KU LEUVEN

FACULTY OF SCIENCE

From Text to Insight:

Computational interpretations of 'Invisible Cities' and its public reflections

Peiyang Huo

Supervisor: Prof. Katrien Verbert
Mentor: Ivania Nadine Donoso Guzmán

Thesis presented in
fulfillment of the requirements
for the degree of Master of Science
in Digital Humanities

Academic year 2022-2023

Preface

This thesis endeavors to bridge the realms of text analytics and visualization, utilizing Italo Calvino's "Invisible Cities" and its public commentary as a primary data source. The main objectives are twofold: firstly, to employ advanced topic modeling techniques to decode abstract literary texts and public reviews, and secondly, to transform these analytical insights into visually compelling narratives. By synthesizing the tools of natural language processing with visualization methodologies, the work aspires to offer a holistic understanding of Calvino's masterpiece, while also showcasing the potential of modern data analysis and presentation techniques.

First and foremost, I would like to express my profound gratitude to my promotor, Katrien Verbert, for not only serving as the promotor of this thesis but also for her invaluable guidance in the courses on data visualization and web information systems. Her teachings and insights have laid a solid foundation for my research journey, and I am immensely grateful for the knowledge and wisdom she has imparted.

Equally, I owe a debt of gratitude to my daily advisor, Ivania Nadine Donoso Guzmán. Her unwavering support has been instrumental at every step of this thesis, from the selection of the topic and the adjustments in direction to the technical strategies and feedback on the final results. I am especially thankful for her meticulous reviews, which have significantly enhanced the quality of my work.

I would like to extend my heartfelt thanks to my group members from the Digital Humanities program: Dawn Zhuang, Ching-Han Kuo, Lee, Shirin Izadpanah, and Agni Vourtsi. Working alongside them on our assignments has been an enriching experience, marked by mutual learning and continuous skill refinement. Their camaraderie and collaboration have been a source of both inspiration and motivation.

I extend my heartfelt gratitude to my family. Their unwavering support has been a cornerstone throughout this journey. To my parents and siblings, thank you for always believing in me and for being my constant source of inspiration and strength.

List of abbreviations and list of symbols

BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag of Words
c-TF-IDF	TF-IDF formula adopted for multiple classes by joining all documents per class
COVID-19	Coronavirus disease 2019
GSDMM	Gibbs Sampling Dirichlet Mixture Model
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
NLP	Natural language processing
NLTK	Natural Language Toolkit
Oulipo	Ouvroir de littérature potentielle (Workshop for Potential Literature)
SVD	Singular Value Decomposition
TF-IDF	Term Frequency - Inverse Document Frequency
UMAP	Uniform Manifold Approximation and Projection
URL	Uniform Resource Locator
VSM	Vector Space Model

Summary

The digital age has bestowed upon us an abundance of textual data, stemming from a myriad of sources, both digital-native and those transitioning from the physical world through the process of digitization. This thesis embarked on a mission to navigate the complexities of this vast textual landscape, aiming to extract meaningful insights and present them in a visually compelling manner.

In this thesis, a brief introduction to the research background and objectives initiates the discussion in the first chapter. The subsequent second chapter delves into the domains of Text Analytics and the pertinent field of text visualization. In the Text Analytics section, this thesis meticulously traces the evolution of this domain, focusing primarily on the topic modeling technique, a popular method to extract topics from texts. And critically analyzes the pros and cons of various methodologies within this domain. For text visualization, we investigate different strategies tailored according to the target audience, be it scholars or the general public. Besides, this chapter conducts an in-depth review of the literature related to Italo Calvino and his seminal work, "Invisible Cities."

In the third chapter, the rationale for selecting Calvino and "Invisible Cities" as the subject of our study is articulated, and describes the data set used in detail.

The fourth chapter, the crux of my research, elaborates on the holistic strategies deployed for text analysis and visualization. Using topic modeling, I interpret "Invisible Cities" from an urban planning perspective, distilling six core topics from the text. These topics and their interrelations are eventually presented to the readers via an interactive interface, wherein the shared themes across stories are depicted using edge bundling. Additionally, I analyze public reviews of the book, identifying that readers mainly approach their comments from five themes: genres, structure, recommend, personal reflection, and quotation. Addressing these five themes, I further delve into topic modeling analysis, uncovering nuanced insights. Beyond this, I stumble upon intriguing findings, such as other authors deemed stylistically similar to Calvino by readers and the most frequently quoted sections from the book. The visualization for this part predominantly employs the Sankey Diagram, complemented with network graphs, bar charts, and histograms to represent our findings.

In the fifth chapter, the discourse centers on the design and implementation of a dedicated website, equipped with a suite of basic interactive features, to effectively showcase the insights gleaned from our text analytics. Keeping in mind that the intended audience for the website is predominantly readers of "Invisible Cities" and fans of Calvino, I enrich the content with an introduction to Calvino and a comprehensive life map, aiming to weave a compelling narrative experience.

Table of contents

Chapter 1

Introduction

The rise of the internet and digital technologies has consistently decreased the costs of publishing, accessing, and storing textual data. Not only are sources of texture data diversifying with digitally native content, but a growing amount of physical text information is also being digitized. In this age of data proliferation, the extraction of valuable insights from a vast sea of data has become paramount. This surge in importance is further fueled by advancements in the field of natural language processing (NLP), equipping computers with the capability to process human language and mine text data effectively.

Simultaneously, the deluge of data introduces challenges in comprehending and analyzing abstract datasets. Visualization emerges as a solution, capitalizing on the human visual system's proficiency in interpreting images and graphics, thereby translating abstract data into an understandable format. In the realm of visualization, effective communication of information stands paramount, making the target audience a critical factor that shapes the visualization's design and purpose.

These two domains share a symbiotic relationship, especially when addressing the general public. Visualization serves as a medium to convey the results of text analytics visually, enhancing comprehension and engagement.

Combining the two fields, this thesis selects Italo Calvino's "Invisible Cities" and its public reviews as the textual data for analytics and visualization. The findings are presented to the audience through a website. The text analytics segment delves into the application of topic modeling on abstract texts and public reviews, along with other text analytics methods employed. The primary objective of visualization is to visually represent the analysis results, which, when paired with textual content, form a cohesive narrative.

Chapter 2

Literature review

2.1 Text analytics

2.1.1 Evolution of text analytics

Text analytics, also referred to as text mining, operates on the principle that researchers can process text automatically without the necessity of reading it (Iezzi & Celardo, 2020). Since the 1940s, this arena has progressively integrated advancements from fields like computer science, linguistics, statistics, and mathematics. Post-1990s, with the surge in data availability, techniques such as web mining emerged, aiming to extract valuable information from the vast expanses of the World-Wide Web. As we transitioned into the new millennium, the field further evolved, introducing advanced processing methodologies like topic modeling (Iezzi & Celardo, 2020) (Figure1).

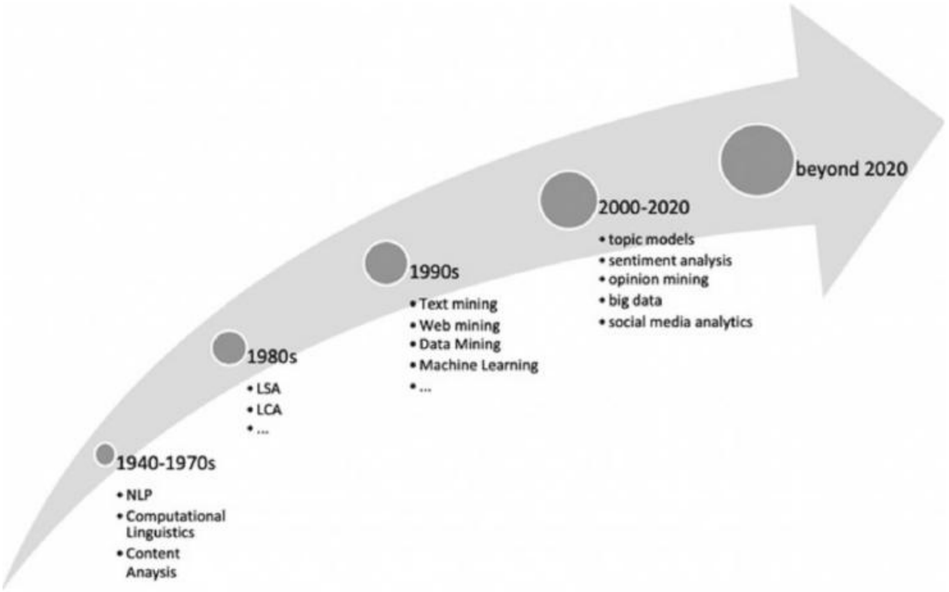


Figure 1: Text analysis timeline (Iezzi & Celardo, 2020)

Iezzi & Celardo (2020) summarize that text analytics generally involves six stages: purpose, data collection, preprocessing, data matrix creation, text analysis, and interpretation. As for analytics tasks, Liu et al. (2018) reviewed papers in the fields of visualization and text analytics, then summarized nine first-level text analytics tasks. Ranked by the number of papers, these tasks are information retrieval, cluster/topic analysis, natural language processing (NLP), classification, exploratory analysis, trend analysis, network analysis, predictive analysis, and outlier analysis. In terms of applications, text analytics is widely used in research, business, healthcare, the internet, and other industries. For example, Oyebode et al. (2021) identified 34 negative themes related to discussions about COVID-19 on social media and offered corresponding recommendations.

2.1.2 Distant reading: an alternative approach to do literary studies

The field of literary criticism traditionally employs a method known as 'close reading'. This approach, established in the early 19th century, necessitates a thorough understanding of semantic connotations by identifying the central theme(s) within a text. Close reading extends beyond the text itself, incorporating analysis of word choice, structure, stylistic elements, argumentation patterns, and the interplay between the author and related events (Jasinski, 2001).

In 2005, Franco Moretti introduced the concept of "distant reading". This method, in contrast to close reading, aims to create an abstract view by transitioning from the examination of textual content to the visualization of global features within a single text or across multiple texts. Concurrently, Jockers (2013) proposed the term 'macro reading', inspired by macroeconomics. This approach focuses on applying statistical methods to perform quantitative analytics on literary texts.

At the application level, Moretti (2005) proposed the term "distant reading" with three representative cases, each exemplifying a different interpretive approach. These approaches include analyzing the variation of fictional themes using graphs, mapping the geographical characteristics of novels, and categorizing detective novels using tree diagrams. Of course, distant reading merely provides a new perspective and cannot replace close reading. Coles and Lein (2013) suggest that distant reading can guide readers in filtering the text for research, thereby complementing traditional close reading techniques.

2.1.3 Topic modeling

Topic modeling is a potent tool for extracting latent semantic information concealed within a corpus (a collection of documents), and it is also a crucial category of text analytics. However, when confronting different specific tasks and data types, it has evolved into various methods.

Traditional Methods and their variations:

The traditional method is based on the Bag of Words (BoW) model. In this model, every document is depicted as a vector, where each dimension stands for a word, and its value indicates the word's frequency within the document (Harris, 1954).This method disregards the order of words in the document and necessitates pre-processing of data by the user to remove noise that might obscure latent topics.

The Vector Space Model (VSM), as the inaugural model developed on a document-term matrix, is a non-probabilistic method that shines in information retrieval (Salton, Wong & Yang, 1975). The Latent Dirichlet Allocation (LDA) method is another popular conventional approach. LDA represents a corpus as a probabilistic mixture of latent topics, assuming that each document encompasses multiple topics and that each word in the document pertains to a topic (Blei, Ng & Jordan, 2003).

However, these methods do not perform well with short texts because their effectiveness diminishes due to the generation of high-dimensional, sparse document vectors. Although there are methods like Latent Semantic Analysis (LSA) or Latent Semantic Indexing (LSI), which utilize Singular Value Decomposition (SVD) to represent text in a low-dimensional vector space, they require significant computing power (Yamunathangam et al., 2021).

Building on traditional unsupervised methods, Jagarlamudi, Iii & Udupa (2012) described a semi-supervised method called *sets of seed words* intended to guideLDA further in a specific direction based on prior knowledge. This method achieves the desired results by setting a seed term, representing prior knowledge, and allowing the model to converge in that direction.

Clustering Based Approaches:

Recognizing the limitations of conventional approaches with short texts, researchers have proposed clustering strategies for topic inference. This strategy assumes each document pertains to a single topic, not a mixture. For instance, the Gibbs Sampling Dirichlet Mixture Model (GSDMM) (Yang, Huang & Cai, 2019).

Deep Learning Methods:

Deep learning has also been integrated into topic modeling, offering another method for text representation. Word2Vec (Mikolov et al., 2013), a well-known 'word embedding' algorithm, proposes that words with similar contexts tend to have related meanings. In 2014, Doc2Vec was introduced as an extension of Word2Vec (Le and Mikolov, 2014), generating vector representations for not only words but also paragraphs or documents, thereby incorporating word sequences.

