythical figure with authority and powers greater than those of normal people – whereas Frodo is identified as a low mimetic hero – a relatable Hobbit with a common sense of humanity. When it comes to the Hero’s Journeys, however, “the plot is complicated by Tolkien’s decision to give the truly epic quest to the more humble quester, while the more personal story of growth and maturity is given to the high mimetic hero” (Caughey 2014, 408).
The first journey shares its narrative form with many typical fairy-tales, which Caughey describes as follows:

… the young hero sets out on a series of adventures which turn out to be much more challenging than expected. He faces obstacles, which are duly overcome by his own skill or with the assistance of various companions, and returns home, laden with treasure, to a life of significantly greater comfort and prestige (Caughey 2014, 409).

Described by Caughey as a “There and Back Again” quest, this fits squarely with the journey of Bilbo Baggins in *The Hobbit*; he leaves the comfort of the Shire and undergoes several dangerous situations, which he manages to overcome – at first with the aid of Gandalf and the Dwarves, then gradually by using his own abilities – and returns to the Shire for a happy ending, having grown wiser and more mature about the world.

The other example of this first quest form is one that bears significant resemblance to many medieval romances – that of Merry and Pippin. Caughey describes them as untried heroes, searching for a “knightly identity”; fittingly, both Hobbits are eventually knighted near the end of their quests. The turning point comes in Fangorn Forest, where they begin to carry themselves rather than being carried – it is even an instance of doubling, when compared to previous failures in the Old Forest. Merry and Pippin return home to the Shire, as Bilbo did, largely unchanged from how they left it.

In stark contrast lies the second journey, that of “The Broken Hero”; in this case, Frodo. The hero may fail in their quest, and is forever altered after their return. Caughey’s reading of the scene of *pathos* for Frodo, standing in the Crack of Doom and wrestling with temptation, illustrates this:

This … reading suggests that the task simply demanded that Frodo use the ugliest and most selfish part of his nature, as represented by Gollum, in addition to the noble and
sacrificial elements represented by Sam, and that with the destruction of Gollum, his personality will never be wholly reintegrated again (Caughey 2014, 411).

The Broken Hero quest does not result in perfect success, but with the discovery of the hero’s own fallibility through the testing of their virtues. Frodo is so wounded by his journey and trials, both physically and emotionally, that he departs Middle-earth forever.

The third and final form of the Hero’s Journey is referred to by Caughey as “The Patriarch’s Quest”, undertaken by Aragorn and Sam:

The third, and arguably most transformative, form of Hero’s Journey in *The Lord of the Rings* is that of the nascent patriarch – the inexperienced young man who emerges from his adventure not merely glorified, but transformed; who moves forward into a state of social and emotional maturity that fits him for the role of husband and father (Caughey 2014, 412).

Aragorn assumes leadership after the apparent death of Gandalf in the Mines of Moria, though he hesitates at times, even admitting failure after the Breaking of the Fellowship. Through the forming of the Three Hunters, however, he is able to display his abilities to unite people of Middle-earth with little in common, in pursuit of the ultimate goal of protecting those who cannot protect themselves, as a leader, a father, and a king. In contrast, the low mimetic figure of Sam (referred to by Tolkien himself in his letters as the “chief hero” of the quest) begins his journey as a humble gardener, growing in maturity as he accompanies Frodo to the point where he is the only character able not only to temporarily carry the Ring in Frodo’s absence, but also return it. He ends his journey as the Mayor of the Shire, effectively a ruler like Aragorn, protecting his people and becoming a father. Caughey concludes by saying that “the multiplicity of heroic journeys, characters, and modes that [entrelacement] makes available provide some of the greatest pleasures that the text has to offer” (Caughey 2014, 415).
Charles A. Huttar, in “Hell and the City: Tolkien and the Traditions of Western Literature”, discusses another aspect to the Hero’s Journey. In classical mythology, there is a three-tiered nature to many myths – the underworld, earth, and heaven – and Tolkien utilizes this symbology in *The Lord of the Rings*, especially the image of descending into the underworld, or hell, and the image of the city. Huttar includes his own interpretation and summary of Frodo’s journey with these things in mind:

The hero’s adventure has three stages, departure, initiation, and return; the first two compromise his journey to hell. He is singled out to be a hero, called to undertake the quest, given a choice to accept or refuse which is at the same time a test of whether or not he really qualifies. He then undergoes a series of further trials, wins through by supernatural aid, enters into the final death-like struggle, and eventually emerges victorious (Huttar 1975, 119).

Huttar follows Frodo’s journey from the womb-like Bag End, through images of hell in the Barrow-Downs, to a brief respite and midpoint in Rivendell, and through the hellish descent into Moria. The darkness and struggle of Moria serve as an experience in which Frodo, as well as the other hobbits, learn valuable lessons – they learn courage, pity, and how to carry on through sorrow in the absence of hope after Gandalf’s fall. Frodo himself emerges with a renewed sense of responsibility. Gandalf’s battle with Durin’s Bane reflects a kind of ascent, returning to life and being rescued by an eagle. Another midpoint of Middle-earth, Lothlórien, follows the journey through Moria before the second journey to “hell” for Frodo: through the Dead Marshes and into Mordor itself. This portion of his adventure is notably different; “… this part of Frodo’s journey to hell is not a matter of landscape alone but of the ring which weighs more and more heavily on him and, still worse, of the growing power of the Eye” (Huttar 1975, 128).

The image of the city takes a different approach that focuses on the morals of *The Lord of the Rings*. The verticality of the cities in Middle-earth, exemplified in towers such as Minas
Tirith, Barad-dûr, and Orthanc, demonstrate the celestial dimension of good as well as the infernal dimension of evil. This also applies to characters, the vast majority of which are neither good nor evil; rather, they possess the potential for both, and are constantly making decisions between them. In Huttar’s words:

… Tolkien’s moral vision is reinforced by his treatment of the inanimate world that furnishes man with the opportunity and challenge to be a subcreator. Objects are not in themselves good or evil, but nearly everything has a potential for both. Mithril is both the greatest of treasures and a deadly bane. Cities are built on an axis that reaches to hell as well as heaven… The dual potential included in the capacity of “making” embraces poetry and jewelry as well as instruments of war and torture… The mastery of fire produces bombs, but also comfortable hearths and homes (Huttar 1975, 139).

This notion of a spatial aspect to morality is further explored by Walter Scheps in “The Fairy-tale Morality of The Lord of the Rings”. Scheps describes the relation between physical and moral direction in Tolkien’s works, with the North and West generally being associated with good, and the South and East with evil; the Grey Havens and the Shire are in northwestern Middle-earth, Mirkwood and the Desolation of Smaug are in the northeast, Mordor is in the southeast, and Gondor is in the southwest. Even the Shire, divided into quadrants, can be seen as a microcosm of Middle-earth with the spread of fear, the appearance of the Black Riders, and the entry of Saruman all from the south and east, with characters like Tom Bombadil and Aragorn introduced with mentions of the west.

Summarizing the relationships of spatial representation, characters’ journeys, and linguistics, Robert Tally writes in Topophrenia of Tolkien’s approach to constructing the geography of The Lord of the Rings:

… [T]he plot of Tolkien’s novel is itself cartographic, even if one were to set aside the helpful maps included at the beginning of each volume. The Lord of the Rings represents a vast space: the imaginary realms of Middle-earth, which features not only a diverse topography and geography … but also a rich, dense historical background that gives these
places a kind of overdetermined significance or surfeit meaning. Each place is not only marked as if on the map but also described, contextualized, interpreted, and woven into a larger geopolitical discourse. Many of the places encountered bear three or four different names, for example, as their geographical or spatial situation is couched with a long history … The more or less linear itineraries of the novel’s protagonists … are set in relation to a much larger geography and history that make their own personal encounters and discoveries all the more meaningful. The “map” of Middle-earth is not merely the sketched drawing included at the front of the book … but is primarily formed through the narrative itself (Talley 2018).

Again, the scholarly consensus on *The Lord of the Rings* is that it forms several interconnected quest narratives that are each tied to a series of spatial locations, most of which serve either as a place of respite or as a tribulation through which the characters must undergo. Regardless of which category any given location falls under, they all present opportunities to discover things about the characters and the world of Middle-earth. This paper aims to take these concepts of how various aspects of the narrative structure are spatially represented and take certain computational approaches to show it.
Data and Methods

Aelib, an open source digital library associated with GitHub, has HTML files of *The Hobbit* and *The Lord of the Rings*; these digital copies of the four novels were read into R, the programming language used to write the scripts that performed analysis and statistical computing. The first step was parsing out the vocabulary – every word that Tolkien uses in the corpus – and creating a simple two-column matrix that demonstrated the number of appearances each word makes. This was expanded by a script that looped through each file, divided the novel into its chapters, and counted each word as it appeared by chapter; this resulted in a larger matrix with each row representing a word, each column representing a chapter in Tolkien’s work, and the intersection points being how many times each word appeared in that chapter.

The total number of columns was eighty-six; this began with one “chapter” of notes at the beginning of *The Hobbit*, followed by its nineteen chapters. Five “chapters” of prologue began *The Lord of the Rings*, with its three volumes comprising the remaining sixty-one chapters. The total number of types, or unique words, was 13,888, though a very small number (less than one percent) were aspects of the HTML files themselves, and were not actual words. The number of tokens (occurrences of any given word), or the total word count calculated, was 578,432, though this includes the notes to *The Hobbit* and the prologue to *The Lord of the Rings*.

Part of the preliminary research done before the scripts were ever written involved compiling a gazetteer, or a geographical index of locations within a given area. In this case, the gazetteer took the form of a list of all the named places in Middle-earth. This list included several categories of locations: the names of larger regions such as Mordor and Gondor, smaller sub-regions such as Southfarthing and the Westfold, city names such as Minas Tirith and
Hobbiton, individual landmarks such as Weathertop and Orthanc, smaller places of note such as the Prancing Pony and Helm’s Gate, locations that no longer exist such as Beleriand and Númenor, geographic features such as the river Anduin and the Misty Mountains, and even references to more abstract concepts such as the Sea and the West. The total number of unique, individually named locations throughout *The Hobbit* and *The Lord of the Rings* was 772.

The gazetteer was used as the basis for a similar set of loops that, rather than counting each word, focused solely on the named locations from the gazetteer. This resulted in a second matrix similar in construct to the first, though each row represented a place in Middle-earth rather than a given word. The columns still represented chapters, and the intersection points displayed how many times each location was named in a given chapter.

These two matrices were used as the basis for the bulk of the analytics, in tandem with a set of geographical coordinates that were taken from a printed map of Middle-earth; simply using a ruler to mark out two numerical axes gave each location a relative latitude and longitude.
Figure 1: The map of Middle-earth used to generate relative latitude and longitude values for each specifically named location in the corpus.

The scale was roughly 1.50cm = 100 miles. Using the centimeter scale as units, the map provided a way to quantify the relative location of each place in Middle-earth; for example, the city of Minas Tirith had a latitude of 6.1 units and a longitude of 17.3 units, while Rivendell had a latitude of 15.3 units and a longitude of 13.0 units. The larger the latitude, the further north the place was located; locations with larger longitudes were located further east. For regions such as the Shire (14.6, 7.4) and Gondor (5.5, 13.5), values were assigned based on the relative center of their areas; for some areas such as Harad (0.8, 22.5), however, this was not possible, and the center of where their names appeared on the map was used instead. The sheer scale of the map
used also meant that the more specific the location, the less accurate relative locations would be; in other words, several individual landmarks such as the Houses of Healing and the White Tower of Ecthelion were both located in Minas Tirith, so all three of these locations had the same latitudes and longitudes. Places with different names, such as Rivendell and Imladris, were also assigned the same latitudes and longitudes.