In practice, these deep learning tools often combine with clustering strategies. For instance, Angelov's Top2Vec (2020), based on Doc2Vec, aims to address weaknesses in LDA and LSA. Top2Vec generates embedding vectors for documents and words, representing semantic information in high-dimensional space. It employs the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm for topic clustering. Unlike conventional approaches, Top2Vec does not require text preprocessing and can determine the number of topics based on data distribution.

Inspired by Top2Vec, Grootendorst (2022) proposed BERTopic, which leverages the BERT (Bidirectional Encoder Representations from Transformers) model, widely used in other NLP tasks. BERTopic uses BERT for document embedding, then combines UMAP (Uniform Manifold Approximation and Projection) and HDBSCAN for clustering. It also introduces a variant of TF-IDF (Term Frequency - Inverse Document Frequency), called c-TF-IDF, which considers a single cluster as one document and applies c-TF-IDF to reveal significant words within the cluster.

Both methods allow users to explore results post-training, such as semantic search and topic similarity calculation. By combining these techniques, more opportunities arise for users to perform advanced analyses.

Online topic modelling:

In addition to static data, emerging methods of topic modeling for dynamic data are also constantly being developed, playing a crucial role in the analysis of social networks. These techniques are typically based on clustering strategies. For example, Rakib, Zeh, and Milios (2021) proposed EStream, which is a method for short text stream clustering. This approach dynamically assigns texts to suitable clusters.

2.2 Data visualization for text analytics

2.2.1 Text visualization

While text analytics is receiving increasing attention, visualization as a means of conveying its results has been researched and applied since the 1990s (Cao and Cui, 2016). Cao and Cui (2016) also categorized text visualization task into five groups: (1) visualizing document similarity, (2) content revelation, (3) visualizing text sentiment, (4) exploration of document corpora, and (5) domain-specific rich corpus data.

In terms of visualization approach, Liu et al. (2018) summarized that "typographic visualizations", "chart visualizations," and "graph visualizations" are the most common primary concepts. Upon a more detailed study of chart visualizations, they found that some simple and effective techniques are becoming increasingly popular, such as "line charts," "scatterplots," and "tables." Additionally, researchers tend to focus on developing more advanced learning methods rather than more complex visualization techniques (Liu et al., 2018).

In the literary field both distant reading and close reading have developed various visualization tools to support scholars. Jänicke and et al. (2015) summarized four approaches for close reading: color, font size, glyphs and connections; and seven approaches for distant reading: structure, heat maps, tag clouds, maps, timeline, graphs, and miscalls.

2.2.2 Narrative visualization

Compared with general data visualization which focuses on transforming abstract data to the audience, narrative visualization focuses on the public. It consists of a series of story pieces with well-designed sequences that serve the story presentation (Lee et al., 2015).

A rising trend in this context is digital storytelling, as seen in the multimedia narratives of major online journalism outlets like The New York Times (Lee et al., 2015). These stories, a fusion of various media and infographics, provide a richer, more engaging format known as "storytelling articles" (Kosara & Mackinlay, 2013). Rather than focusing on the latest news, these narratives dig deeper, shedding light on intricate, data-driven issues. Two primary factors drive this approach: the high costs involved and the transformative power of data visualization, which turns complex information into visually engaging and easily digestible content, ultimately enhancing cognition (Ware, 2019).

Infographics serve as a key tool in the realm of online journalism, transforming intricate data into understandable narratives by blending graphics, text, and images. Seyser and Zeiller's study (2018) observed a preference for certain visualization techniques in Rolling Narrative Articles, namely line charts, Sankey diagrams, timelines, pie charts, and bar charts. While some interactive elements were included, they mostly provided basic functionalities, like a play button.

However, infographics are more than just data translators. They can enhance the narrative structure, be adapted to support the story, and even incorporate principles from Gestalt theory to boost visual appeal (Seyser & Zeiller, 2018). Yet, the success of an infographic depends heavily on its meticulous application and integration within the narrative structure. Otherwise, the intended effect can backfire, confusing the audience rather than enlightening them. Hullman et al. (2013), through a qualitative analysis, pointed out that sequencing choices in narrative visualizations can impact the understanding and memory of the reader. They suggested that creators consider the readers' prior knowledge and employ a variety of transition types and sequencing strategies.

2.3 Close reading about Italo Calvino and *Invisible Cities*

Italo Calvino, an Italian writer and journalist born in 1923, is regarded as one of the most important and innovative writers of the 20th century. He was nominated for the Nobel Prize in Literature in 1985 but passed away the same year. His writing style, characterized by playful and imaginative narratives, often blends elements of fantasy and reality. His later works frequently meld elements of science fiction and postmodernism, with "Invisible Cities" serving as a prime example (Licata, 2021). Ricci (2001) characterizes Calvino's narratives as "image-centric," meaning they describe real or imagined visual representations in literary form.

Calvino's works underwent several transformations throughout his career, with "Invisible Cities", completed in France, being deeply influenced by 'Oulipo'. Oulipo, a French literary group founded by writers and mathematicians, is an abbreviation for 'Ouvroir de littérature potentielle', translating to 'Workshop of Potential Literature' in English (Botta, 1997). The group is known for creating new possibilities for literary expression through constraints based on mathematical or linguistic principles. As a member of this group, Calvino wrote three works heavily influenced by its principles: "If on a winter's night a traveler," "Invisible Cities", and "The Castle of Crossed Destinies."

From a mathematical perspective, Marello (1986) suggests that the book constructed a parallelogrammatic mathematical structure within a symmetric narrative framework. Broadly, the narrative is divided into three parts: Build, Maintain, and Erase, symbolizing the past, present, and future respectively (Figure 2, above). In terms of the narrative approach, as old city themes conclude, new ones are introduced, thus creating a cycle (Figure 2, under). This dual structure of narrative and sequence led Breiner (1988) to comment that the book seems to hesitate on some threshold between modernism and postmodernism.

In terms of content, Case and Gaggiotti (2016) decipher the meaning of the metaphor of the city that cities serve as metaphors for human society but embodying two contrasting expressions. On one hand, cities can be simplified into a series of objectives and functions, making them seem lifeless. On the other hand, cities offer a wealth of imagination, providing a foundation for the ethics of various human organizations. Similarly, Modena (2011) analyzes this book and the author's life experiences from the perspective of architecture and urban planning, pointing out that Calvino's work as an editor exposed him to a wide range of related literature, including Kevin Lynch's "Image of the City". This book proposes the method of remapping cities based on subjective feelings, which aligns with Calvino's image-centric writing style.

Furthermore, this book reflects Calvino's interdisciplinary thinking and his contributions to urban reform. The urgent need for city and social reconstruction is predicated on a humanistic investigation of the city itself (Modena, 2011).

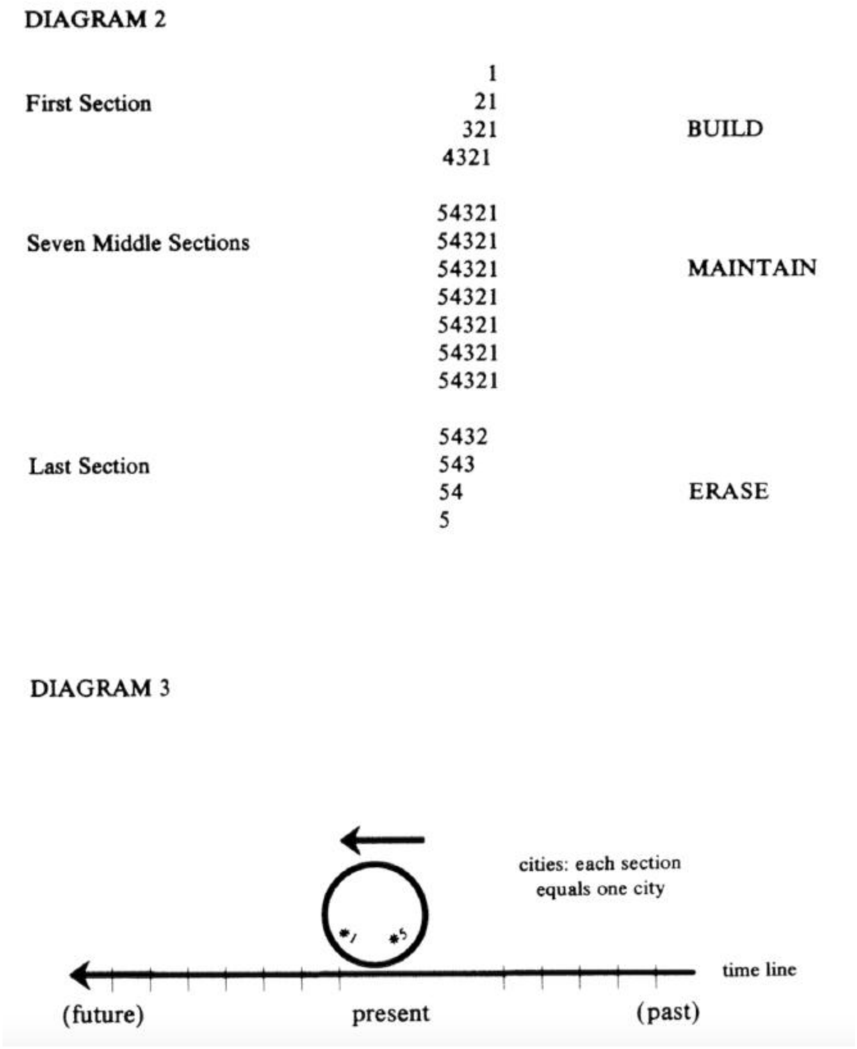


DIAGRAM 2

	1	
First Section	21	
	321	BUILD
	4321	
	54321	
Seven Middle Sections	54321	
	54321	MAINTAIN
	54321	
	54321	
	54321	
	54321	
	5432	
Last Section	543	
	54	ERASE
	5	

DIAGRAM 3

Figure 2: Diagram summarized by Marello (1986)

Chapter 3

Methodology

3.1 Selection of the Author

Calvino's influence extends beyond Literature and permeates the Built Environment, particularly evident in his celebrated work, "Invisible Cities" that first published in 1972. This book continues to be a cornerstone in the disciplines of architecture and urban planning. Moreover, his insightful parables consistently spark discussions on social media. Such cross-disciplinary impact is not unique to Calvino; works like "The Death and Life of Great American Cities" by journalist Jane Jacobs (1961) and "Arrival City" by journalist Doug Saunders (2010) have similarly resonated in the urban planning filed. Their keen insights enable them to convey their understanding to the public. However, compared to these straightforward works, "Invisible Cities" conceals its insights within complex fables. Some readers find Calvino's style challenging, as it necessitates time and attention to immerse oneself in the intricate images he constructs.

Thanks to advancements in text analytics and deep learning, we now have alternative approaches to literary studies. Concurrently, social media and other internet applications provide a rich source of data about discussions and shared thoughts regarding this book. These developments offer a promising avenue for interpretation, merging literary analysis with public reviews. Therefore, an attempt will be made to analyze this abstract literary work, which has mathematical structural characteristics, along with its reviews, through the lens of text mining.

3.2 Research question

This thesis aims to amalgamate text analytics and data visualization techniques, employing computational interpretation on Italo Calvino's Invisible Cities and its public reviews, and then communicating the results to the public through multimedia. In terms of text analytics, the focus will primarily be on the application of topic modeling techniques to guide the interpretation of abstract literary works and to mine more information from their reviews. Apart from topic modeling, this thesis will also employ other text analytics tools to supplement the analysis. Ultimately, the analysis results will be appropriately visualized and assembled into narratives for presentation on a website.

3.3 Data

The data processing and analysis for this study were conducted using Python and Tableau.

The data will be analyzed are the book of Invisible Cities what is the English version translated by William Weaver in 1974 and public reviews from goodreads. This book is a fictionalized dialogue between two historical figures, Marco Polo and Kubla Khan. Marco Polo, as an explorer, sharing Kubla Khan about fifty-five fictional cities outside of Kubla Khan's empire. All of them are short and abstract, with an average length of 331 words, and the longest of which is only 763 words. This book consists of 9 chapters, except the beginning and the end which contain 10 stories, the remaining 7 chapters contain 5 stories each. Each chapter begins and ends with a conversation between these people, which can be thought of as a supplement to the comprehension of the book. Interestingly, the nine chapters are not actually named, but instead the authors have come up with eleven themes, namely, 'Cities & Memory', 'Cities & Desire', ' Cities & Signs', 'Thin Cities', 'Trading Cities', 'Cities & Eyes ', 'Cities & Names', 'Cities & the Dead', 'Cities & the Sky ', 'Continuous Cities' and 'Hidden Cities'. Each theme has 5 narratives. The analytics of the book is focused on these 55 narratives.

The dataset collected from goodreads, includes reviews, ratings, and review timestamps. These reviews are not specific to any particular version of the book, but pertain to the book itself, and are not restricted to any particular language. The dataset contains a total of 5,899 reviews from 2007 to February 20th, 2023. As all the data is collected from a well-regulated platform, these reviews are presumed to adhere to the platform's code of conduct, thus considered high quality. Preliminary processing and basic statistical analysis revealed that the dataset, includes reviews in 45 languages as detected by Google's 'langdetect' library in Python. The majority of reviews are in English (4,120), with other languages contributing over 100 reviews including Italian, Spanish, Portuguese, Turkish, and Arabic ((Figure 3). The average review length is 95 words, with 716 reviews exceeding 200 words and the longest review stretching to 3,284 words.

Given the advanced state of Natural Language Processing (NLP) for English and considering the substantial English-speaking fan base of the book, this study will utilize the English version of the book. The analysis of reviews will also focus primarily on those written in English.

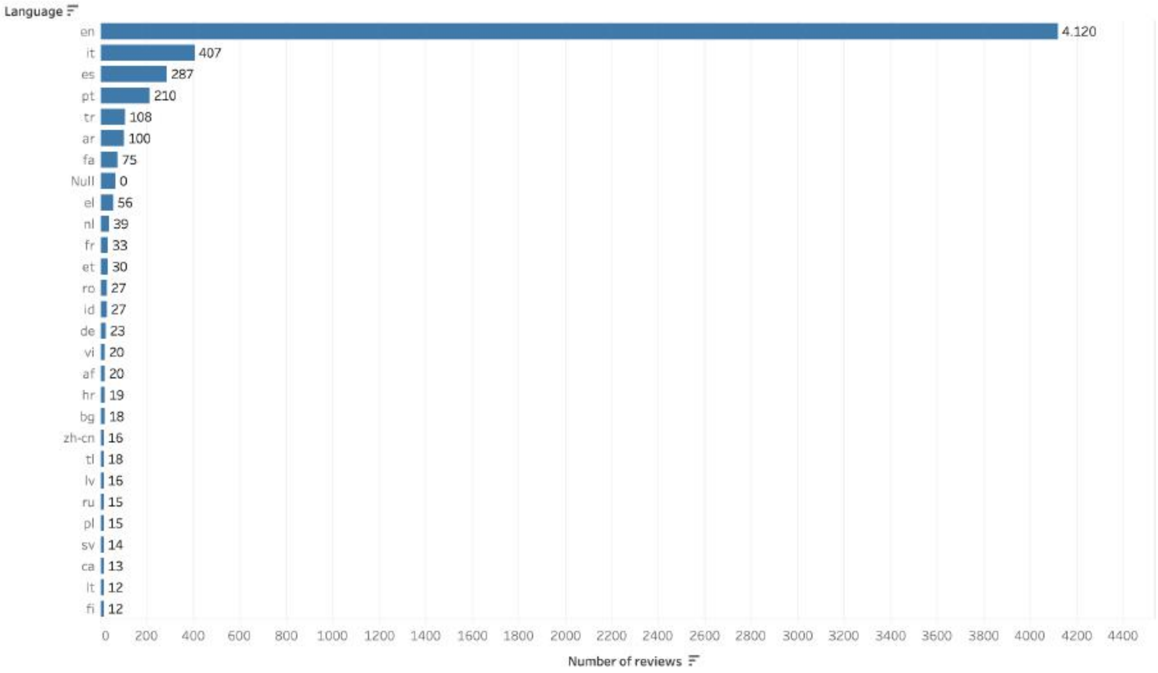


Figure 3: count of detected languages

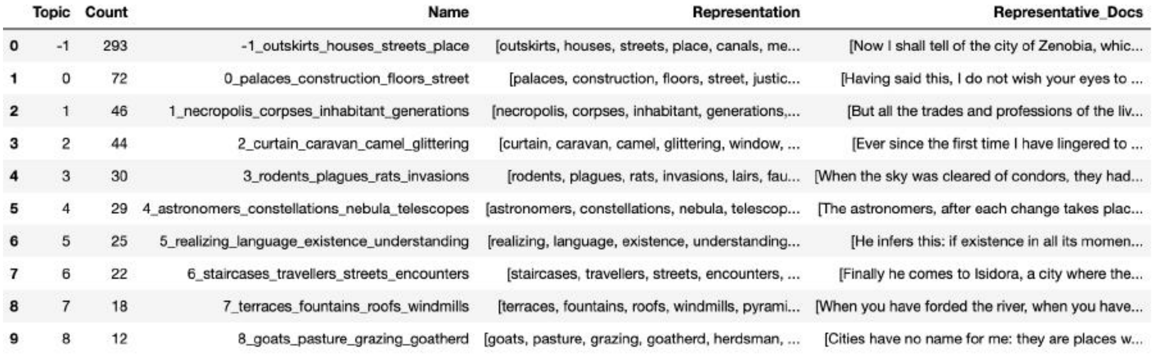
Chapter4

Text analytics and visualization

This chapter focuses on the text analytics and visualization strategy applied to both the book "Invisible Cities" and its reviews. Split into four sections, the initial two are dedicated to the book itself. Section 4.1 delves deep into the utilization of BERTopic for uncovering latent topics, highlighting its distinctive features and its edge over traditional LDA methods. Section 4.2 then elucidates the visualization methodologies adopted for the results derived from 4.1, exploring strategies to effectively convey these insights to the end-users. The latter half of the chapter, sections three and four, steer the focus toward the analytics and visualization processes related to the book's reviews.

4.1 Text analytics on Invisible Cities

The textual analysis of the book was applied only to the stories; the dialogue between chapters is not included in the analysis. Inspired by the book's poetic writing style, it is assumed that each sentence contains at most one topic, rather than a mixture of topics, so that each story can still represent multiple topics. Corpus is structured on a sentence-by-sentence basis, rather than on a story-by-story basis, treating them as short texts. Therefore, BERTopic, a method that incorporates deep learning, was chosen. Figure 4 shows the result after applying BERTopic and Figure 5 shows the 2D clustering visualization.



	Topic	Count	Name	Representation	Representative_Docs
0	-1	293	-1_outskirts_houses_streets_place	[outskirts, houses, streets, place, canals, me...	[Now I shall tell of the city of Zenobia, whic...
1	0	72	0_palaces_construction_floors_street	[palaces, construction, floors, street, justic...	[Having said this, I do not wish your eyes to ...
2	1	46	1_necropolis_corpses_inhabitant_generations	[necropolis, corpses, inhabitant, generations,...	[But all the trades and professions of the liv...
3	2	44	2_curtain_caravan_camel_glittering	[curtain, caravan, camel, glittering, window, ...	[Ever since the first time I have lingered to ...
4	3	30	3_rodents_plagues_rats_invasions	[rodents, plagues, rats, invasions, lairs, fau...	[When the sky was cleared of condors, they had...
5	4	29	4_astronomers_constellations_nebula_telescopes	[astronomers, constellations, nebula, telescop...	[The astronomers, after each change takes plac...
6	5	25	5_realizing_language_existence_understanding	[realizing, language, existence, understanding...	[He infers this: if existence in all its momen...
7	6	22	6_staircases_travellers_streets_encounters	[staircases, travellers, streets, encounters, ...	[Finally he comes to Isidora, a city where the...
8	7	18	7_terraces_fountains_roofs_windmills	[terraces, fountains, roofs, windmills, pyrami...	[When you have forded the river, when you have...
9	8	12	8_goats_pasture_grazing_goatherd	[goats, pasture, grazing, goatherd, herdsman, ...	[Cities have no name for me: they are places w...

Figure 4: Result of BERTopic applying on book (the customized stop word includes 'berenices', 'lares', 'penates', 'inhabitants', 'city', 'cities' and all 55 cities’ name)

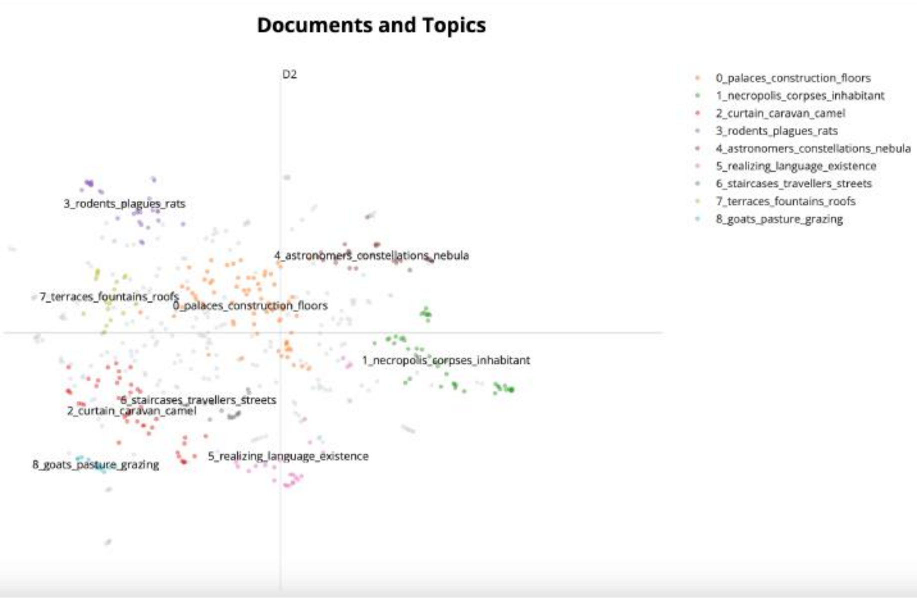


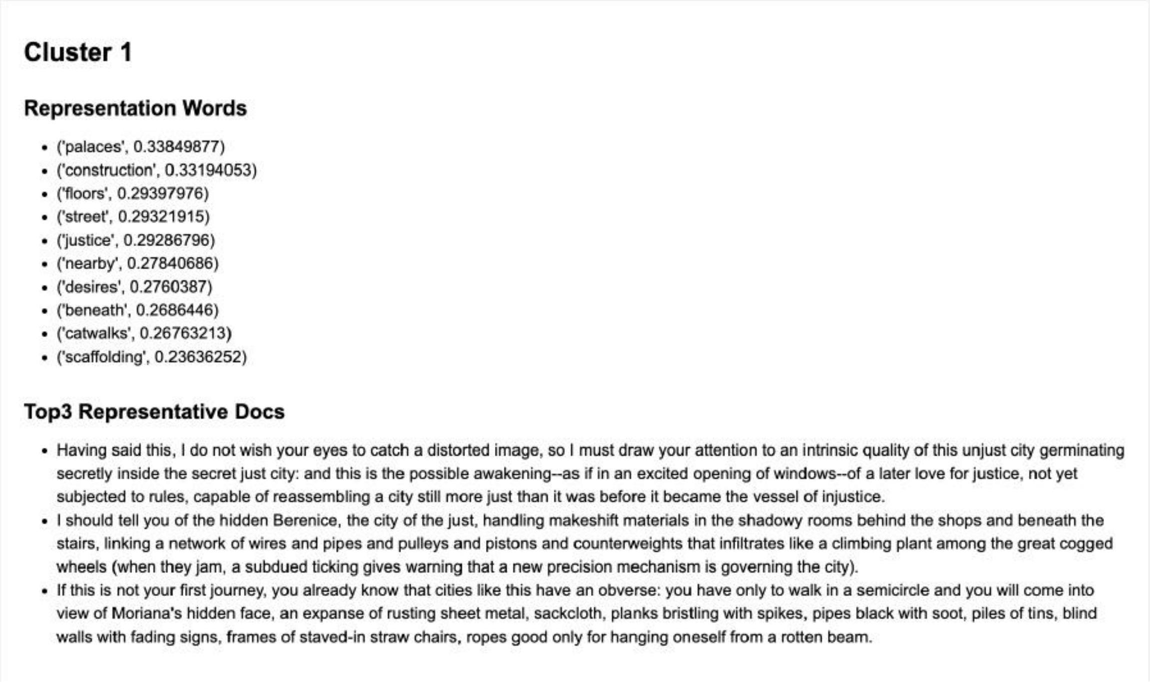
Figure 5: Visualization of 2D clustering

After applying BERTopic, there are 9 clusters were identified confidently, then the topic was summarized and the name was assigned through reading the top three most representative documents (sentences). After this process, these documents (sentences) are mapped to the original location, then checking the topic distribution on each story. In the end, there are six topics are identified in which two clusters are manually merged into one topic, and two clusters are kept open to identify.

The summarization is based on personal ideas and reference to other close reading work.

Cluster 1

Named Complexity. The first topic encompasses the largest cluster of documents. Upon examining the representative documents, all three highlighted sentences emphasize the dual nature and complexity of a city, with a primary focus on its negative aspects (Figure 6). This narrative strongly resonates with the societal context of the time, during which the rapidly growing real estate industry was introducing a new urban landscape. On the other hand, behind the physical renewal of cities lay social upheaval and an oversight of social dimensions in urban planning. As such, this topic is named 'Complexity' to underscore the intricate nature of cities as organic entities, rather than merely physical constructs dictated by buildings. This realization began to germinate about 10 years before this book came out. Correspondingly, the representative words also depict a city in a state of physical renewal and construction.