Once the baseline of vocabulary, locations, and coordinates were set, the next step was creating functions to perform relevant analytical calculations, the first of which was measuring pointwise mutual information, or PMI. This is a measure of association between two variables, or how likely it is that two things (in this case, words) coincide, using their individual probabilities and joint distribution.

\[
\text{pmi}(x; y) = \log \frac{p(x, y)}{p(x) p(y)}
\]

In the PMI formula, shown above, \(p(x)\) represents the probability that any word, chosen randomly, will be of type \(x\); \(p(y)\) represents the probability that any word, chosen randomly, appears in chapter \(y\). The joint probability distribution is given as \(p(x)\) times \(p(y)\), and \(p(x, y)\) represents the actual value observed in the relevant matrix. In total, the PMI of any given word shows whether it is more or less frequent than expected. PMI works well with regards to measuring semantic similarity in a word-context matrix, performing better than a wide variety of alternative weighting methods (Turney & Pantel, 2010, 157).

Another computation was that of cosine similarity, a measure of similarity between vectors. When used in analyzing text, the dimensional value of vectors representing words correspond to the number of times each word appears in the text; cosine similarity can be used to measure how similar two texts are with regards to their content.
\[
\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}
\]

In the similarity formula, shown above, \(A\) and \(B\) are two non-zero vectors, with \(A_i\) and \(B_i\) being components of said vectors, respectively. The numerator is the sum of the products of the components of each vector, while the denominator is the product of the square roots of each vector component squared and summed together. The formula utilizes the Euclidean dot product and results in a value between negative one and positive one; a result of negative one indicates that the two words are exactly opposite to one another, with positive one indicating the two words are one and the same.

In order to avoid words seeming to be highly similar simply because they appear frequently, the cosine similarity formula divides by the respective length of the vectors and provides a length normalization to obtain the cosine of the angle between them (Clark 2015, 504). Measuring cosine similarity for a given location over the location matrix gives other named locations that coincide most frequently with the specified place. This becomes clearer in practice; for example, measuring the cosine similarity of Bree over the location matrix gives the following results:

<table>
<thead>
<tr>
<th>Location</th>
<th>Similarity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bree</td>
<td>1.0000000</td>
</tr>
<tr>
<td>Bree-land</td>
<td>0.9147455</td>
</tr>
<tr>
<td>Staddle</td>
<td>0.8644356</td>
</tr>
<tr>
<td>Archet</td>
<td>0.8424889</td>
</tr>
<tr>
<td>Greenway</td>
<td>0.8275395</td>
</tr>
</tbody>
</table>
When measuring a given word over the overall word matrix (as opposed to only focusing on locations), broader comparisons can be made, and the calculation results in the words that are the most similar to the given word. For example, measuring the cosine similarity of the word “tree” produces the following results:

<table>
<thead>
<tr>
<th>Word</th>
<th>Similarity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>tree</td>
<td>1.0000000</td>
</tr>
<tr>
<td>branches</td>
<td>0.8613180</td>
</tr>
<tr>
<td>trunk</td>
<td>0.8374366</td>
</tr>
<tr>
<td>trees</td>
<td>0.8003490</td>
</tr>
<tr>
<td>rustle</td>
<td>0.7612032</td>
</tr>
<tr>
<td>blending</td>
<td>0.7446865</td>
</tr>
<tr>
<td>boughs</td>
<td>0.7124064</td>
</tr>
</tbody>
</table>

It is clear that the results of measuring the cosine similarity of a given location will produce other locations that are somehow related, be it geographically or thematically. In a similar manner, measuring a word over the word matrix produces words that are closely related to one another. Given, however, that locations are also words, cosine similarity can be used to measure locations over the word matrix, as well. Extending the previous example, Bree (taken from the location matrix), when measured over the word matrix rather than the location matrix, produces the following:

<table>
<thead>
<tr>
<th>Word</th>
<th>Similarity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>bree</td>
<td>0.9994855</td>
</tr>
<tr>
<td>greenway</td>
<td>0.8928279</td>
</tr>
<tr>
<td>word</td>
<td>score</td>
</tr>
<tr>
<td>----------</td>
<td>---------</td>
</tr>
<tr>
<td>butterbur</td>
<td>0.8754266</td>
</tr>
<tr>
<td>landlord</td>
<td>0.8591487</td>
</tr>
<tr>
<td>staddle</td>
<td>0.8585841</td>
</tr>
<tr>
<td>landers</td>
<td>0.8576275</td>
</tr>
<tr>
<td>gatekeeper</td>
<td>0.8291039</td>
</tr>
</tbody>
</table>

This essentially gives a list of keywords for any given place. The implications of this measurement deserve further discussion, and more comparisons of this nature will be explored in the dedicated Analysis section of the paper.

In order to make the geographical coordinates useful, a third computation was made, this time to find geographical centroids. This process multiplies the word matrix by the location matrix in order to give each word (converted to a vector) a geographic footprint derived from the latitude and longitude of each location with which it coincided. The end result allowed this data to be mapped onto a coordinate plane, with each word’s location on the graph being affected by where it was used most often in the text.

Another way to geographically visualize Tolkien’s vocabulary is to graph each word’s longitude from the geographic centroid alongside its radius, calculated using the Euclidean distance of the vectors. This measures how dispersed the words are going from west to east in Middle-earth.
In the graph above, words like “denethor” and “bree” have very low dispersion, meaning that the places in which these words are referenced – or the places with which they coincide – are all extremely close together; Denethor is mentioned most often in reference to Minas Tirith, and Bree is a town that is relatively self-contained in the narrative. On the other hand, words like “éowyn” and “strider” have very high dispersion, meaning that they coincide with a wide variety of locations across Middle-earth. Strider in particular is a good example of this; Aragorn is introduced into the narrative in Bree under the name “Strider”, and though his true name is eventually revealed, he travels all the way from Bree to Rivendell being referred to frequently as
“Strider”. Even when Merry and Pippin are separated from him and travel to Fangorn, they refer to him as “Strider”, and Aragorn himself uses the name when meeting the Riders of Rohan. When Sam awakens in Ithilien after the Ring is destroyed, the name “Strider” is used – the word appears all across Middle-earth, giving it a relatively high dispersion.

This is simply a surface level analysis of the scripts; understanding the methods used to gather the data is necessary to support further synthesis of the information. Individual comparisons that can be made, and how the data can be extrapolated to relate to the narrative, will be explored further in the dedicated Analysis section.
Analysis

The geographic centroids of words mapped onto a coordinate plane are displayed below in Figure 3; this analysis will be the starting point to examine the world in which characters are moving in a geosemantic sense. The paths taken by the characters and the locations they encounter are tied to specific words, a concept that will be elaborated on below.

Due to the nature of the calculations and the high frequency of words being used, the majority of meaningful vocabulary clustered near the center of the map; however, a clear shape can be seen in the graph above, demonstrated with the line of best fit, stretching from the
northwest to the southeast. Perhaps unsurprisingly, this is the overall direction that the narrative of *The Lord of the Rings* takes – beginning in the Shire to the northwest, traveling across the Misty Mountains and through Rohan, and ending in Gondor and Mordor in the southeast.

Words that are central to the story tend to be located in the center of the cluster, such as “Frodo” (with a longitude of 12.9 and a latitude of 10.7) or “ring” (12.5, 10.5); some words are more self-contained within relevant locations, such as “marsh” (18.2, 11.7), which is drawn towards the geographic location of the Dead Marshes (17.3, 8.1), or “Durthang” (18.1, 7.5), a fortress in Mordor, which is drawn towards Mordor itself.

In contrast, there are some locations that still fall near the center; for example, the geographic centroid of the Shire (12.7, 10.5) is, on the coordinate plane, nearly halfway between the Shire and Mordor. This tends to occur with locations that have high frequencies – not only due to the tendency of high-frequency locations to gravitate towards the center of the plane, but also because they are being referred to in a wide variety of circumstances. The Shire is not only described when the characters are physically there, but also when Frodo and Sam think back on their home while trekking through Mordor and scaling the slopes of Mount Doom. Similarly, Mordor (13.0, 10.7) is not only described when the characters are physically there, but also when members of the Fellowship think of the journey ahead and their ultimate destination, even when Gandalf speaks of it in the Shire. The broader regions of Middle-earth are referred to more often and in a wider variety of contexts, making them more likely to be drawn into the center of the cluster.

Aside from words that are centrally located, there are many words that lie on the fringes of Middle-earth, far to the east or west; these words tend to be tied to specific locations or events in
the story. Some of the more frequent words that occur relatively far to the west (from longitudes ranged roughly from 6 to 9) include some that are associated with the early parts of the narrative. As a simple example, the word *mushrooms* (8.9, 13.9) appears nine times throughout the novel, seven instances of which appear in Chapter 4, fittingly titled “A Shortcut to Mushrooms”; this chapter is set in the outskirts of the Shire, just before the four hobbits cross the Brandywine River into Buckland. Many of the words that are drawn to the west are location names themselves: the Loudwater (8.3, 14.3), or the River Bruinen that flows near Rivendell; the Withywindle (8.5, 13.2), the river that runs through the Old Forest; and the Westfarthing (8.6, 11.4) of the Shire are all located in the western reaches of Middle-earth, so it is appropriate that they are also found in the lower longitudes relative to other words. For the purposes of a more literary analysis, the word *farmers* (8.2, 14.5) is also located far to the west, a fitting descriptor for the laidback people of the Shire. Though there are almost certainly farms to feed the people of Rohan and Gondor, the word “farmers” is only associated with hobbits from the Shire, supporting the homely feel of the entire region, especially when compared to the other, more impressive lands to the east.

Other interesting words that appear far to the west of the central cluster are evocative of specific narrative events. The names of Thingol (6.6, 12.5), the King of Doriath in the First Age, and of Tinúviel (6.6, 12.5), the name given to his daughter Lúthien, have the exact same coordinates, and are both used only in the scene at Weathertop, where Strider tells the hobbits the tale of Beren and Lúthien. This scene is significant; not only does it give a sense of the depth of Tolkien’s mythology, but it cements the hobbits once again as being very grounded in the Shire; these mythical figures, who were powerful and influential in the First Age of Middle-earth, are
simply a legend to them. From a geographical perspective, Goldberry (7.2, 13.3), the wife of
Tom Bombadil, is also located far to the west, where she and Bombadil stay in their realm; this is
fitting, as they do not interfere with the rest of Middle-earth, and she is never seen outside the
area where she resides. When Bombadil rescues the hobbits from the Barrow-Downs, he reminds
them that “Tom has his house to mind, and Goldberry is waiting!” (FR, “Fog on the Barrow-
Downs”).

These words all evoke warmer feelings, being associated with the Shire, the house of Tom
Bombadil, and the Elves of the First Age, all things that could be considered good in Middle-
earth – and the West is normally associated with the good. The East, however, is the opposite,
and the words that are located furthest to the east on the coordinate plane (from longitudes
ranged roughly from 16 to 19) reflect that. The word troubling (18.5, 7.3) is used when Frodo
and Sam are in the Dead Marshes, and later when they are traversing Mordor itself. The words
driver (18.1, 7.5) and durthang (18.1, 7.5) are in reference to the Orc slave-drivers and a fortress
within Mordor, respectively. Other negative words have extremely similar coordinates, being
used heavily or even exclusively in association with Mordor – words like starve (18.1, 10.7),
lash (18.1, 7.5), badness (18.1, 7.5), grassless (18.1, 7.5), brutes (17.8, 7.1), waterless (17.6,
7.6), tortured (17.6, 7.6), and defile (17.4, 7.0).

These words showing the evil of Mordor is clear; when Frodo and Sam draw near to Mount
Doom, Sam hears “the dreadful cries of tortured things” (RK, “Mount Doom”). When Aragorn
marches on the Black Gate, it is said to hide “the shadowy defile beyond” that was “tunnelled by
teeming broods of evil things” (RK, “The Black Gate Opens”). The Words of the West, one could
say, compared to the Words of the East, show a clear path across Middle-earth that echoes the
overall journey taken by the two primary protagonists, Frodo and Aragorn. They travel west to east, and the good of the former fades to the evil of the latter as they traverse Middle-earth.

The words and places found in Middle-earth are shown to form a clear geography, and there are narrative and thematic threads that run through semantically similar words. This can be taken into account for specific locations, as well, using cosine similarity. Calculating the cosine similarity of words provides a wealth of information by measuring how semantically similar they are to one another. The words with higher similarity scores coincide more frequently with the word to which they are being compared. As was previously mentioned, two words that have a similarity score of positive one are one and the same; thus, in the tables and analysis that follow, these results are excluded.
Table 4. Frequent Named Locations with Semantically Similar Locations

<table>
<thead>
<tr>
<th>Location</th>
<th>Similar Locations and Similarity Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Shire</strong></td>
<td></td>
</tr>
<tr>
<td>Buckland</td>
<td>Ered Nimrais 0.7080882</td>
</tr>
<tr>
<td>Michel Delving</td>
<td>0.7131518</td>
</tr>
<tr>
<td>Bag End</td>
<td>Ithilien 0.6765778</td>
</tr>
<tr>
<td>Brandywine</td>
<td>Lossarnach 0.6489313</td>
</tr>
<tr>
<td>Hobbiton</td>
<td>Minas Tirith 0.6384341</td>
</tr>
<tr>
<td></td>
<td>Mindolluin 0.6380255</td>
</tr>
<tr>
<td><strong>Gondor</strong></td>
<td></td>
</tr>
<tr>
<td>Ered Nimrais</td>
<td>Barad-dûr 0.7159829</td>
</tr>
<tr>
<td>Ithilien</td>
<td>Gorgoroth 0.6814617</td>
</tr>
<tr>
<td>Lossarnach</td>
<td>Dark Tower 0.6433909</td>
</tr>
<tr>
<td>Minas Tirith</td>
<td>Orodruin 0.6429616</td>
</tr>
<tr>
<td></td>
<td>Black Gate 0.5892472</td>
</tr>
<tr>
<td><strong>Mordor</strong></td>
<td></td>
</tr>
<tr>
<td>Meduseld</td>
<td>Nan Curunir 0.7756563</td>
</tr>
<tr>
<td>Edoras</td>
<td>Wellinghall 0.7204994</td>
</tr>
<tr>
<td>Mark</td>
<td>Fangorn 0.6768723</td>
</tr>
<tr>
<td>Entwash</td>
<td>Brown Lands 0.6113360</td>
</tr>
<tr>
<td>Gondor</td>
<td>Entwash 0.5730949</td>
</tr>
<tr>
<td><strong>Isengard</strong></td>
<td></td>
</tr>
<tr>
<td>Ford of Bruinen</td>
<td>Ron Curunir 0.6924914</td>
</tr>
<tr>
<td>Bruinen</td>
<td>Brown Lands 0.6113360</td>
</tr>
<tr>
<td>Weathertop</td>
<td>Entwash 0.5730949</td>
</tr>
<tr>
<td>House of Elrond</td>
<td>Hall of Fire 0.5125958</td>
</tr>
<tr>
<td>Hall of Fire</td>
<td>0.5125958</td>
</tr>
<tr>
<td><strong>Rivendell</strong></td>
<td></td>
</tr>
<tr>
<td>Rauros</td>
<td>Gondor 0.6384341</td>
</tr>
<tr>
<td>Númenor</td>
<td>Tol Brandir 0.5935504</td>
</tr>
<tr>
<td>Tol Brandir</td>
<td>Ered Nimrais 0.5792584</td>
</tr>
</tbody>
</table>

As was shown in Table 1, specific places can be measured over the location matrix to semantically compare them against one another. Table 4 displays the named locations in Middle-earth that appear more than one hundred times, with semantically similar words that are strictly other named locations. For most of these examples (as was the case for Table 1), the results are places that are found within or near the initial location; Michel Delving and Hobbiton are cities within the Shire, with Bag End being the Baggins’ home within Hobbiton itself. Buckland is just outside the Shire, and the Brandywine is the river that borders between the two. Similarly, Barad-dûr, Gorgoroth, Orodruin, and the Black Gate are all found within Mordor.
These two regions in particular appear to be relatively self-contained, at least when compared to the smaller city of Minas Tirith, which – surprisingly – is heavily related to the falls of Rauros, the island of Tol Brandir near it, and the long-fallen island of Númenor. Perhaps the departure of Boromir near the aforementioned falls weighted Rauros and Tol Brandir more heavily toward Minas Tirith; far more intriguing, however, is the semantic connection that Minas Tirith shares with Númenor. Tolkien drew strong connections between the two locations through the White Tree – originally Nimloth, the White Tree of Númenor, from which the dead White Tree of Gondor is descended, it serves as a symbol for the line of kings from Isildur down to Aragorn. Númenor having such a relatively strong semantic similarity to Minas Tirith strengthens the connections that Tolkien drew from both a linguistic and a geographic perspective.

As has already been demonstrated in Table 2, individual words have their own lists of similar vocabulary. This has been explored in Table 4 with locations, but any word in the corpus can be examined and compared to others. In order to see how the characters of *The Lord of the Rings* are connected to the world of Middle-earth and the overall narrative, their names can be measured over the word matrix.

*Table 5. The Fellowship with Semantically Similar Vocabulary*

<table>
<thead>
<tr>
<th>Character</th>
<th>Similar Words and Similarity Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Frodo</strong></td>
<td>sam 0.7767146, didn’t 0.7397747, trying 0.7354412, keep 0.7289882, shan’t 0.7039926</td>
</tr>
<tr>
<td><strong>Gandalf</strong></td>
<td>wizard 0.7671267, staff 0.6771592, glamdring 0.6390158, stand 0.6353180, hammer 0.6294722</td>
</tr>
<tr>
<td><strong>Sam</strong></td>
<td>frodo 0.7767146, didn’t 0.7610867, isn’t 0.7409293, mr 0.7337675, gamgee 0.7127476</td>
</tr>
</tbody>
</table>
Table 5 displays the nine members of the Fellowship of the Ring and the five most semantically similar words to each of them. The strongest relationships between characters become abundantly clear; as Frodo and Sam not only spend the entire novel together, but are also frequently mentioned in the same breath by others referring to them and their journey, their names appear as the most semantically similar word to one another. In a similar vein, the Three Hunters all have semantically similar names to one another; Aragorn, Legolas, and Gimli forge their own bond and travel together in search of Merry and Pippin.

Words with high similarities can also be used to develop insights about the journey of the character and thematic connections can be drawn (though reading too much into every word is a dangerous and slippery slope). For Frodo, the words *keep* and *trying* are evocative of his sense of duty and perseverance, continuing on with his quest despite his expectation of failure and the weight of the Ring upon him.
For Gimli and Legolas, the word *dwarf* is strongly associated with both characters, whereas the word *elf* is not. This is likely due to the fact that in *The Lord of the Rings*, various members of Fellowship comes across elves relatively frequently – Frodo meets Gildor in the woods of the Shire, the Fellowship forms in Rivendell, and Aragorn leads them to Lothlórien – but as Gimli is the only living dwarf that has any relevance to the narrative, the majority of references to dwarves are in direct reference to him and to the Three Hunters.

Gandalf, of course, is the most prominent wizard in the novel, and it is no surprise that he is strongly associated with his staff and his sword, Glamdring. Merry and Pippin have their full names in their results, with seemingly random words appearing near them; however, in the case of Pippin, the words *betide, embrasure*, and *ernil* (an Elvish word meaning “prince”) only appear in chapters that take place in Minas Tirith, the location where Pippin is finally “knighted”. The results for Merry are less clear, though this could be that because Merry is “knighted” in Rohan, and with Rohan being a realm in which the narrative is constantly in motion – the riding of horses, the fighting of battles, and Merry’s eventual journey away from Rohan into Gondor – there are less established words to tie him down, so to speak.

Lastly, Boromir’s results have some clear connections, the strongest of which is his tie to the company, or the Fellowship itself. Other characters have the word *company* associated with them (or in Aragorn’s case, *companions*) further along in their results, but only Boromir has the word take the top spot. One could suggest this giving him the strongest ties to the company; after all, his death has a profound impact on Aragorn and helps catalyze the breaking of the Fellowship. The other four words, however, evoke different images and tell a clearer story: one of Boromir at the falls of Rauros, unable to overcome the burden of the Ring weighing on his
mind, falling to the temptation to take it from Frodo, then redeeming himself in death and being sent over the falls with honor in a funeral boat.

Table 6. Other Characters with Semantically Similar Vocabulary

<table>
<thead>
<tr>
<th>Character</th>
<th>Similar Words and Similarity Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faramir</td>
<td>stewards 0.8967964, risked 0.8500669, captain 0.8454338, perceiving 0.8427889, dínen 0.8101560</td>
</tr>
<tr>
<td>Saruman</td>
<td>orthanc 0.8117029, persuade 0.7753027, conceal 0.7686051, foaled 0.7686051, grievances 0.7686051</td>
</tr>
<tr>
<td>Éomer</td>
<td>mark 0.8706302, théoden 0.8637906, eorl 0.8199018, king 0.8109946, household 0.7807262</td>
</tr>
<tr>
<td>Théoden</td>
<td>ride 0.8967258, edoras 0.8822574, éomer 0.8637906, mark 0.8627714, king 0.8567600</td>
</tr>
<tr>
<td>Sauron</td>
<td>crave 0.8435983, insolence 0.7875551, foreseen 0.7849458, isildur 0.7758541, destroyed 0.7746595</td>
</tr>
<tr>
<td>Éowyn</td>
<td>women 0.8656680, woman 0.8147677, king 0.8063194, begs 0.8002305, flashing 0.7921499</td>
</tr>
<tr>
<td>Galadriel</td>
<td>celeborn 0.8266605, gift 0.7700119, gem 0.7303585, overlaid 0.7064388, blows 0.6800466</td>
</tr>
</tbody>
</table>

Aside from the Fellowship, there are numerous supporting characters (and antagonists) that play crucial roles in the narrative of *The Lord of the Rings*. Table 6 displays seven of these characters with each of their five most semantically similar words, and once again, several thematic connections are clearly demonstrated through computational data.

Faramir is not only the son of the current Steward of Gondor (who ascends to his father’s position briefly before allowing Aragorn his rightful place on the throne), but also the Captain of the Rangers of Ithilien, explaining the associations of both *stewards* and *captain*. The other three
results, however, are much more worthy of discussion. The word *dínen* is Sindarin for “silent”; while its association with Faramir likely comes from it being a portion of a street name in Minas Tirith, Faramir’s home, it is also a fitting (if unintended) adjective to describe the man who has spent his life in Boromir’s shadow. In contrast, perhaps, to his image as the more cautious of the two sons of Denethor, the word *risked* appears very strongly associated with Faramir. As Beregond tells Pippin in *The Return of the King*:

> [Faramir] is bold, more bold than many deem; for in these days men are slow to believe that a captain can be wise and learned in the scrolls of lore and song, as he is, and yet a man of hardihood and swift judgement in the field. But such is Faramir. Less reckless and eager than Boromir, but not less resolute (*RK*, “Minas Tirith”).