Cluster 1

Representation Words

- ('palaces', 0.33849877)
- ('construction', 0.33194053)
- ('floors', 0.29397976)
- ('street', 0.29321915)
- ('justice', 0.29286796)
- ('nearby', 0.27840686)
- ('desires', 0.2760387)
- ('beneath', 0.2686446)
- ('catwalks', 0.26763213)
- ('scaffolding', 0.23636252)

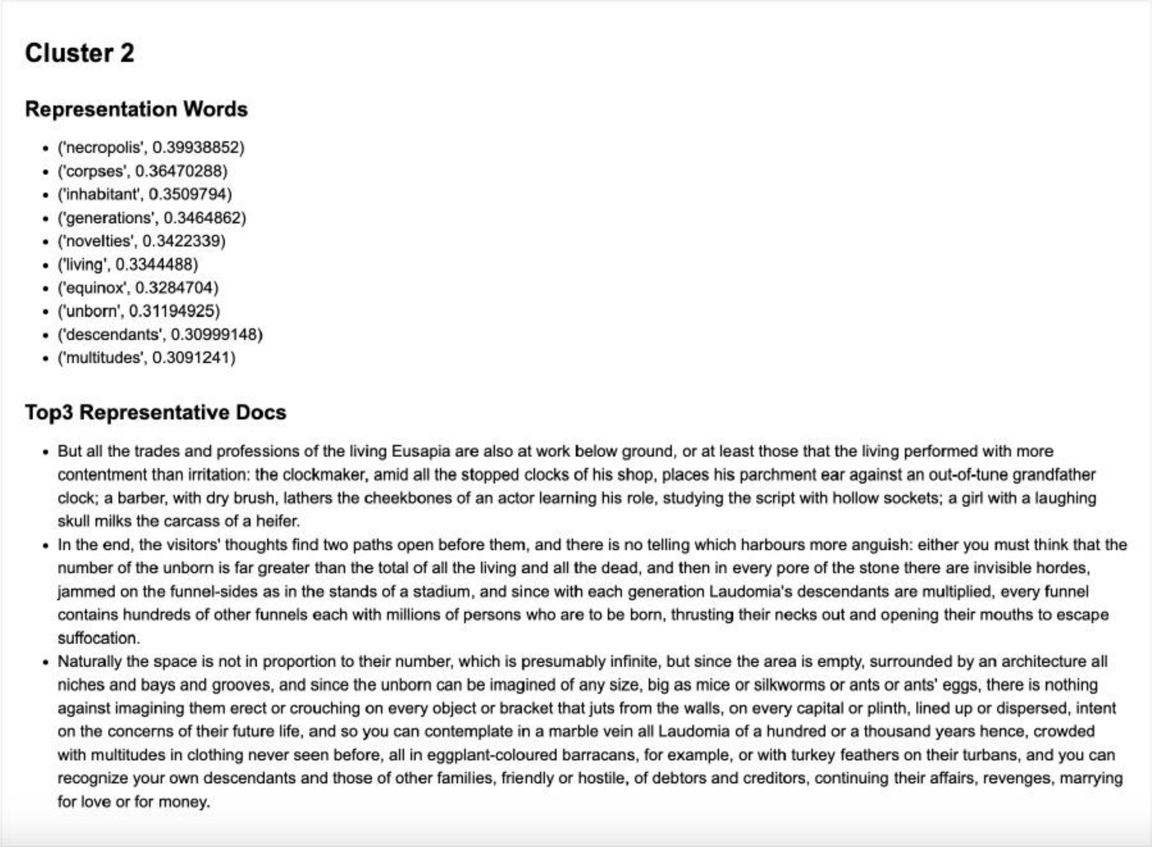
Top3 Representative Docs

- Having said this, I do not wish your eyes to catch a distorted image, so I must draw your attention to an intrinsic quality of this unjust city germinating secretly inside the secret just city: and this is the possible awakening--as if in an excited opening of windows--of a later love for justice, not yet subjected to rules, capable of reassembling a city still more just than it was before it became the vessel of injustice.
- I should tell you of the hidden Berenice, the city of the just, handling makeshift materials in the shadowy rooms behind the shops and beneath the stairs, linking a network of wires and pipes and pulleys and pistons and counterweights that infiltrates like a climbing plant among the great cogged wheels (when they jam, a subdued ticking gives warning that a new precision mechanism is governing the city).
- If this is not your first journey, you already know that cities like this have an obverse: you have only to walk in a semicircle and you will come into view of Moriana's hidden face, an expanse of rusting sheet metal, sackcloth, planks bristling with spikes, pipes black with soot, piles of tins, blind walls with fading signs, frames of staved-in straw chairs, ropes good only for hanging oneself from a rotten beam.

Figure 6: representation words and Top 3 representative documents of Cluster 1 (score after each word means the probability of related to this topic)

Cluster 2

Named Escape. The documents within this cluster, set against the backdrop of a fictitious ghost world, reflect the work landscape within modern industrial cities (Figure 7). The descriptions are predominantly negative, filled with endless repetition. Even though the first document's description seems positive, it refers to choices made after death, indirectly mirroring a sense of weariness and desire to escape from work in real life. Considering the representative words, these metaphors map onto two facets: the escape from present life and an unrealistic fantasy for a better one.



Cluster 2

Representation Words

- ('necropolis', 0.39938852)
- ('corpses', 0.36470288)
- ('inhabitant', 0.3509794)
- ('generations', 0.3464862)
- ('novelties', 0.3422339)
- ('living', 0.3344488)
- ('equinox', 0.3284704)
- ('unborn', 0.31194925)
- ('descendants', 0.30999148)
- ('multitudes', 0.3091241)

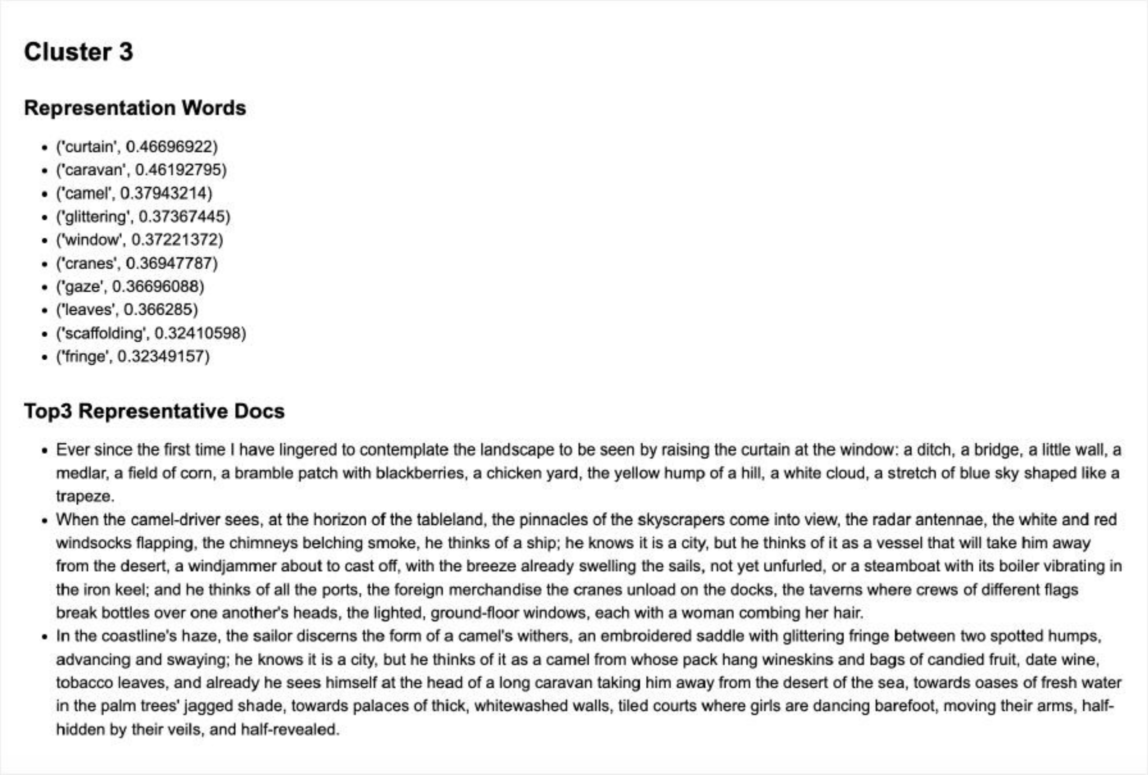
Top3 Representative Docs

- But all the trades and professions of the living Eusapia are also at work below ground, or at least those that the living performed with more contentment than irritation: the clockmaker, amid all the stopped clocks of his shop, places his parchment ear against an out-of-tune grandfather clock; a barber, with dry brush, lathers the cheekbones of an actor learning his role, studying the script with hollow sockets; a girl with a laughing skull milks the carcass of a heifer.
- In the end, the visitors' thoughts find two paths open before them, and there is no telling which harbours more anguish: either you must think that the number of the unborn is far greater than the total of all the living and all the dead, and then in every pore of the stone there are invisible hordes, jammed on the funnel-sides as in the stands of a stadium, and since with each generation Laudomia's descendants are multiplied, every funnel contains hundreds of other funnels each with millions of persons who are to be born, thrusting their necks out and opening their mouths to escape suffocation.
- Naturally the space is not in proportion to their number, which is presumably infinite, but since the area is empty, surrounded by an architecture all niches and bays and grooves, and since the unborn can be imagined of any size, big as mice or silkworms or ants or ants' eggs, there is nothing against imagining them erect or crouching on every object or bracket that juts from the walls, on every capital or plinth, lined up or dispersed, intent on the concerns of their future life, and so you can contemplate in a marble vein all Laudomia of a hundred or a thousand years hence, crowded with multitudes in clothing never seen before, all in eggplant-coloured barracans, for example, or with turkey feathers on their turbans, and you can recognize your own descendants and those of other families, friendly or hostile, of debtors and creditors, continuing their affairs, revenges, marrying for love or for money.

Figure 7: representation words and Top 3 representative documents of Cluster 2 (score after each word means the probability of related to this topic)

Cluster 3

Named Gaze. The documents within this cluster underscore the beauty seen through the gaze of different roles, yet these ideals greatly diverge from the roles' actual identities (Figure 8). The term 'Gaze' was chosen to encapsulate an individual's perception of others or groups, which can also be interpreted as desire or longing. It's noteworthy that in this context, 'Gaze' does not carry any special connotation in terms of power structure, as this concept was primarily introduced in 1975 by Michel Foucault in "Discipline and Punish: The Birth of the Prison", which was published later than this book. When examining the representative words, it is challenging to align these metaphors with the theme deduced from the documents, suggesting a need for further close reading.



Cluster 3

Representation Words

- ('curtain', 0.46696922)
- ('caravan', 0.46192795)
- ('camel', 0.37943214)
- ('glittering', 0.37367445)
- ('window', 0.37221372)
- ('cranes', 0.36947787)
- ('gaze', 0.36696088)
- ('leaves', 0.366285)
- ('scaffolding', 0.32410598)
- ('fringe', 0.32349157)

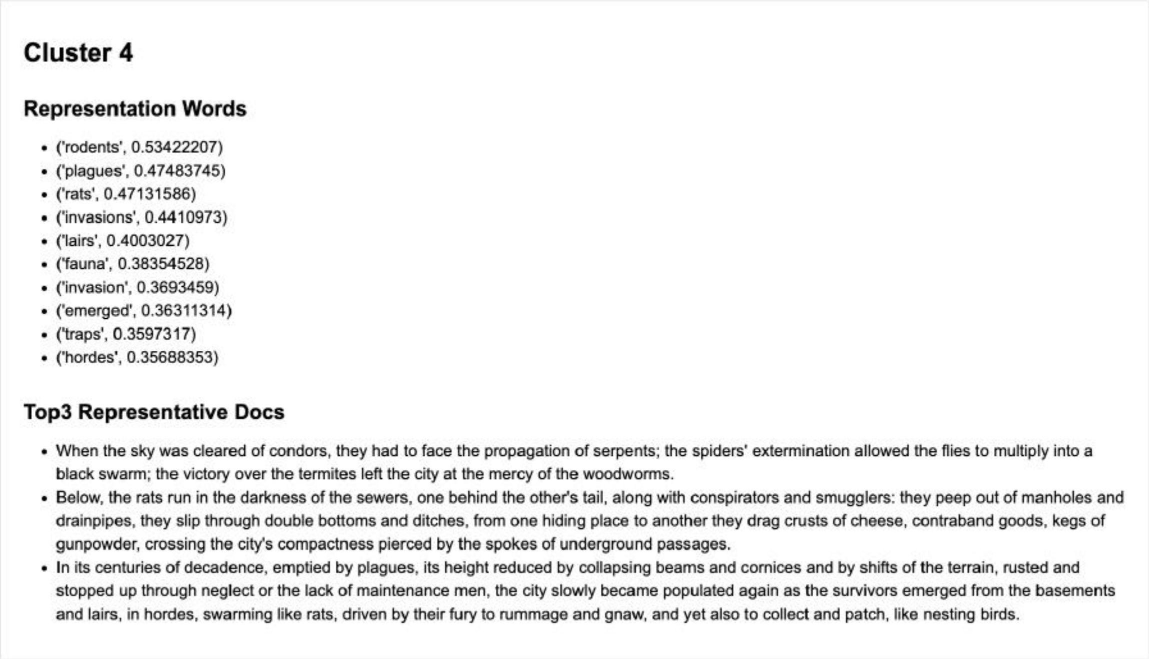
Top3 Representative Docs

- Ever since the first time I have lingered to contemplate the landscape to be seen by raising the curtain at the window: a ditch, a bridge, a little wall, a medlar, a field of corn, a bramble patch with blackberries, a chicken yard, the yellow hump of a hill, a white cloud, a stretch of blue sky shaped like a trapeze.
- When the camel-driver sees, at the horizon of the tableland, the pinnacles of the skyscrapers come into view, the radar antennae, the white and red windsocks flapping, the chimneys belching smoke, he thinks of a ship; he knows it is a city, but he thinks of it as a vessel that will take him away from the desert, a windjammer about to cast off, with the breeze already swelling the sails, not yet unfurled, or a steamboat with its boiler vibrating in the iron keel; and he thinks of all the ports, the foreign merchandise the cranes unload on the docks, the taverns where crews of different flags break bottles over one another's heads, the lighted, ground-floor windows, each with a woman combing her hair.
- In the coastline's haze, the sailor discerns the form of a camel's withers, an embroidered saddle with glittering fringe between two spotted humps, advancing and swaying; he knows it is a city, but he thinks of it as a camel from whose pack hang wineskins and bags of candied fruit, date wine, tobacco leaves, and already he sees himself at the head of a long caravan taking him away from the desert of the sea, towards oases of fresh water in the palm trees' jagged shade, towards palaces of thick, whitewashed walls, tiled courts where girls are dancing barefoot, moving their arms, half-hidden by their veils, and half-revealed.