The other attribute of Faramir exemplified by a semantically similar word is that of perceptiveness. The word *perceiving* may indicate Faramir’s ability not only to see the truth in Frodo and Sam’s explanations about the Ring, but also his immediate realization of the true identity of Aragorn in the Houses of Healing.

Other character insights that can be gleaned from this particular data set center around the two main antagonists of the novel, Sauron and Saruman. The word most closely associated with Sauron is *crave*, indicative of his lust for power and hunger for domination over Middle-earth. Sauron craves the One Ring, and will do anything to obtain it. The words *isildur* and *destroyed* are likely considered semantically similar to Sauron due simply to the frequent discussion on seeing Sauron destroyed, and mention of Isildur – either in reference to the end of the Second Age, when Isildur cut the One Ring from Sauron’s hand, or when referring to the One Ring itself as “Isildur’s Bane”. The word *foreseen* appearing in the results suggests Sauron’s propensity for planning, though the irony is heavy in light of the one thing he could not have foreseen – two
hobbits trekking through his own lands to destroy the One Ring, rather than use it against him as he expected.

Saruman is, of course, associated with his tower Orthanc, but the words **persuade, conceal, and grievances** are of particular interest. The voice of Saruman is well documented to have a power over those who hear it, as he uses his wizardry to sway people to his side. He even tries to persuade Gandalf to join him, though Gandalf refuses – after learning that Saruman had concealed his possession of a **palantír** from the rest of the White Council. All of Saruman’s plans were kept from his allies as he feigned allegiance to Sauron with the intent of ruling Middle-earth with power of the One Ring. These plans lead to the grievances committed against the people of Rohan, through his manipulation of Théoden and the events that lead to the battle at Helm’s Deep.

Interestingly, the three characters that hail from Rohan – Éomer, Théoden, and Éowyn – have some similarities in their results. They all share the word **king**, which appears to be associated heavily with Rohan rather than Gondor (given the Steward that sits on the throne at Minas Tirith in the absence of their king), and Éomer and Théoden in particular share not only one another’s names, but the word **mark**, in reference to the Riddermark (the Rohirrim’s name for Rohan itself). The word most closely associated with Théoden is **ride**, which comes as no surprise given Rohan’s connection to horses and the proclivity of Théoden himself to ride into battle (both at Helm’s Deep and the Pelennor Fields).

Éowyn deserves special mention aside from her shared results with the other two; both the word **women** and its singular form **woman** appear as highly similar words. It is no coincidence that both words are so strongly associated with Éowyn; she is, in her own words to the Witch-
king on the Pelennor Fields, “no living man” (RK, “The Battle of the Pelennor Fields”), and is thus able to slay him. The pain of Éowyn being unable to accompany the men of Rohan into battle is central to her character, and her womanhood ends up as the very thing that gives her the power to kill the Witch-king, a crucial moment in the narrative.

Finally, Galadriel’s results, particularly the first three words, give an insight into her role in the story. Celeborn is the name of her husband, the Lord of Lórien; Lothlórien is where she resides and meets the Fellowship after their journey through Moria. The word gem is frequently used in association with elves, and is used to describe the Three Rings untouched by Sauron, one of which Galadriel herself wields. But perhaps Galadriel’s most important role in the narrative is her giving of gifts to each member of the Fellowship, especially to Frodo and Sam. The lembas bread keep them alive in Mordor, and the Elven cloaks help keep them hidden on their journey. The Elven rope given to Sam becomes particularly useful in traversing the Emyn Muil, but the most crucial item they receive is the Phial of Galadriel: used not only in fending off the attacks of Shelob, but also to overcome the will of the Two Watchers at the Tower of Cirith Ungol. All together, the gifts of Galadriel are imperative in the success of the Quest by helping Frodo and Sam travel through the heart of Sauron’s realm, making the semantic similarity of the word gift that much more potent.

Table 7. Polyonymous People with Semantically Similar Vocabulary

<table>
<thead>
<tr>
<th>Name</th>
<th>Similar Words and Similarity Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gollum</td>
<td>sméagol 0.9311162 precious 0.8827871 yess 0.8264254 nice 0.8248856 yes 0.7889800</td>
</tr>
<tr>
<td>Sméagol</td>
<td>gollum 0.9311162 nice 0.8780458 precious 0.8609962 yess 0.8260633 slinker 0.8044042</td>
</tr>
<tr>
<td></td>
<td>wizard staff glamdring stand hammer</td>
</tr>
</tbody>
</table>
Tolkien has a habit of creating people and places with multiple names; Rivendell is also called Imladris, Gandalf is also called Mithrandir, and Mount Doom is also called Orodruin. Each separate name has a distinct meaning and is used in distinct contexts, and Table 5 displays the semantic differences between various people in Middle-earth with multiple names.

Four of these instances are of particular interest. The first is Gollum, though not because of vast semantic differences with his alter-ego Sméagol – on the contrary, the two are extraordinarily similar. It is interesting in and of itself that his two names are so closely related, while the other examples (as will be explored below) do not share this. Perhaps the personalities of Gollum and Sméagol are so closely linked, and the two are so irrevocably tied together, that even semantically, it is impossible to untangle one from the other. Also appearing in both names’ results (aside from the words that he speaks frequently, the slithering yess and nice), is the other object that cannot be sundered from his being – precious. Gollum’s connection to the One Ring is so strong, his own name for it ranks among the highest semantically similar words to both sides of himself.

<table>
<thead>
<tr>
<th>Name</th>
<th>1st Name</th>
<th>2nd Name</th>
<th>3rd Name</th>
<th>4th Name</th>
<th>5th Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gandalf</td>
<td>0.7671267</td>
<td>0.6771592</td>
<td>0.6390158</td>
<td>0.6353180</td>
<td>0.6294722</td>
</tr>
<tr>
<td>Mithrandir</td>
<td>beregond</td>
<td>citadel</td>
<td>halfling</td>
<td>duties</td>
<td>betide</td>
</tr>
<tr>
<td></td>
<td>0.8963902</td>
<td>0.8892151</td>
<td>0.8713649</td>
<td>0.8645035</td>
<td>0.8571764</td>
</tr>
<tr>
<td>Strider</td>
<td>weathertop</td>
<td>ferny</td>
<td>comfortless</td>
<td>crosses</td>
<td>presses</td>
</tr>
<tr>
<td></td>
<td>0.9181024</td>
<td>0.8293248</td>
<td>0.8241551</td>
<td>0.8238455</td>
<td>0.7938950</td>
</tr>
<tr>
<td>Aragorn</td>
<td>legolas</td>
<td>gimli</td>
<td>nonetheless</td>
<td>judge</td>
<td>must</td>
</tr>
<tr>
<td></td>
<td>0.8457541</td>
<td>0.8282479</td>
<td>0.7582307</td>
<td>0.7245775</td>
<td>0.6987855</td>
</tr>
<tr>
<td>Elessar</td>
<td>elfstone</td>
<td>especial</td>
<td>warden</td>
<td>healers</td>
<td>healer</td>
</tr>
<tr>
<td></td>
<td>0.8990467</td>
<td>0.7665790</td>
<td>0.7650582</td>
<td>0.7560882</td>
<td>0.7495625</td>
</tr>
</tbody>
</table>
The second polyonymous person is one that has already been covered to an extent—Gandalf’s semantic attributes have been discussed, and the difference seen between his more commonly used name and the Sindarin name of Mithrandir appears heavily tied to location. While words strongly associated with Gandalf are broader words that relate to him as a character, when he is referred to as Mithrandir, it is strongly weighted towards Gondor and the city of Minas Tirith. The citadel (one of the most important landmarks in the city), beregend (the name of the soldier in Minas Tirith that helped Gandalf save Faramir’s life), halfling (the word used for hobbits that appears frequently in the lands of Gondor), duties (the men of Gondor are strongly beholden to their duties, with some even obeying Denethor in his madness), and betide (a word also strongly associated with Pippin, and only used in chapters that are set in Minas Tirith). This suggests that the word Mithrandir is strongly tied to Minas Tirith, and if that is indeed the case, then their relative latitudes and longitudes should be close to one another. As it happens, mithrandir has a longitude of 14.4 and a latitude of 9.3, while minas tirith (calculated using the geodata for the word tirith, given that it only ever coincides with minas in reference to the city, while minas is used in reference to other locations such as Minas Anor and Minas Ithil) has a longitude of 14.5 and a latitude of 9.6. The two words plot incredibly close to one another on the coordinate plane, and are both weighted heavily towards the southeastern end of the word cluster from Figure 1, the direction of Gondor and the White City itself.

Aragorn also stands out as a semantically interesting character, with not just two, but three different names used across the novel. When he is introduced to the narrative, he is referred to as Strider, and so under this name he is associated with the events that unfold on the journey from Bree to Rivendell; fenny (the name of the man that acted as a spy for the Nazgûl in Bree),
weathertop (the location at which Strider fended off the Nazgûl themselves and Frodo was stabbed by the Morgul blade), and crosses (indicative of the crossing of the Bruinen, to safety and towards Rivendell) stand out in particular. For the majority of the novel, he is referred to as Aragorn, and his lengthy journey alongside Gimli and Legolas weighs the three names heavily towards one another, as has been discussed. His royal name, Elessar, is extremely similar to the word elfstone, which is what “Elessar” means in Quenya. The words warden, healers, and healer are all reminiscent of the Houses of Healing in Minas Tirith, where Aragorn finally agrees to enter the city to help save the men and women afflicted by the Black Breath of the Nazgûl.

Before this, “Elessar” is only used seven times, and all in crucial moments in the narrative: once by Galadriel after the Fellowship’s journey through Moria and into Lothórien; once as they pass by the statues of Isildur and Anárion at the Falls of Rauros; once as the Three Hunters reveal themselves to the Riders of Rohan; twice when Gandalf tells of his return to Middle-earth; once when Aragorn proclaims mastery over the Shadow Host with the Grey Company; and once when he emerges from the ships at the Battle of the Pelennor Fields to help turn the tide against the armies of Mordor. Elessar is clearly an important name for Aragorn, and it is fitting that the return of the King to Gondor is accompanied by the use of the royal name, and one that is so heavily associated with Minas Tirith.

This portion of analysis is meant to show that the semantics of Tolkien’s work align with his narrative structure, with glimpses into the narrative space itself. It demonstrates how tied certain characters and events are to specific locations or regions. With all this in mind, the concepts previously introduced with Table 2 can be more deeply explored; finding keywords for places that serve as crucial locations in the story will further illustrate the importance of Middle-
earth’s geography to the narrative.

Table 8. Frequently Named Locations with Semantically Similar Vocabulary

<table>
<thead>
<tr>
<th>Location</th>
<th>Similar Words and Similarity Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bag End</strong></td>
<td></td>
</tr>
<tr>
<td>Bag End</td>
<td>bagshot 0.9199458</td>
</tr>
<tr>
<td></td>
<td>admittance 0.8559210</td>
</tr>
<tr>
<td></td>
<td>stocks 0.8559210</td>
</tr>
<tr>
<td></td>
<td>carts 0.8229759</td>
</tr>
<tr>
<td><strong>Shire</strong></td>
<td></td>
</tr>
<tr>
<td>Shire</td>
<td>year 0.7419132</td>
</tr>
<tr>
<td></td>
<td>hobbits 0.6625460</td>
</tr>
<tr>
<td></td>
<td>frodo 0.6513328</td>
</tr>
<tr>
<td></td>
<td>called 0.6401848</td>
</tr>
<tr>
<td><strong>Bree</strong></td>
<td></td>
</tr>
<tr>
<td>Bree</td>
<td>greenway 0.8928279</td>
</tr>
<tr>
<td></td>
<td>gatekeeper 0.8291039</td>
</tr>
<tr>
<td></td>
<td>inhabitants 0.7851658</td>
</tr>
<tr>
<td></td>
<td>accounting 0.7708456</td>
</tr>
<tr>
<td><strong>Weathertop</strong></td>
<td></td>
</tr>
<tr>
<td>Weathertop</td>
<td>strider 0.8885272</td>
</tr>
<tr>
<td></td>
<td>tinúviel 0.8131434</td>
</tr>
<tr>
<td></td>
<td>bolsters 0.8131434</td>
</tr>
<tr>
<td></td>
<td>conical 0.8131434</td>
</tr>
<tr>
<td>Rivendell</td>
<td>bruinen 0.6979195</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Rivendell</td>
<td>glorfindel 0.6685963</td>
</tr>
<tr>
<td>Rivendell</td>
<td>gravest 0.6372138</td>
</tr>
<tr>
<td>Rivendell</td>
<td>wound 0.6093988</td>
</tr>
<tr>
<td>Moria</td>
<td>anvil 0.8895135</td>
</tr>
<tr>
<td>Moria</td>
<td>bewilder 0.8658382</td>
</tr>
<tr>
<td>Moria</td>
<td>creek 0.8658382</td>
</tr>
<tr>
<td>Moria</td>
<td>darksome 0.8658382</td>
</tr>
<tr>
<td>Rohan</td>
<td>war 0.8411767</td>
</tr>
<tr>
<td>Rohan</td>
<td>proud 0.7717276</td>
</tr>
<tr>
<td>Rohan</td>
<td>my 0.7584183</td>
</tr>
<tr>
<td>Rohan</td>
<td>may 0.7485747</td>
</tr>
<tr>
<td>Isengard</td>
<td>nan 0.7730207</td>
</tr>
<tr>
<td>Isengard</td>
<td>leisurely 0.6993997</td>
</tr>
<tr>
<td>Isengard</td>
<td>quickbeam 0.6493095</td>
</tr>
<tr>
<td>Isengard</td>
<td>hom 0.6310495</td>
</tr>
<tr>
<td>Minas Tirith</td>
<td>boromir 0.7760613</td>
</tr>
<tr>
<td>Minas Tirith</td>
<td>surely 0.6731995</td>
</tr>
</tbody>
</table>
Table 8 displays significant locations that appear in *The Lord of the Rings* in the general order that they are visited, and the twenty words that are most semantically similar to each one with their respective similarity scores (again, excluding the location name itself, which would have a similarity score of positive one). While examining all of these would be too much for the purposes of this paper, three of the more interesting locations – Rivendell, Rohan, and Mordor – will be discussed further, and examined as three stages of the characters’ overall journey.

The first location, Rivendell, is a place of respite for the hobbits after their grueling journey from the comfort of the Shire through the wilderness, culminating at Weathertop and the flight to and across the Ford of Bruinen. As the House of Elrond, Rivendell (or Imladris) is one of the three remaining havens for Elves in Middle-earth (the others being Lothlórien and the Woodland
Realm). When Frodo inquires if it is safe, Gandalf explains that “there is a power in Rivendell to withstand the might of Mordor, for a while” (*FR*, “Many Meetings”). Like everything in Middle-earth, however, the safety of Rivendell will not last forever, and so the Council of Elrond meets to decide the appropriate course of action to be taken. The words in Table 8 that are associated with Rivendell paint a clear picture of the location’s purpose in the narrative. Frodo uses Rivendell as a place of healing, to recover from his encounter with the Nazgûl (or ringwraiths); Gandalf tells him that “there was some fragment of the blade still in the closed wound”, but that the wound is “mending fast … Elrond has cured you” (*FR*, “Many Meetings”). Despite the perils that lay ahead of them, “such was the virtue of the land of Rivendell that soon all fear and anxiety was lifted from their minds” (*FR*, “The Ring Goes South”).

Rivendell serves another purpose for Strider, now properly called Aragorn. It is here that the full extent of his lineage is revealed, when Gandalf explains to Frodo that Aragorn is not “just a Ranger”, but “the last remnant … of the great people, the Men of the West” and a member of the “race of the Kings from over the Sea” (*FR*, “Many Meetings”). Bilbo tells Frodo why Aragorn is referred to as “the Dúnadan”, it being Elvish for “Man of the West” or “Númenorean”. It is in Rivendell, also, that the shards of Narsil, the sword of Elendil that Isildur used to cut the One Ring from Sauron’s hand, are reforged into Andúril, Flame of the West. The transition of the character from Strider to Aragorn is completed in Rivendell with the reforging of the blade that was broken.

Though relatively safe, Rivendell is also the place where Frodo begins to see the effects of the Ring he bears. When Bilbo asks to see it again, Frodo “[feels] a strange reluctance” before “a shadow … [falls] between them” and Bilbo is seen as a creature reminiscent of Gollum (*FR*, “Many Meetings”).
“Many Meetings”). Later, at the meeting of the Council of Elrond, Gandalf instructs Frodo to bring out the Ring and show it to everyone present, and Frodo again feels “a great reluctance to reveal the Ring, and a loathing of its touch” (FR, “The Council of Elrond”). This culminates in Frodo as the protagonist reluctantly accepting his burden and the quest before him:

A great dread fell on him, as if he was awaiting the pronouncement of some doom that he had long foreseen and vainly hoped might after all never be spoken. An overwhelming longing to rest and remain at peace by Bilbo’s side in Rivendell filled all his heart. At last with an effort he spoke, and wondered to hear his own words, as if some other will was using his small voice (FR, “The Council of Elrond”).

Rivendell is crucial to the narrative for these three reasons in particular; it is a place for Frodo to heal, a place for Aragorn to realize his legacy, and a place for Frodo to accept the burden of the Ring. It is located in the western reaches of Middle-earth, and serves as a major starting point for Frodo and Aragorn’s journeys to Mordor and Gondor, respectively. The central part of Aragorn’s journey leads fittingly through central Middle-earth, in the kingdom of Rohan. He leads the Three Hunters in pursuit of Merry and Pippin, reluctantly leaving Frodo and Sam to their own quest.

Rohan is a kingdom consistently embroiled in war, first against Isengard and eventually against Mordor alongside Gondor; Aragorn says to Éomer that “open war lies before him, with Sauron or against him”, and Éomer tells of Gandalf’s warning that “sudden war was preparing in Isengard” (TT, “The Riders of Rohan”). The Rohirrim that reside there are described by Aragorn as “proud and wilful, but they are true-hearted, generous in thought and deed; bold but not cruel; wise but unlearned, writing no books but singing many songs” (TT, “The Riders of Rohan”). As the Riders of Rohan, horses are extremely important to their culture, and horses from Rohan “were of great stature, strong and clean-limbed; their grey coats glistened, their long tails flowed
in the wind, their manes were braided on their proud necks” (TT, “The Riders of Rohan”). The words war, proud, and horses all appear in Table 8 as being semantically similar to Rohan, and again, are able to paint a clear picture of the importance of the location to the narrative. When meeting the Rohirrim, Aragorn introduces himself as Elessar for the first time, showing his reforged blade and momentarily appearing as a majestic king of old. This is indicative of another step in Aragorn’s journey to accepting his role as King of Gondor.

Rohan also marks the importance of the Rohirrim themselves, who ride to the aid of Gondor and prove the two kingdoms’ loyalty to one another. The Riders’ rescue of Minas Tirith renews the closeness between Gondor and Rohan through battle, and the courage and heroism displayed by them at Helm’s Deep as well as the Pelennor Fields is inspiring, showing Sauron that the world of Men can be united once more to overthrow him. Though Gondor is a different kingdom, it too appears in Table 8, making the connection between it and Rohan quantifiable.

The word battle also appears, further cementing the importance of battles to Rohan and its people; Helm’s Deep is a fight for their peoples’ survival, and at the Pelennor Fields (though located in Gondor), the Riders come to save the day in battle.

Finally, the third stage of Frodo’s journey takes him into the Land of Shadow itself: Mordor. Its role in the narrative is clear: reaching it is the goal of the quest, to go into the heart of Mordor and destroy the Ring in the fires of Mount Doom. One would expect the word evil to be strongly associated with Mordor, the center of evil in Middle-earth during the Third Age. The tower of Barad-dûr serves as a bastion of evil and the stronghold of Sauron’s power; when debating whether to ride to the aid of Minas Tirith, Théoden laments that “[s]o great a power as the Dark Lord seems now to wield might well contain us in battle before the City”, showing that
Sauron’s foes know how powerful he is (RK, “The Muster of Rohan”). Barad-dûr itself is a symbol of the Dark Lord’s might, and as the Dark Tower, it contrasts greatly with the White Tower of Minas Tirith; “out from the Dark Tower there crept the veils of Shadow that Sauron wove about himself (RK, “Mount Doom”). Other words from Table 8 appear in direct reference to Sauron himself; when Sam is forced to carry the Ring for a time, it tempts him, and he considers the option to “challenge the Power that sat in its dark hold beyond the valley of shadows” (RK, “The Tower of Cirith Ungol”). Frodo mentions seeing “a map of Mordor that was made before the Enemy came back here” (RK, “The Land of Shadow”), and Ithilien is said to have fallen “under the shadow of their Enemy” (RK, “Minas Tirith”).

The word *fear* is indicative of Sauron’s main power over the free people of Middle-earth; “yet another weapon, swifter than hunger, the Lord of the Dark Tower had: dread and despair” (RK, “The Siege of Gondor”). When he is defeated at last, his mind is shaken free “[f]rom all his policies and webs of fear and treachery” to focus solely on the fate of the Ring; ironically, Sauron is subject to fear himself in his final moments (RK, “Mount Doom”). Strange, perhaps, is the association of the word *hope* with such a hopeless place as Mordor, though there are distinct differences in some of its occurrences. Frodo laments to Sam as they trek through Mordor, telling him that he will follow Sam “[a]s long as you’ve got any hope left. Mine is gone” (RK, “The Land of Shadow”). When the Ring is destroyed, however, the Captains of the West see the weakened armies of Mordor, and “their hearts were filled with a new hope in the midst of darkness”; similarly, Frodo and Sam sleep after the completion of their quest, and “in the morning they rose again in hope and peace” (RK, “The Field of Cormallen”).
The purpose of Frodo journeying into Mordor is to see the Ring destroyed, completing his quest narrative. It is fitting, then, that the word *destroyed* should associate closely with Mordor, the former being the goal and the latter the destination. Gandalf makes it clear when he says that “[i]f [the Ring] is destroyed, then [Sauron] will fall; and his fall will be so low that none can foresee his arising ever again … And so a great evil of this world will be removed” (*RK*, “The Last Debate”). The destruction of the Ring will lead directly to the destruction of Sauron, and as Gandalf says, “[i]n wisdom or folly it has been sent away to be destroyed, lest it destroy us” (*RK*, “The Last Debate”). The Ring is an object of evil, and is tied to Sauron’s power.