Figure 8: representation words and Top 3 representative documents of Cluster 3 (score after each word means the probability of related to this topic)

Cluster 4

Named Disasters. The documents in this cluster depict a city being gradually destroyed or the world post-destruction (Figure 9). These imaginative descriptions focus on the negative aspects, the metaphoric expressions seeming more like the author's insights into the hidden darkness beneath apparent prosperity, thereby awakening readers' vigilance towards surface-level booms. In conjunction with the representative words, these metaphors convey the author's concern that the unnoticed shadows are slowly eroding our cities.



Cluster 4

Representation Words

- ('rodents', 0.53422207)
- ('plagues', 0.47483745)
- ('rats', 0.47131586)
- ('invasions', 0.4410973)
- ('lairs', 0.4003027)
- ('fauna', 0.38354528)
- ('invasion', 0.3693459)
- ('emerged', 0.36311314)
- ('traps', 0.3597317)
- ('hordes', 0.35688353)

Top3 Representative Docs

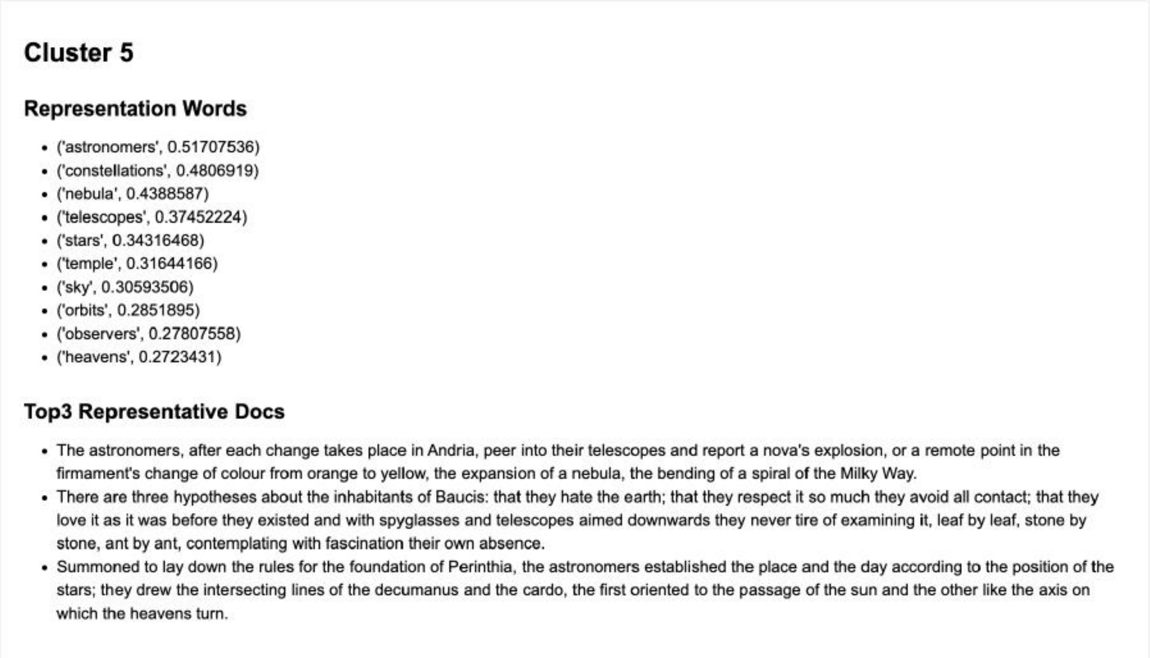
- When the sky was cleared of condors, they had to face the propagation of serpents; the spiders' extermination allowed the flies to multiply into a black swarm; the victory over the termites left the city at the mercy of the woodworms.
- Below, the rats run in the darkness of the sewers, one behind the other's tail, along with conspirators and smugglers: they peep out of manholes and drainpipes, they slip through double bottoms and ditches, from one hiding place to another they drag crusts of cheese, contraband goods, kegs of gunpowder, crossing the city's compactness pierced by the spokes of underground passages.
- In its centuries of decadence, emptied by plagues, its height reduced by collapsing beams and cornices and by shifts of the terrain, rusted and stopped up through neglect or the lack of maintenance men, the city slowly became populated again as the survivors emerged from the basements and lairs, in hordes, swarming like rats, driven by their fury to rummage and gnaw, and yet also to collect and patch, like nesting birds.

Figure 9: representation words and Top 3 representative documents of Cluster 4 (score after each word means the probability of related to this topic)

Cluster 5

Named The Fatal Conceit, this title is borrowed from Friedrich Hayek's (1988) classic work of liberal thought, subtitled *The Errors of Socialism*. The documents in this cluster abstractly describe the absurd actions of a group blindly worshipping rationality. The naming of this theme encompasses two aspects. On one hand, it reflects Italo Calvino's ideological shift following his disappointment over the Soviet Union's invasion of Hungary in 1956 and his subsequent departure from the Italian Communist Party in 1957. On the other hand, it critiques the socialist totalitarian society, which exercises control over society through conceited rationality, an unsustainable approach for the development of cities and societies. This notion aligns with the resurgence of liberalism in the 1970s.

In relation to the representative words, these terms signify the advancement of astronomy, expressing a trend of technological reverence (Figure10). This adulation of technology is accompanied by naive social conjectures, such as the simplistic post-WWII understanding that poverty was merely the result of impoverished living conditions. This led to numerous slum clearance projects and the rapid rise of high-rise social housing, both of which sowed the seeds for future social issues.



Cluster 5

Representation Words

- ('astronomers', 0.51707536)
- ('constellations', 0.4806919)
- ('nebula', 0.4388587)
- ('telescopes', 0.37452224)
- ('stars', 0.34316468)
- ('temple', 0.31644166)
- ('sky', 0.30593506)
- ('orbits', 0.2851895)
- ('observers', 0.27807558)
- ('heavens', 0.2723431)

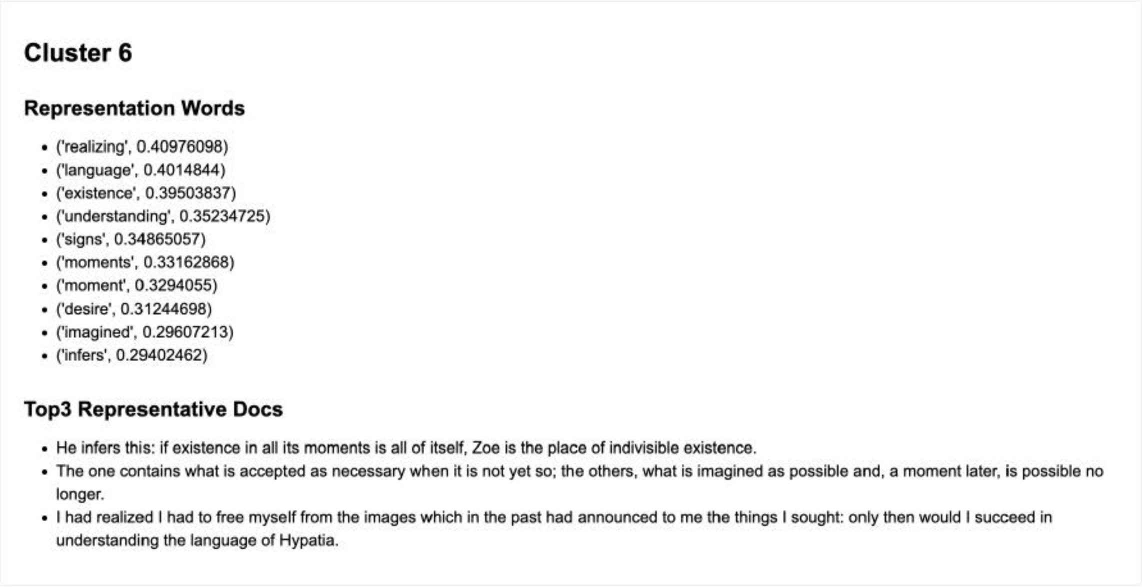
Top3 Representative Docs

- The astronomers, after each change takes place in Andria, peer into their telescopes and report a nova's explosion, or a remote point in the firmament's change of colour from orange to yellow, the expansion of a nebula, the bending of a spiral of the Milky Way.
- There are three hypotheses about the inhabitants of Baucis: that they hate the earth; that they respect it so much they avoid all contact; that they love it as it was before they existed and with spyglasses and telescopes aimed downwards they never tire of examining it, leaf by leaf, stone by stone, ant by ant, contemplating with fascination their own absence.
- Summoned to lay down the rules for the foundation of Perinthia, the astronomers established the place and the day according to the position of the stars; they drew the intersecting lines of the decumanus and the cardo, the first oriented to the passage of the sun and the other like the axis on which the heavens turn.

Figure 10: representation words and Top 3 representative documents of Cluster 5 (score after each word means the probability of related to this topic)

Cluster 6

Named Existentialism. The documents in this cluster are more abstract, yet they carry a strong philosophical undertone. The central sentence of this cluster embodies existentialist philosophical speculation (Figure 11). Existentialism also implies a critique of rationalism, favoring decisions could not be made based pure rationality. When examining the representative words, these abstract terms make it challenging to summarize the topic, however, when combined with the documents, they align well under the topic of existentialism.



Cluster 6

Representation Words

- ('realizing', 0.40976098)
- ('language', 0.4014844)
- ('existence', 0.39503837)
- ('understanding', 0.35234725)
- ('signs', 0.34865057)
- ('moments', 0.33162868)
- ('moment', 0.3294055)
- ('desire', 0.31244698)
- ('imagined', 0.29607213)
- ('infers', 0.29402462)

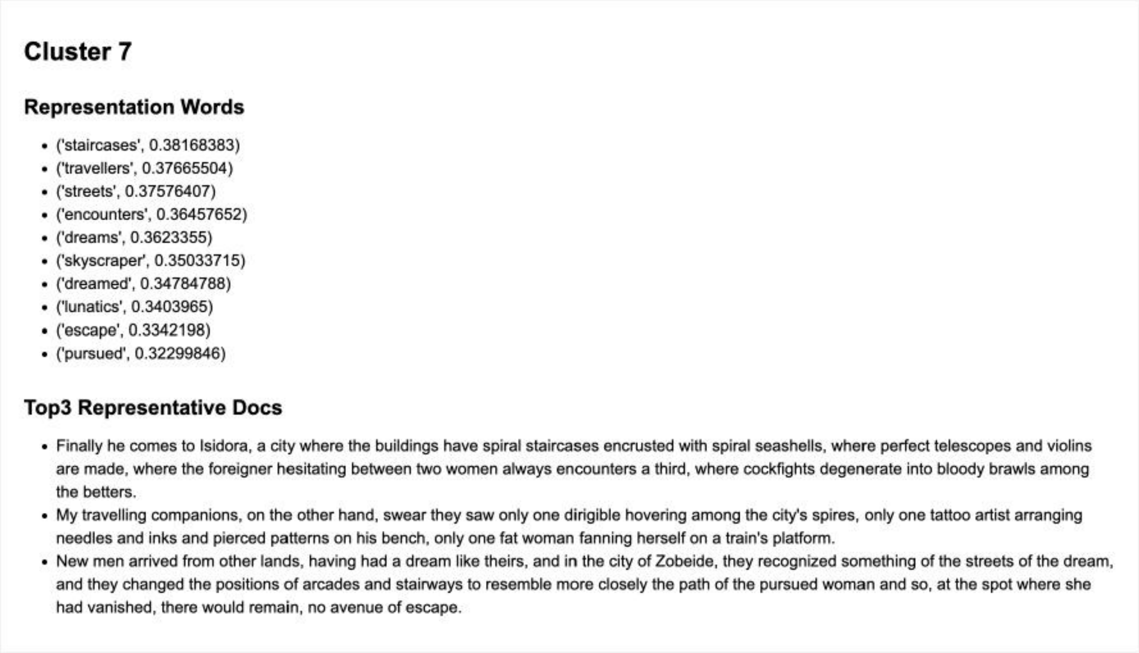
Top3 Representative Docs

- He infers this: if existence in all its moments is all of itself, Zoe is the place of indivisible existence.
- The one contains what is accepted as necessary when it is not yet so; the others, what is imagined as possible and, a moment later, is possible no longer.
- I had realized I had to free myself from the images which in the past had announced to me the things I sought: only then would I succeed in understanding the language of Hypatia.