Tolkien tends to use color as a metric for morality. Green is associated with living things, trees, and the beauty of the natural world; Aragorn even reprimands one of the Rohirrim by claiming the green earth to be “a mighty matter of legend, though you tread it under the light of day” (*TT*, “The Riders of Rohan”). Gold and silver are often associated with the grace of the Elves; Lothlórien is even called the “Land of the Valley of Singing Gold” and the “Golden Wood”, and Elrond and Arwen wear silver in Rivendell. Black, however, is a color that Tolkien associates with evil; Mordor is the “Black Land”, its entrance is the “Black Gate”, the ringwraiths are the “Black Riders”, and their leader is the “Black Captain”. The depth of evil in Mordor is demonstrated by its association with the word *black*, and Frodo and Sam are keenly aware of it as they trek through the barren lands:

But far worse than all such perils was the ever-approaching threat that beat upon them as they went: the dreadful menace of the Power that waited, brooding in deep thought and sleepless malice behind the dark veil about its Throne. Nearer and nearer it drew, looming blacker, like the oncoming of a wall of night at the last end of the world (*RK*, “Mount Doom”).
Frodo also sees the peak of Barad-dûr while climbing Mount Doom; “… and then he saw, rising black, blacker and darker than the vast shades amid which it stood, the cruel pinnacles and iron crown of the topmost tower of Barad-dûr” (RK, “Mount Doom”). Despite the evil of Mordor, the quest succeeds (though due only to the interference of Gollum), and the characters’ journeys end with a triumphant return to the west, with Aragorn being crowned as the King of Gondor and accepting his place on the throne, and Frodo returning west to the Shire – then going even further west, sailing from the Grey Havens and departing Middle-earth forever.
Conclusion

Applying the technique of distant reading through statistical models provides a different look at the geography and narrative structure of Tolkien’s work than one would have simply by reading through it. Tolkien was able to create a story and world in which the narrative structure is complemented by the semantic ties of specific words to one another, with particular emphasis on locations and how they affect the journeys of the characters. The mythology and culture he created and applied to Middle-earth has deep thematic connections to the overall narrative, and both of these have similar connections to the geography of the world itself. As quoted in the Previous Scholarship section above, “Tolkien’s imagination is not just mythopoeic, but also geographic” (Holmes 2014, 136). The meaning of this becomes clearer in light of the quantitative connections that can be drawn between locations in Middle-earth and the narrative and themes of *The Lord of the Rings* through distant reading. Characters travel through the space of Middle-earth, and each location they come across not only serves an important role in the narrative, but also makes the world feel richer, more alive, and more connected with its own internal laws and legends.

This paper not only seeks to provide a deeper understanding of how the narrative space influences the characters and the story, but also to demonstrate the capability of spatial text analyses to be performed on fictional worlds such as Middle-earth. Other authors, such as George R. R. Martin or C. S. Lewis, may have similar aspects to their writings in Westeros and Narnia, respectively. There is also more to Middle-earth that is not included in this analysis: the history of the First and Second Ages are covered in *The Silmarillion*, along with the full extent of Tolkien’s mythos and a chronicling of how the geography of Middle-earth changes over time.
These would all be relevant to further semantic analysis, but the time horizons and scope of *The Silmarillion* excluded it from this paper. Based on these results, however, it can be said that the geography of Middle-earth, the path taken by the narrative, and the overall themes associated with various locations were all intentionally placed by Tolkien himself, and these connections can now be quantitatively shown through a computational and statistical approach.


Appendix A: Maps, Graphs, and Tables

Figure 1
Figure 2
Table 1. Given Location Measured over Location Matrix

<table>
<thead>
<tr>
<th>Location</th>
<th>Similarity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bree</td>
<td>1.0000000</td>
</tr>
<tr>
<td>Bree-land</td>
<td>0.9147455</td>
</tr>
<tr>
<td>Staddle</td>
<td>0.8644356</td>
</tr>
<tr>
<td>Archet</td>
<td>0.8424889</td>
</tr>
<tr>
<td>Greenway</td>
<td>0.8275395</td>
</tr>
<tr>
<td>Bree-hill</td>
<td>0.8170181</td>
</tr>
<tr>
<td>Combe</td>
<td>0.7869258</td>
</tr>
</tbody>
</table>

Table 2. Given Word Measured over Word Matrix

<table>
<thead>
<tr>
<th>Word</th>
<th>Similarity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>tree</td>
<td>1.0000000</td>
</tr>
<tr>
<td>branches</td>
<td>0.8613180</td>
</tr>
<tr>
<td>trunk</td>
<td>0.8374366</td>
</tr>
<tr>
<td>trees</td>
<td>0.8003490</td>
</tr>
<tr>
<td>rustle</td>
<td>0.7612032</td>
</tr>
<tr>
<td>blending</td>
<td>0.7446865</td>
</tr>
<tr>
<td>boughs</td>
<td>0.7124064</td>
</tr>
</tbody>
</table>

Table 3. Given Location Measured over Word Matrix

<table>
<thead>
<tr>
<th>Word</th>
<th>Similarity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>bree</td>
<td>0.9994855</td>
</tr>
<tr>
<td>greenway</td>
<td>0.8928279</td>
</tr>
<tr>
<td>butterbur</td>
<td>0.8754266</td>
</tr>
<tr>
<td>landlord</td>
<td>0.8591487</td>
</tr>
<tr>
<td>staddle</td>
<td>0.8585841</td>
</tr>
<tr>
<td>landers</td>
<td>0.8576275</td>
</tr>
<tr>
<td>gatekeeper</td>
<td>0.8291039</td>
</tr>
<tr>
<td>Location</td>
<td>Similar Locations and Similarity Scores</td>
</tr>
<tr>
<td>----------</td>
<td>-----------------------------------------</td>
</tr>
<tr>
<td><strong>Shire</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shire</td>
</tr>
<tr>
<td><strong>Gondor</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gondor</td>
</tr>
<tr>
<td><strong>Mordor</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mordor</td>
</tr>
<tr>
<td><strong>Rohan</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rohan</td>
</tr>
<tr>
<td><strong>Isengard</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Isengard</td>
</tr>
<tr>
<td><strong>Rivendell</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rivendell</td>
</tr>
<tr>
<td><strong>Minas Tirith</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Minas Tirith</td>
</tr>
</tbody>
</table>
Table 5. The Fellowship with Semantically Similar Vocabulary

<table>
<thead>
<tr>
<th>Character</th>
<th>Similar Words and Similarity Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frodo</td>
<td>sam 0.7767146 didn’t 0.7397747 trying 0.7354412 keep 0.7289882 shan’t 0.7039926</td>
</tr>
<tr>
<td>Gandalf</td>
<td>wizard 0.7671267 staff 0.6771592 glamdring 0.6390158 stand 0.6353180 hammer 0.6294722</td>
</tr>
<tr>
<td>Sam</td>
<td>frodo 0.7767146 didn’t 0.7610867 isn’t 0.7409293 mr 0.7337675 gamgee 0.7127476</td>
</tr>
<tr>
<td>Aragorn</td>
<td>legolas 0.8457541 gimli 0.8282479 nonetheless 0.7582307 judge 0.7245775 must 0.6987855</td>
</tr>
<tr>
<td>Pippin</td>
<td>peregrin 0.8564995 betide 0.8258391 embrasure 0.8258391 ernil 0.8258391 fists 0.8258391</td>
</tr>
<tr>
<td>Merry</td>
<td>score 0.6991391 meriadoc 0.6707345 fellows 0.6647251 call 0.6638438 stared 0.6463807</td>
</tr>
<tr>
<td>Gimli</td>
<td>legolas 0.9139816 aragorn 0.8282479 dwarf 0.7563165 company 0.7247781 judge 0.7122482</td>
</tr>
<tr>
<td>Legolas</td>
<td>gimli 0.9139816 aragorn 0.8457541 dwarf 0.7290931 nonetheless 0.7152580 judge 0.7122482</td>
</tr>
<tr>
<td>Boromir</td>
<td>company 0.6789346 rauros 0.6623723 cannot 0.6387438 heavily 0.6330999 boat 0.6320360</td>
</tr>
</tbody>
</table>
Table 6. Other Characters with Semantically Similar Vocabulary

<table>
<thead>
<tr>
<th>Character</th>
<th>Similar Words and Similarity Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faramir</td>
<td>stewards 0.8967964, risked 0.8500669, captain 0.8454338, perceiving 0.8427889, dínen 0.8101560</td>
</tr>
<tr>
<td>Saruman</td>
<td>orthanc 0.8117029, persuade 0.7753027, conceal 0.7686051, foaled 0.7686051, grievances 0.7686051</td>
</tr>
<tr>
<td>Éomer</td>
<td>mark 0.8706302, théoden 0.8637906, eorl 0.8199018, king 0.8109946, household 0.7807262</td>
</tr>
<tr>
<td>Théoden</td>
<td>ride 0.8967258, edoras 0.8822574, éomer 0.8637906, mark 0.8627714, king 0.8567600</td>
</tr>
<tr>
<td>Sauron</td>
<td>crave 0.8435983, insolence 0.7875551, foreseen 0.7849458, isildur 0.7758541, destroyed 0.7746595</td>
</tr>
<tr>
<td>Éowyn</td>
<td>women 0.8656680, woman 0.8147677, king 0.8063194, begs 0.8002305, flashing 0.7921499</td>
</tr>
<tr>
<td>Galadriel</td>
<td>celeborn 0.8266605, gift 0.7700119, gem 0.7303585, overlaid 0.7064388, blows 0.6800466</td>
</tr>
<tr>
<td>Name</td>
<td>Similar Words and Similarity Scores</td>
</tr>
<tr>
<td>-----------</td>
<td>------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Gollum</td>
<td>sméagol 0.9311162, precious 0.8827871, yess 0.8264254, nice 0.8248856, yes 0.7889800</td>
</tr>
<tr>
<td>Sméagol</td>
<td>gollum 0.9311162, nice 0.8780458, precious 0.8609962, yess 0.8260633, slinker 0.8044042</td>
</tr>
<tr>
<td>Gandalf</td>
<td>wizard 0.7671267, staff 0.6771592, glamdring 0.6390158, stand 0.6353180, hammer 0.6294722</td>
</tr>
<tr>
<td>Mithrandir</td>
<td>beregond 0.8963902, citadel 0.8892151, halfling 0.8713649, duties 0.8645035, betide 0.8571764</td>
</tr>
<tr>
<td>Strider</td>
<td>weathertop 0.9181024, ferny 0.8293248, comfortless 0.8241551, crosses 0.8238455, presses 0.7938950</td>
</tr>
<tr>
<td>Aragorn</td>
<td>legolas 0.8457541, gimli 0.8282479, nonetheless 0.7582307, judge 0.7245775, must 0.6987855</td>
</tr>
<tr>
<td>Elessar</td>
<td>elfstone 0.8990467, especial 0.7665790, warden 0.7650582, healers 0.7560882, healer 0.7495625</td>
</tr>
<tr>
<td>Location</td>
<td>Similar Words and Similarity Scores</td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Bag End</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Location</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Bag End</td>
<td>bagshot</td>
</tr>
<tr>
<td></td>
<td>admittance</td>
</tr>
<tr>
<td></td>
<td>stocks</td>
</tr>
<tr>
<td></td>
<td>carts</td>
</tr>
<tr>
<td>Shire</td>
<td>year</td>
</tr>
<tr>
<td></td>
<td>hobbits</td>
</tr>
<tr>
<td></td>
<td>frodo</td>
</tr>
<tr>
<td></td>
<td>called</td>
</tr>
<tr>
<td>Bree</td>
<td>greenway</td>
</tr>
<tr>
<td></td>
<td>gatekeeper</td>
</tr>
<tr>
<td></td>
<td>inhabitants</td>
</tr>
<tr>
<td></td>
<td>accounting</td>
</tr>
<tr>
<td>Weathertop</td>
<td>strider</td>
</tr>
<tr>
<td></td>
<td>tinúviel</td>
</tr>
<tr>
<td></td>
<td>bolsters</td>
</tr>
<tr>
<td></td>
<td>conical</td>
</tr>
<tr>
<td>Rivendell</td>
<td>bruinen</td>
</tr>
<tr>
<td></td>
<td>glorfindel</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------</td>
</tr>
<tr>
<td></td>
<td>0.6685963</td>
</tr>
<tr>
<td>gravest</td>
<td>0.6372138</td>
</tr>
<tr>
<td></td>
<td>0.6216418</td>
</tr>
<tr>
<td>wound</td>
<td>0.6093988</td>
</tr>
<tr>
<td></td>
<td>0.6034673</td>
</tr>
</tbody>
</table>