Figure 11: representation words and Top 3 representative documents of Cluster 6 (score after each word means the probability of related to this topic)

Cluster 7

Named Desire. The documents in this cluster not only depict some worldly desires in detail, but they also mirror how the pursuit of these desires can lead to lost goals and eventual failure in the policy-making process (Figure 12). However, during the mapping process, this topic is similar to Cluster 3, therefore, it has been merged into that topic.



Cluster 7

Representation Words

- ('staircases', 0.38168383)
- ('travellers', 0.37665504)
- ('streets', 0.37576407)
- ('encounters', 0.36457652)
- ('dreams', 0.3623355)
- ('skyscraper', 0.35033715)
- ('dreamed', 0.34784788)
- ('lunatics', 0.3403965)
- ('escape', 0.3342198)
- ('pursued', 0.32299846)

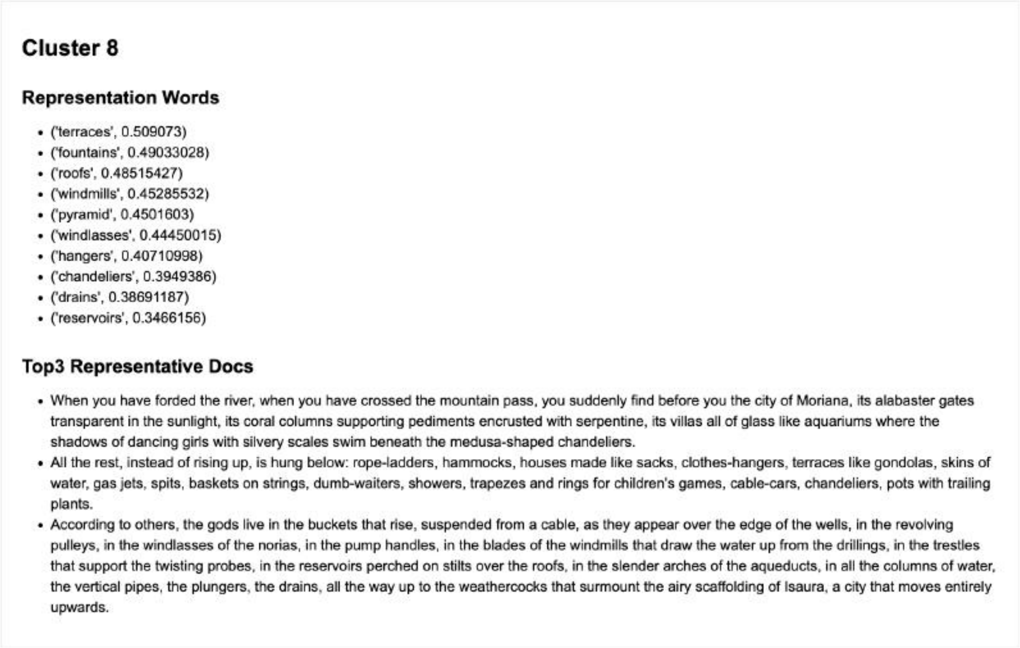
Top3 Representative Docs

- Finally he comes to Isidora, a city where the buildings have spiral staircases encrusted with spiral seashells, where perfect telescopes and violins are made, where the foreigner hesitating between two women always encounters a third, where cockfights degenerate into bloody brawls among the betters.
- My travelling companions, on the other hand, swear they saw only one dirigible hovering among the city's spires, only one tattoo artist arranging needles and inks and pierced patterns on his bench, only one fat woman fanning herself on a train's platform.
- New men arrived from other lands, having had a dream like theirs, and in the city of Zobeide, they recognized something of the streets of the dream, and they changed the positions of arcades and stairways to resemble more closely the path of the pursued woman and so, at the spot where she had vanished, there would remain, no avenue of escape.

Figure 12: representation words and Top 3 representative documents of Cluster 7 (score after each word means the probability of related to this topic)

Cluster 8 & 9

Unnamed clusters. Due to the difficulty in summarizing a theme from the author's related background, the decision was made to forego naming and explaining these clusters. However, this does not imply that the cluster loses its significance. As the ultimate product will be presented to the public, an open approach is adopted for this cluster (Figure 13 & 14).



Cluster 8

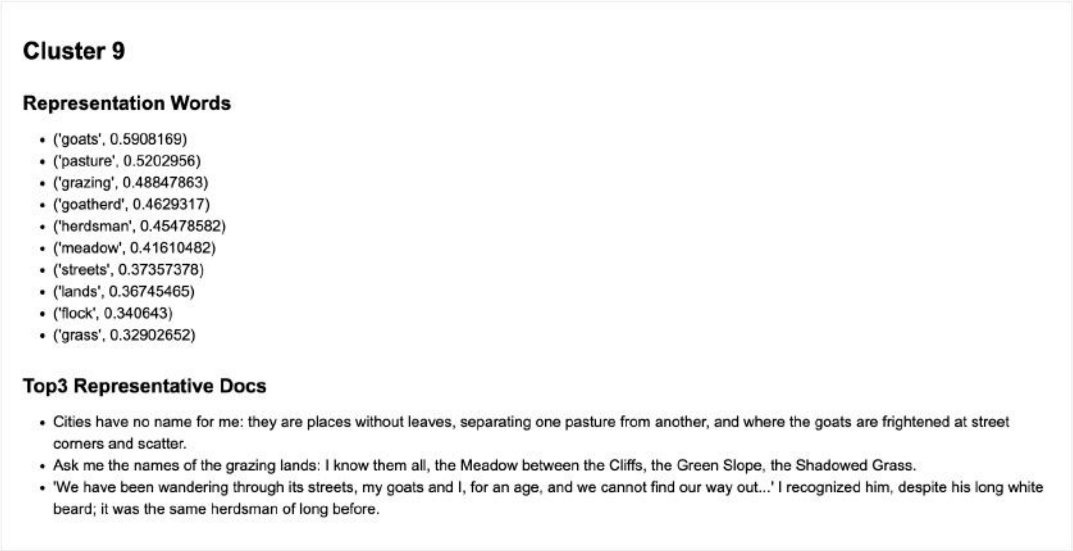
Representation Words

- ('terraces', 0.509073)
- ('fountains', 0.49033028)
- ('roofs', 0.48515427)
- ('windmills', 0.45285532)
- ('pyramid', 0.4501603)
- ('windlasses', 0.44450015)
- ('hangers', 0.40710998)
- ('chandeliers', 0.3949386)
- ('drains', 0.38691187)
- ('reservoirs', 0.3466156)

Top3 Representative Docs

- When you have forded the river, when you have crossed the mountain pass, you suddenly find before you the city of Moriana, its alabaster gates transparent in the sunlight, its coral columns supporting pediments encrusted with serpentine, its villas all of glass like aquariums where the shadows of dancing girls with silvery scales swim beneath the medusa-shaped chandeliers.
- All the rest, instead of rising up, is hung below: rope-ladders, hammocks, houses made like sacks, clothes-hangers, terraces like gondolas, skins of water, gas jets, spits, baskets on strings, dumb-waiters, showers, trapezes and rings for children's games, cable-cars, chandeliers, pots with trailing plants.
- According to others, the gods live in the buckets that rise, suspended from a cable, as they appear over the edge of the wells, in the revolving pulleys, in the windlasses of the norias, in the pump handles, in the blades of the windmills that draw the water up from the drillings, in the trestles that support the twisting probes, in the reservoirs perched on stilts over the roofs, in the slender arches of the aqueducts, in all the columns of water, the vertical pipes, the plungers, the drains, all the way up to the weathercocks that surmount the airy scaffolding of Isaura, a city that moves entirely upwards.

Figure 13: representation words and Top 3 representative documents of Cluster 8 (score after each word means the probability of related to this topic)



Cluster 9

Representation Words

- ('goats', 0.5908169)
- ('pasture', 0.5202956)
- ('grazing', 0.48847863)
- ('goatherd', 0.4629317)
- ('herdsman', 0.45478582)
- ('meadow', 0.41610482)
- ('streets', 0.37357378)
- ('lands', 0.36745465)
- ('flock', 0.340643)
- ('grass', 0.32902652)

Top3 Representative Docs

- Cities have no name for me: they are places without leaves, separating one pasture from another, and where the goats are frightened at street corners and scatter.
- Ask me the names of the grazing lands: I know them all, the Meadow between the Cliffs, the Green Slope, the Shadowed Grass.
- 'We have been wandering through its streets, my goats and I, for an age, and we cannot find our way out...' I recognized him, despite his long white beard; it was the same herdsman of long before.

Figure 14: representation words and Top 3 representative documents of Cluster 9 (score after each word means the probability of related to this topic)

Discussion:

The computational results of BERTopic can provide indirect references for understanding abstract literary works. Through the analysis of "Invisible Cities", a work renowned for its imagination and metaphors, we can summarize the following characteristics and explore the possibilities of interpreting literary works through computational methods:

It is an indirect method, and the interpretation is subjective. This indirect interpretation differs from traditional LDA methods; it still requires users to summarize topics through close reading. As seen in the interpretation of each topic above, this interpretation is based on the prior knowledge of the author. Therefore, different perspectives may yield different results when facing the same cluster; if an individual cannot summarize a theme, it doesn't mean it is themeless for everyone.

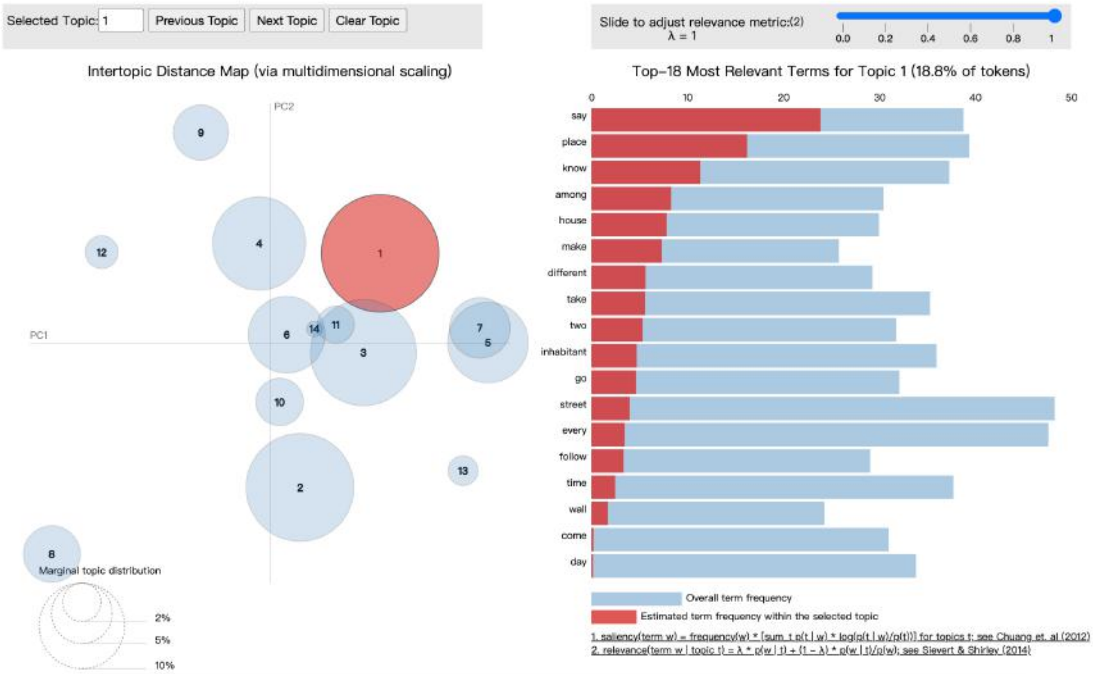
During the interpretation process, representative words and documents may complement each other, or they may not coincide. For instance, in Topic 1, there is a strong correlation between the theme and the words, but the correlation is weaker in Topic 3. The reason mainly lies in the method's characteristics: representative documents and words are obtained by first clustering to get clusters, and then calculating the representative words from the clusters. Therefore, the main interpretation should be through reading the representative documents rather than the representative words.

In addition to providing additional explanations for the topic, it also offers a way to understand metaphors. Due to the fictitious nature of the book, the descriptions are more reflective of real life, which can easily confuse readers. In the interpretation process, representative words in this case can be understood as clusters of metaphors and images. Once the theme is summarized from the clustered documents, the implied meanings of these metaphors become clearer.

The computational results may require manual correction. Since the interpretation of a topic is subjective, two distinctive clusters may be interpreted as similar, which requires a holistic correction after summarizing the topics. For instance, Cluster 7 and Cluster 3 express similar topic, so they were ultimately merged into one topic. At the same time, computed clusters do not necessarily have interpretable themes, such as Cluster 8 and 9.

Compared to LDA, it was challenging for the author to extract latent topics from the results. During the application process, the author treated individual stories as documents, creating a corpus comprised of documents with unigram and bigram. The stop words included the names of 55 cities and the basic stop words from NLTK (Natural Language Toolkit). Figure 15 showcases one topic from the results, from which it was difficult to summarize any potential topic. Consequently, this approach was eventually abandoned.

In combination with these characteristics, BERTopic, a kind of secondary interpretation provides a new way of close reading. In this particular case, the text is very abstract, the traditional method doesn't guide the model well to get better topic results, as most of the words are metaphors themselves. Probability and statistical methods find it hard to replace the advantage of close reading in this regard. However, BERTopic helps users cluster sentences that may express similar meanings, thereby assisting in topic induction and summarization.



Selected Topic: 1

Slide to adjust relevance metric:(2)
λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 15: LDA results visualized using pyLDAvis (Topic 1 selected to display the most relevant terms)

4.2 Visualization of topic modeling results

Building on the previous section's exposition of clustering, this segment will focus on how to convey the results of topic modeling. In practice, some scholars have responded to this need. For instance, Sievert and Shirley (2014) proposed the LDAviz system to aid users in visualizing topic-term relationships to understand the LDA model. Similarly, BERTopic offers built-in functions for visual interpretation of results. For a researcher, these tools are invaluable for determining the number of topics, interpreting those topics, and regenerating knowledge. However, for the reader, they do not convey the interpreted results directly and imply a steeper learning curve. This gap necessitates further visualization design tailored to the end-user.

Regarding visualization design, the objective is to convey to readers the potential topics shared between narratives. Thus, the ultimate design emphasizes narratives rather than standalone sentences, recognizing that readers interact with and interpret stories, not disjointed phrases.

After interpretation and adjustments, the analysis resulted in 8 topics and 2 unnamed clusters intended for visualization. However, topic 1 was not presented in the visualization because it represents the core theme of the book, and it will be elaborated in the textual description. Additionally, cluster 7 was amalgamated into cluster 3.

To express the intricate relationships between results, inspired by Bostock's (2012) interactive Chord diagram visualizing Uber ride data (Figure16), this thesis employs a similar interactive narrative approach and graphic representation. The circular appearance also hints at the book's non-linear nature. However, while the Chord diagram is apt for showing data flow, it's not suitable for this project. As a result, edge bundling was chosen to highlight narratives that share the same topics (Figure 17).

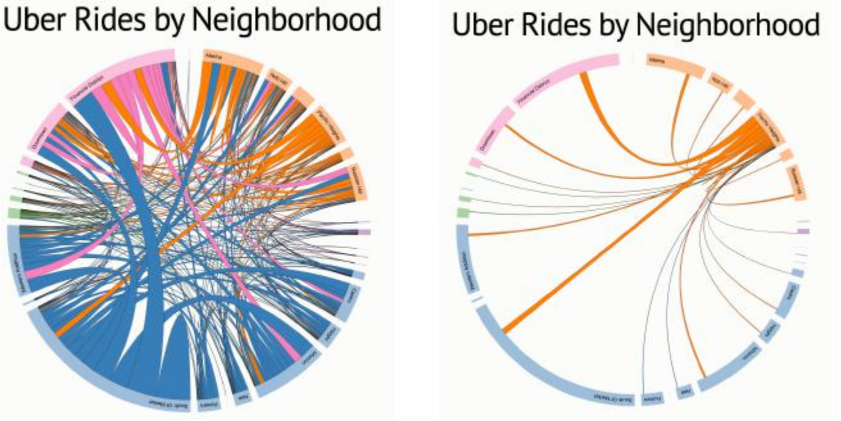
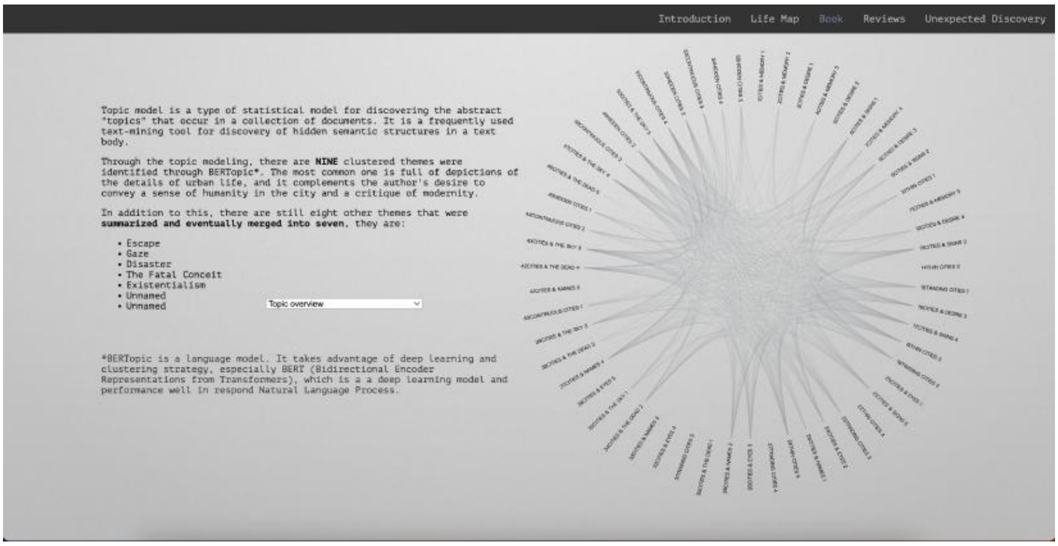


Figure 16: Chord diagram of Uber Rides by Neighborhood (Bostock, 2012)

The outcome is a visualization of topic modeling results using edge bundling, complemented by an interactive interface for interpretation (Figure 17). Initially, different colors were considered to distinguish various topics, but this approach resulted in low readability and failed to convey information effectively. Instead now, when explaining a single topic, highlighted nodes are efficient to convey that they share the same topic. Thus, for links, a singular color was chosen, adjusting transparency as options change (Figure 18).



Topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. It is a frequently used text-mining tool for discovery of hidden semantic structures in a text body.

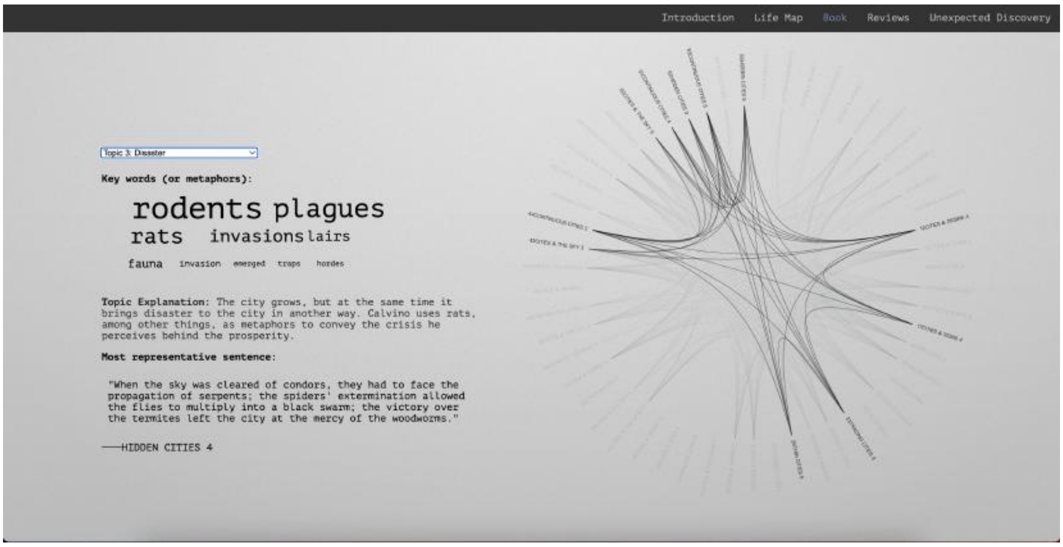
Through the topic modeling, there are **NINE** clustered themes were identified through BERTopic*. The most common one is full of depictions of the details of urban life, and it complements the author's desire to convey a sense of humanity in the city and a critique of modernity.

In addition to this, there are still eight other themes that were **summarized and eventually merged into seven**, they are:

- Escape
- Gaze
- Disaster
- The Fatal Conceit
- Existentialism
- Unnamed
- Unnamed

*BERTopic is a language model. It takes advantage of deep learning and clustering strategy, especially BERT (Bidirectional Encoder Representations from Transformers), which is a a deep learning model and performance well in respond Natural Language Process.

Figure 17: website interface (topic overview status)



Key words (or metaphors):

rodents plagues
rats invasions lairs
fauna invasion emerged traps hordes

Topic Explanation: The city grows, but at the same time it brings disaster to the city in another way. Calvino uses rats, among other things, as metaphors to convey the crisis he perceives behind the prosperity.

Most representative sentence:

"When the sky was cleared of condors, they had to face the propagation of serpents; the spiders' extermination allowed the flies to multiply into a black swarm; the victory over the termites left the city at the mercy of the woodworms."

——HIDDEN CITIES 4

Figure 18: website interface (when one topic is selected)

Apart from highlighting narratives that share topics, representative words and documents are also presented to the reader. Word clouds were chosen to depict representative words, with word size indicating its relevance to the topic (Figure 19). Representative documents are conveyed directly in text form, but with layout adjustments and source citations to intuitively inform the reader of it is a quotation from the book (Figure 20).

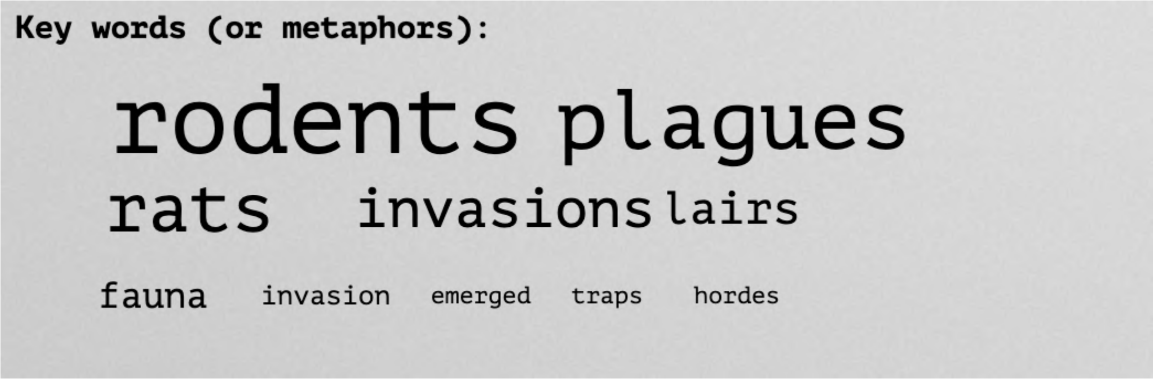
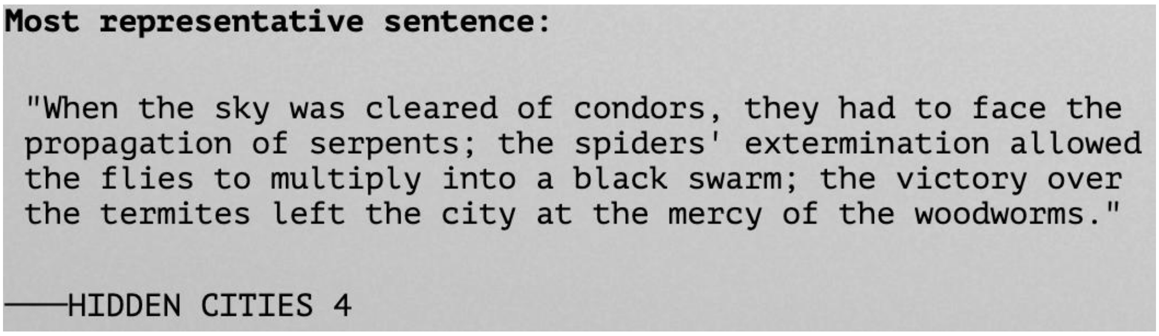


Figure 19: word could for representative words


Most representative sentence:

"When the sky was cleared of condors, they had to face the propagation of serpents; the spiders' extermination allowed the flies to multiply into a black swarm; the victory over the termites left the city at the mercy of the woodworms."