| Moria               | anvil      | annon   | attracted | berúthiel | beth |
|                     | 0.8895135  | 0.8658382 | 0.8658382 | 0.8658382  |      |
|                     | bewilder   | buckler | churning  | cooler     | covets |
|                     | 0.8658382  | 0.8658382 | 0.8658382 | 0.8658382  |      |
|                     | creek      | daeron  | dammed   | dan        | dares |
|                     | 0.8658382  | 0.8658382 | 0.8658382 | 0.8658382  |      |
|                     | darksome   | decipher| disquietingly | doggedly | dolven |
|                     | 0.8658382  | 0.8658382 | 0.8658382 | 0.8658382  |      |

| Rohan               | war        | speak    | shall    | yet        | nay   |
|                     | 0.8411767  | 0.8235498 | 0.7890092 | 0.7817102  | 0.7777228 |
|                     | proud      | will     | gondor   | man        | tidings |
|                     | 0.7717276  | 0.7663195 | 0.7661723 | 0.764086   | 0.7612044 |
|                     | my         | is       | lord     | battle     | hour   |
|                     | 0.7584183  | 0.7531144 | 0.7520321 | 0.7514579  | 0.7502625 |
|                     | may        | son      | set      | horses     | aragorn |
|                     | 0.7485747  | 0.7442446 | 0.7431275 | 0.7379821  | 0.7351003 |

| Isengard            | nan        | ents     | curunir  | saruman    | treebeard |
|                     | 0.7730207  | 0.7329282 | 0.7296711 | 0.7290215  | 0.7251475 |
|                     | leisurely  | striding | ent      | wellinghall | isen    |
|                     | 0.6993997  | 0.6696268 | 0.6650387 | 0.6612244  | 0.6508884 |
|                     | quickbeam  | chatter  | furnace  | hollowed   | dales   |
|                     | 0.6493095  | 0.6311687 | 0.6311687 | 0.6311687  | 0.6310495 |
|                     | hom        | hasty    | aye      | bowl       | roused   |
|                     | 0.6310495  | 0.6272925 | 0.6225403 | 0.6189983  | 0.6170501 |

<p>| Minas Tirth         | boromir    | maybe    | gondor   | hour       | yet    |
|                     | 0.7760613  | 0.7162375 | 0.7103452 | 0.7100327  | 0.6857233 |
|                     | surely     | need     | anduin   | trial      | halflings |
|                     | 0.6731995  | 0.6714900 | 0.6713712 | 0.6694675  | 0.6672862 |
|                     | rauros     | may      | our      | halfling   | words   |
|                     | 0.6616034  | 0.6559257 | 0.6555772 | 0.6551451  | 0.6546402 |</p>
<table>
<thead>
<tr>
<th></th>
<th>part</th>
<th>set</th>
<th>númenor</th>
<th>me</th>
<th>save</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.6523117</td>
<td>0.6504780</td>
<td>0.6482726</td>
<td>0.6474147</td>
<td>0.6472638</td>
</tr>
</tbody>
</table>

**Gondor**

<table>
<thead>
<tr>
<th></th>
<th>yea</th>
<th>halfling</th>
<th>men</th>
<th>rohan</th>
<th>honour</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.8077411</td>
<td>0.7807475</td>
<td>0.7693074</td>
<td>0.7661723</td>
<td>0.7574521</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>proud</th>
<th>tidings</th>
<th>errand</th>
<th>son</th>
<th>minas</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.7553325</td>
<td>0.7527349</td>
<td>0.7487391</td>
<td>0.7457798</td>
<td>0.7434889</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>lords</th>
<th>lord</th>
<th>maybe</th>
<th>city</th>
<th>save</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.7389203</td>
<td>0.7379994</td>
<td>0.7369089</td>
<td>0.7337513</td>
<td>0.7336817</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>yet</th>
<th>worthy</th>
<th>anduin</th>
<th>hour</th>
<th>tirth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.7270591</td>
<td>0.7234507</td>
<td>0.7190570</td>
<td>0.7177815</td>
<td>0.7128942</td>
</tr>
</tbody>
</table>

**Mordor**

<table>
<thead>
<tr>
<th></th>
<th>evil</th>
<th>its</th>
<th>sauron</th>
<th>fear</th>
<th>gorgoroth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.7345005</td>
<td>0.7237474</td>
<td>0.7133250</td>
<td>0.7029030</td>
<td>0.7007808</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>tower</th>
<th>westward</th>
<th>power</th>
<th>towers</th>
<th>black</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.6986935</td>
<td>0.6927736</td>
<td>0.6842445</td>
<td>0.6789814</td>
<td>0.6736350</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>hope</th>
<th>doubt</th>
<th>overthrow</th>
<th>cloak</th>
<th>might</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.6710605</td>
<td>0.6629501</td>
<td>0.6624103</td>
<td>0.6546882</td>
<td>0.6508370</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>roads</th>
<th>hide</th>
<th>march</th>
<th>destroyed</th>
<th>enemy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.6507436</td>
<td>0.6496316</td>
<td>0.6437491</td>
<td>0.6423854</td>
<td>0.6402233</td>
</tr>
</tbody>
</table>
install.packages("xml2")
library(xml2)
html_loc = "~/Desktop/tolkien"
setwd(html_loc)

vocab = c()
filenames = dir()
for (i in 1:length(filenames)) {
txt = read_html(filenames[i])
full_text = xml_text(txt)
words = strsplit(full_text, split = "\W")
words = unlist(words)

words = tolower(words)
words = words[words != ""]
freqs = table(words)

terms = names(freqs)
vocab = c(vocab, terms)
}
vocab = unique(vocab)
wordcount = matrix(0, nrow = length(vocab), ncol = 1)
rownames(wordcount) = vocab

# The Hobbit

txt = read_html("hobbit.html")
nodes = xml_find_all(txt, "//\*/")
elements = xml_name(nodes)
hits = which(elements == "h2")
for (i in 1:(length(hits))) {
    start = hits[i]
    if (i < length(hits)) {
        end = hits[i+1]
    } else {
        end = length(nodes)
    }
    chapter_paragraphs = nodes[start:end]
    chapter = xml_text(chapter_paragraphs)
    ord = which(elements[start:end] != "p")
    chapter = chapter[-ord]
    words = strsplit(chapter, split = "\W")
    words = unlist(words)
words = tolower(words)
words = words[words != ""]
freqs = table(words)

terms = names(freqs)

vec = rep(0, length(vocab))
names(vec) = vocab
vec[terms] = freqs

wordcount = cbind(wordcount, vec)
}

totals = colSums(wordcount)
hits = which(totals > 10)
labels = paste("hobbit_", 1:length(hits), sep = "")
wordcount = wordcount[,hits]
colnames(wordcount) = labels

# The Fellowship of the Ring

wordcount_lor1 = matrix(0, nrow = length(vocab), ncol = 1)
rownames(wordcount_lor1) = vocab

txt = read_html("lor1.html")
nodes = xml_find_all(txt, "/\*/")
elements = xml_name(nodes)
hits = which(elements == "h3")
for (i in 1:(length(hits))) {
  start = hits[i]
  if (i < length(hits)) {
    end = hits[i+1]
  } else {
    end = length(nodes)
  }
  chapter_paragraphs = nodes[start:end]
  chapter = xml_text(chapter_paragraphs)
  ord = which(elements[start:end] != "p")
  chapter = chapter[-ord]
  words = strsplit(chapter, split = "\\W")
  words = unlist(words)
  words = tolower(words)
  words = words[words != ""]
freqs = table(words)

terms = names(freqs)

vec = rep(0, length(vocab))
names(vec) = vocab
vec[terms] = freqs

wordcount_lor1 = cbind(wordcount_lor1, vec)
}

totals = colSums(wordcount_lor1)

hits = which(totals > 50)

labels = paste("for", 1:length(hits), sep = "")

wordcount_lor1 = wordcount_lor1[, hits]
colnames(wordcount_lor1) = labels

# The Two Towers

wordcount_lor2 = matrix(0, nrow = length(vocab), ncol = 1)
rownames(wordcount_lor2) = vocab

txt = read_html("lor2.html")
nodes = xml_find_all(txt, "/\/*")

elements = xml_name(nodes)
hits = which(elements == "h3")
for (i in 1:(length(hits))) {
  start = hits[i]
  if (i < length(hits)) {
    end = hits[i+1]
  } else {
    end = length(nodes)
  }
  chapter_paragraphs = nodes[start:end]
  chapter = xml_text(chapter_paragraphs)
  ord = which(elements[start:end] != "p")
  chapter = chapter[-ord]
  words = strsplit(chapter, split = "\W")
  words = unlist(words)
  words = tolower(words)
  words = words[words != ""]
  freqs = table(words)

  terms = names(freqs)
vec = rep(0, length(vocab))
names(vec) = vocab
vec[terms] = freqs

wordcount_lor2 = cbind(wordcount_lor2, vec)
}

totals = colSums(wordcount_lor2)
hits = which(totals > 50)
labels = paste("tt_", 1:length(hits), sep = "")
wordcount_lor2 = wordcount_lor2[, hits]
colnames(wordcount_lor2) = labels

# The Return of the King

wordcount_lor3 = matrix(0, nrow = length(vocab), ncol = 1)
rownames(wordcount) = vocab

txt = read_html("lor3.html")
nodes = xml_find_all(txt, "/\/*")
elements = xml_name(nodes)
hits = which(elements == "h3")
for (i in 1:(length(hits))) {
  start = hits[i]
  if (i < length(hits)) {
    end = hits[i+1]
  } else {
    end = length(nodes)
  }
  chapter_paragraphs = nodes[start:end]
  chapter = xml_text(chapter_paragraphs)
  ord = which(elements[start:end] != "p")
  chapter = chapter[-ord]
  words = strsplit(chapter, split = "\W")
  words = unlist(words)

  # Mark proper nouns or other text processing

  words = tolower(words)
  words = words[words != ""]
  freqs = table(words)
  terms = names(freqs)
vec = rep(0, length(vocab))
names(vec) = vocab
vec[terms] = freqs

wordcount_lor3 = cbind(wordcount_lor3, vec)
}
totals = colSums(wordcount_lor3)
hits = which(totals > 50)
labels = paste("rotk ", 1:length(hits), sep = "")
wordcount_lor3 = wordcount_lor3[, hits]
colnames(wordcount_lor3) = labels

word_mat = cbind(wordcount, wordcount_lor1, wordcount_lor2, wordcount_lor3)

locs = read.csv("~/Desktop/tolkien/tolkienlocations.csv", stringsAsFactors = F)
places = locs[,1]
places = unique(places)

setwd("~/Desktop/tolkien")
txt = read_html("hobbit.html")
nodes = xml_find_all(txt, "/\/*")
elements = xml_name(nodes)
hits = which(elements == "h2")
for (i in 1:(length(hits))) {
  start = hits[i]
  if (i < length(hits)) {
    end = hits[i+1]
  } else {
    end = length(nodes)
  }
  chapter_paragraphs = nodes[start:end]
  chapter = xml_text(chapter_paragraphs)
  ord = which(elements[start:end] != "p")
  chapter = chapter[-ord]
  chapter = paste(chapter, collapse = " ")
  freqs = rep(0, length(places))  # Creating a blank vector
  for (j in 1:length(places)) {
    results = gregexpr(places[j], chapter)[[1]]  # This finds all
    the places in the string where the place-string appears
    if (results == -1) {
      freq = 0
    } else {  
      freq = length(results)
      places[j] = places[j]  
      chapter = gsub(places[j], " ", chapter)  
    }
    freqs[j] = freq
  }
  freqs = do.call(rbind, freqs)  
  wordcount_lor3 = cbind(wordcount, wordcount_lor1, wordcount_lor2, wordcount_lor3)
}

wordcount_lor3 = wordcount_lor3[, hits]
colnames(wordcount_lor3) = labels

freqs = rep(0, length(places))  # Creating a blank vector
for (j in 1:length(places)) {
  results = gregexpr(places[j], chapter)[[1]]  # This finds all
  the places in the string where the place-string appears
  if (results == -1) {
    freq = 0
  } else {  
    freq = length(results)
    places[j] = places[j]  
    chapter = gsub(places[j], " ", chapter)  
  }
  freqs[j] = freq
}
wordcount_lor3 = cbind(wordcount, wordcount_lor1, wordcount_lor2, wordcount_lor3)

freq = length(results)
}
freqs[j] = freq
}
if(i == 1) {
   places_mat = matrix(freqs, length(freqs), 1)
} else {
   places_mat = cbind(places_mat, freqs)
}
}
rownames(places_mat) = places

setwd("~/Desktop/tolkien")
txt = read_html("lor1.html")
nodes = xml_find_all(txt, "/\*\*/")
elements = xml_name(nodes)
hits = which(elements == "h3")
for (i in 1:(length(hits))) {
   start = hits[i]
   if (i < length(hits)) {
      end = hits[i+1]
   } else {
      end = length(nodes)
   }
   chapter_paragraphs = nodes[start:end]
   chapter = xml_text(chapter_paragraphs)
   ord = which(elements[start:end] != "p")
   chapter = chapter[-ord]
   chapter = paste(chapter, collapse = " ")
   freqs = rep(0, length(places))
   for (j in 1:length(places)) {
      results = gregexpr(places[j], chapter)[[1]]
      if (results == -1) {
         freq = 0
      } else {
         freq = length(results)
      }
      freqs[j] = freq
   }
   if(i == 1) {
      places_mat_1 = matrix(freqs, length(freqs), 1)
   } else {
      places_mat_1 = cbind(places_mat_1, freqs)
   }
}
rownames(places_mat_1) = places

setwd("~/Desktop/tolkien")
txt = read_html("lor2.html")
nodes = xml_find_all(txt, "/\/*")
elements = xml_name(nodes)
hits = which(elements == "h3")
for (i in 1:(length(hits))) {
    start = hits[i]
    if (i < length(hits)) {
        end = hits[i+1]
    } else {
        end = length(nodes)
    }
    chapter_paragraphs = nodes[start:end]
    chapter = xml_text(chapter_paragraphs)
    ord = which(elements[start:end] != "p")
    chapter = chapter[-ord]
    chapter = paste(chapter, collapse = " ")
    freqs = rep(0, length(places))
    for (j in 1:length(places)) {
        results = gregexpr(places[j], chapter)[[1]]
        if (results == -1) {
            freq = 0
        } else {
            freq = length(results)
        }
        freqs[j] = freq
    }
    if(i == 1) {
        places_mat_2 = matrix(freqs, length(freqs), 1)
    } else {
        places_mat_2 = cbind(places_mat_2, freqs)
    }
}
rownames(places_mat_2) = places

setwd("~/Desktop/tolkien")
txt = read_html("lor3.html")
nodes = xml_find_all(txt, "/\/*")
elements = xml_name(nodes)
hits = which(elements == "h3")
for (i in 1:(length(hits))) {
    start = hits[i]
    if (i < length(hits)) {

end = hits[i+1]
else {
    end = length(nodes)
}

chapter_paragraphs = nodes[start:end]
chapter = xml_text(chapter_paragraphs)
ord = which(elements[start:end] != "p")
chapter = chapter[-ord]
chapter = paste(chapter, collapse = " ")
freqs = rep(0, length(places))
for (j in 1:length(places)) {
    results = gregexpr(places[j], chapter)[[1]]
    if (results == -1) {
        freq = 0
    } else {
        freq = length(results)
    }
    freqs[j] = freq
}
if(i == 1) {
    places_mat_3 = matrix(freqs, length(freqs), 1)
} else {
    places_mat_3 = cbind(places_mat_3, freqs)
}
rownames(places_mat_3) = places

places_mat = cbind(places_mat, places_mat_1, places_mat_2,
places_mat_3)

totals = colSums(places_mat)
hits = which(totals > 3)
labels = paste(1:length(hits))
places_mat = places_mat[, hits]
colnames(places_mat) = labels
colnames(places_mat)[1] = paste("h_notes", sep = "")
colnames(places_mat)[2:20] = paste("hobbit_", 1:19, sep = "")
colnames(places_mat)[21:25] = paste("p_", 1:5, sep = "")
colnames(places_mat)[26:47] = paste("fotr_", 1:22, sep = "")
colnames(places_mat)[48:67] = paste("ttt_", 1:20, sep = "")
colnames(places_mat)[68:86] = paste("rotk_", 1:19, sep = "")

colnames(word_mat)[1] = paste("h_notes", sep = "")
colnames(word_mat)[2:20] = paste("hobbit_", 1:19, sep = "")
colnames(word_mat)[21:25] = paste("p_", 1:5, sep = "")
colnames(word_mat)[26:47] = paste("fotr_", 1:22, sep = "")
colnames(word_mat)[48:67] = paste("ttt_", 1:20, sep = "")
colnames(word_mat)[68:86] = paste("rotk_", 1:19, sep = "")

save(word_mat, file = "word_mat.rda")
save(places_mat, file = "places_mat.rda")

setwd("~/Desktop/Thesis Stuff")
load("places_mat.rda")
load("word_mat.rda")
chaps = read.csv("chaptersettings.csv")
latlon = read.csv("latlongs.csv")
locs = read.csv("tolkienlocations.csv")

plot(latlon$Lon, latlon$Lat, cex = 0)
text(latlon$Lon, latlon$Lat, labels = latlon$Location)

# Correlating with textual data
places = latlon$Location

# Calculating centroid for a given chapter
vec = places_mat[places, 1]
vec = vec / sum(vec)
mean_lon = sum(latlon$Lon * vec)
mean_lat = sum(latlon$Lat * vec)

# Compiling place-keyword matrix
ord = which(rowSums(places_mat) == 0)
places_mat = places_mat[-ord,]
ord = which(rowSums(word_mat) == 0)
word_mat = word_mat[-ord,]
pw = places_mat %*% t(word_mat)

# PMI function
ppmi = function(mat) {
  total = sum(mat, na.rm = T)
  pcoll = apply(mat, 2, function(x) sum(x, na.rm = T) / total)
  prow = apply(mat, 1, function(x) sum(x, na.rm = T) / total)
  pmat = as.matrix(prow) %*% t(pcoll)
  mat = mat / total
  mat = mat / pmat
  mat = apply(mat, 1, log)
mat = t(mat)
mat[mat < 0] = 0
mat[is.na(mat)] = 0
mat[is.infinite(mat)] = 0
return(mat)
}

pw_ppmi = ppmi(pw)

# Cosine similarity

semantic_similarity = function(mat, term, output = "top") {
cos_sim = function(x, y) {
x %*% y/(sqrt(x %*% x) * sqrt(y %*% y))
}
if (length(term) == 1) {
if (term %in% rownames(mat) == F) {
  stop("The term is not among the rownames of your matrix. This can happen
  for several reasons: 1) the term is just not included in your
data, 2) the matrix does not have rownames, or 3) you are trying
to calculate over the columns, in which case you need to transpose
  the matrix with t(mat).")
}
vec = mat[,term]
} else {
  vec = term
}
results = unlist(apply(mat, 1, cos_sim, vec))
if (output == "top") {
  results = sort(results, decreasing = T)[1:12]
}
return(results)
}

##### Work with lonlat data ######
geo_coords = latlon[,c(1,3,4)]
colnames(geo_coords) = c("NAME", "LAT", "LON")

footprint = function(mat, term, coords) {
  vec = mat[,term]
  vec[is.na(vec)] = 0
vec = vec[vec > 0]
vec = vec[names(vec) %in% coords[,"NAME"]]

hits = which(coords[,"NAME"] %in% names(vec))
#vec = vec[coords[hits,"NAME"]]
ord = which(is.na(vec))
if (length(ord) > 0) {
    hits = hits[-ord]
    vec = vec[-ord]
}

mean_lon = sum(coords[hits,"LON"] * vec[vec > 0]) / sum(vec)
mean_lat = sum(coords[hits,"LAT"] * vec[vec > 0]) / sum(vec)
euc_dist = function(x, y) { sum(sqrt(x - y)^2) }
distances = c()
for (i in hits) {
    lat = coords[i,"LAT"]
    lon = coords[i,"LON"]
    d = euc_dist(x = c(mean_lon, mean_lat), y = c(lon, lat))
    distances = c(distances,d)
}

mean_dist = sum(distances * vec, na.rm = T) / sum(vec, na.rm = T)
sd_dist = sd(distances, na.rm = T)
results = list()
results$n = length(vec)
results$freq = sum(vec, na.rm = T)
results$sd = sd(vec, na.rm = T)
results$lon = mean_lon
results$lat = mean_lat
results$radius = mean_dist
results$stdev = sd_dist
return(results)
}

# Gathering geosemantic data for each word

geodata = matrix(0,nrow(word_mat), 7)
rownames(geodata) = rownames(word_mat)
for (i in 1:nrow(geodata)) {
    print(i)
    res = footprint(mat = pw, term = rownames(word_mat)[i], coords = geo_coords)
    geodata[i,] = unlist(res)
}
colnames(geodata) = names(res)
# Mapping the data

```r
ord = order(geodata[, "freq"], decreasing = T)[1:1000]
plot(geodata[ord, "lon"], geodata[ord, "lat"], cex = 0)
text(geodata[ord, "lon"], geodata[ord, "lat"], labels = rownames(geodata[ord, ]))

fit = lm(geodata[ord, "lat"] ~ geodata[ord, "lon"])
abline(a = fit$coefficients[1], b = fit$coefficients[2], lwd = 4, col = "dark green")

ord = order(geodata[, "freq"], decreasing = T)[1:500]
plot(geodata[ord, "lon"], geodata[ord, "radius"], cex = 0)
text(geodata[ord, "lon"], geodata[ord, "radius"], labels = rownames(geodata[ord, ]))

ord = order(geodata[, "freq"], decreasing = T)[1:500]
plot(geodata[ord, "radius"], geodata[ord, "lat"], cex = 0)
text(geodata[ord, "radius"], geodata[ord, "lat"], labels = rownames(geodata[ord, ]))
```