——HIDDEN CITIES 4

Figure 20: the most representative document of topic

The choice of visualization method is not definitive; it must be flexibly considered to cater to various end-users. Drawing from experiences in a previous project, the author and project team members evaluated two unique visualization methods, both about visualization of topic modeling results. Figure 21 depicts two approaches to topic modeling descriptions of film and TV productions on the Netflix platform to discern the evolution of underlying themes. On the left, a tree map portrays topic distribution, where colors represent different topics, and the size of each segment indicates the degree of relevance of topic terms. The more intricate design on the right allows users to sort related terms by clicking on the theme on the left. The presence of both red and gray indicates changes over two different time periods.

In the decision stage, the left method directly conveys topic distribution but offers limited exploration. In contrast, the method on the right is more intricate, demanding more time for comprehension, yet offering richer exploratory possibilities. Considering that the ultimate readership consists of film and TV producers, it was assumed that they would be more willing to invest time in learning how to use the visualization tool. Furthermore, they are likely to favor a tool that allows for deeper exploration and extraction of more information. Hence, the approach on the right was ultimately chosen.

The visualization approach for this thesis incorporated lessons from two previous attempts. Both efforts relied too heavily on the visualization itself to convey information, overlooking the synergistic role of text and other elements in tandem with visuals. Moreover, although the more intricate method on the right was eventually chosen, it still mainly required reading to gather information and lacked visual appeal compared to the left-hand approach. Consequently, in this thesis, a single representation wasn't used to express all the information. Instead, an interactive interface was chosen, segmenting the final information into edge bundling and word cloud visuals, accompanied by structured text to elucidate the content. This approach ensured both visual allure and comprehensive content interpretation.

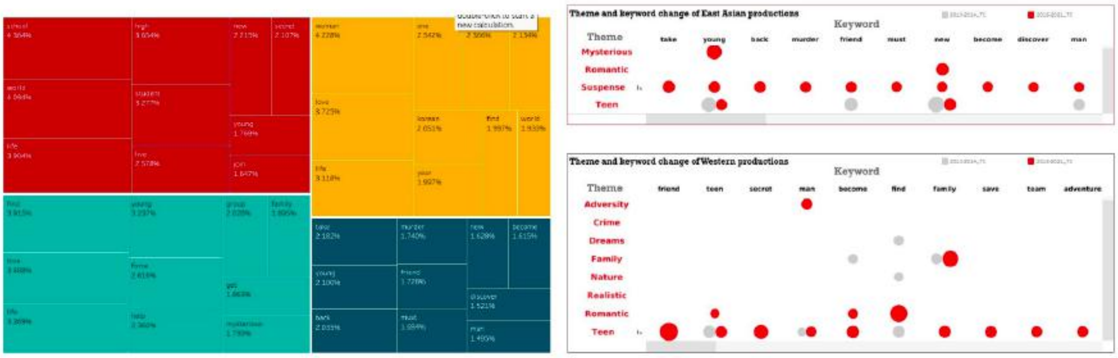


Figure 21: visualization of Netflix

4.3 Text analytics on public's reviews

This part will focus on applying text analytics to English reviews and then summarize information from what reviewers discussed. The first part filtered reviews and prepared data for topic modeling. The second part applied BERTopic in filtered reviews, then the third part explored different subtopics results from the last part.

4.3.1 Review filter

Even though the platform has a review mechanism to filter malicious texts, the dataset still needs to be further processed to reduce noise. First, the deep learning model is utilized to filter those toxic reviews. There are only a few reviews detected, they are using obscenities to express extreme emotion both positively and negatively. Then, hypotheses are made by reading a subset of reviews each sentence of each review contains at most one topic. So, still, adopting a clustering strategy and all reviews are separated into sentences, preparing for the BERTopic.

Also, there another phenomenon that could be noise for topic modeling is the quotation, but it inspired another approach to do analytics. In the beginning, through the regular expression, words between “ ” are detected as quotations and removed them. However, what is founded from the results is still considerable quotation noise in the dataset. For the result of topic modeling, it increased the topic number and require users to spend time on determining them. However, these quotation clusters reveal potential analytics clue that which sentences are quoted the most.

4.3.2 Topic modeling

In this part, this project still use BERTopic, since not all sentences have related topic, and traditional methods struggle in finding more detailed topics.

The final clustering result includes 103 topics (limit to a topic have at least 20 sentences). Through reading each topic then summary and induction, there are five themes among them, they are Genre Related (20 clusters), Structure Related (14 clusters), Personal Reflections (14 clusters), Recommend Related (21 clusters) and Quotation (25).

The naming and summarization of each theme are as follows:

Genre Related

This theme mainly describes readers' experiences with the author's writing style, primarily focusing on admiration for the author's imagination and poetic style. In addition to imagination and poetry, words such as "prose-like" are commonly used, as well as references to similar authors and works, such as Jorge Luis Borges and Alan Lightman's 'Einstein's Dreams'.

Structure Related:

This theme mainly discusses the tightly woven, almost mathematical structure of the book.

Personal Reflections:

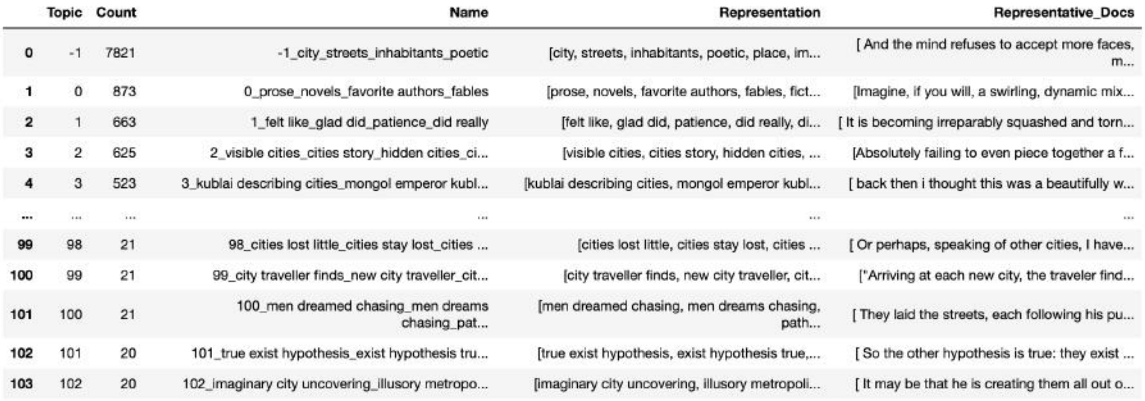
This theme primarily comprises sentences expressing the reviewers' personal experiences, whether negative or positive.

Recommendation Related:

Sentences related to this theme mainly focus on whether or not the book is recommended, or contain very direct expressions of likes and dislikes, generally without specific personal insights.

Quotation:

Sentences in this theme mainly consist of quotations from the original text by readers. Although some quotations were removed before topic modeling, they still make up a considerable number.

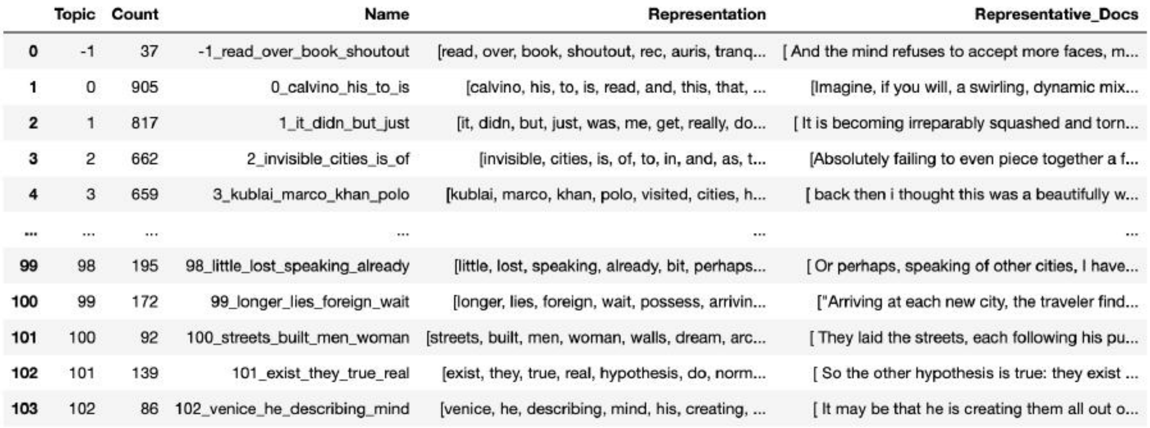


	Topic	Count	Name	Representation	Representative_Docs
0	-1	7821	-1_city_streets_inhabitants_poetic	[city, streets, inhabitants, poetic, place, im...	[And the mind refuses to accept more faces, m...
1	0	873	0_prose_novels_favorite authors_fables	[prose, novels, favorite authors, fables, fict...	[Imagine, if you will, a swirling, dynamic mix...
2	1	663	1_felt like_glad did_patience_did really	[felt like, glad did, patience, did really, di...	[It is becoming irreparably squashed and torn...
3	2	625	2_visible cities_cities story_hidden cities_ci...	[visible cities, cities story, hidden cities, ...	[Absolutely failing to even piece together a f...
4	3	523	3_kublai describing cities_mongol emperor kubl...	[kublai describing cities, mongol emperor kubl...	[back then i thought this was a beautifully w...
...	...	...	...	...	...
99	98	21	98_cities lost little_cities stay lost_cities ...	[cities lost little, cities stay lost, cities ...	[Or perhaps, speaking of other cities, I have...
100	99	21	99_city traveller finds_new city traveller_cit...	[city traveller finds, new city traveller, cit...	["Arriving at each new city, the traveler find...
101	100	21	100_men dreamed chasing_men dreams chasing_pat...	[men dreamed chasing, men dreams chasing, path...	[They laid the streets, each following his pu...
102	101	20	101_true exist hypothesis_exist hypothesis tru...	[true exist hypothesis, exist hypothesis true,...	[So the other hypothesis is true: they exist ...
103	102	20	102_imaginary city uncovering_illusory metropo...	[imaginary city uncovering, illusory metropoli...	[It may be that he is creating them all out o...

Figure 22: first result of BERTopic on the review dataset (Customed stop word list

include:
'travels','literature','italo','travel','calvinos','calvin','calvino','reading','read','novel','books
','book','venice','invisible','polo','marco','khan','polos')

Figure 22 shows an appropriate topic result, but still have high outlier, through the built-in reduce outlier function of this library, Figure 23 shows the final topic distribution.



	Topic	Count	Name	Representation	Representative_Docs
0	-1	37	-1_read_over_book_shoutout	[read, over, book, shoutout, rec, auris, tranq...	[And the mind refuses to accept more faces, m...
1	0	905	0_calvino_his_to_is	[calvino, his, to, is, read, and, this, that, ...	[Imagine, if you will, a swirling, dynamic mix...
2	1	817	1_it_didn_but_just	[it, didn, but, just, was, me, get, really, do...	[It is becoming irreparably squashed and torn...
3	2	662	2_invisible_cities_is_of	[invisible, cities, is, of, to, in, and, as, t...	[Absolutely failing to even piece together a f...
4	3	659	3_kublai_marco_khan_polo	[kublai, marco, khan, polo, visited, cities, h...	[back then i thought this was a beautifully w...
...	...	...	...	...	...
99	98	195	98_little_lost_speaking_already	[little, lost, speaking, already, bit, perhaps...	[Or perhaps, speaking of other cities, I have...
100	99	172	99_longer_lies_foreign_wait	[longer, lies, foreign, wait, possess, arrivin...	["Arriving at each new city, the traveler find...
101	100	92	100_streets_built_men_woman	[streets, built, men, woman, walls, dream, arc...	[They laid the streets, each following his pu...
102	101	139	101_exist_they_true_real	[exist, they, true, real, hypothesis, do, norm...	[So the other hypothesis is true: they exist ...
103	102	86	102_venice_he_describing_mind	[venice, he, describing, mind, his, creating, ...	[It may be that he is creating them all out o...

Figure 23: result of BERTopic on the review dataset after reduction outlier

In terms of topic count, Figure 24 shows an overview. 'Quotation' comes out on top with 25 clusters, accounting for about 17.8 percent of all review sentences, with 3,247 related sentences. 'Recommend Related' topics are the second most frequent with 21 clusters and 3,392 related sentences, making up about 18.8 percent of all review sentences. 'Genre Related' topics have 20 themes, but with slightly more related sentences at 3,613, accounting for about 20 percent of all review sentences. Both 'Structure Related' and 'Personal Reflection Related' themes have 14 clusters each. 'Structure Related' topics have the highest count of related sentences at 3,732, making up about 21 percent of all review sentences. 'Personal Reflection' related sentences total 3,076, accounting for about 17 percent of all review sentences.

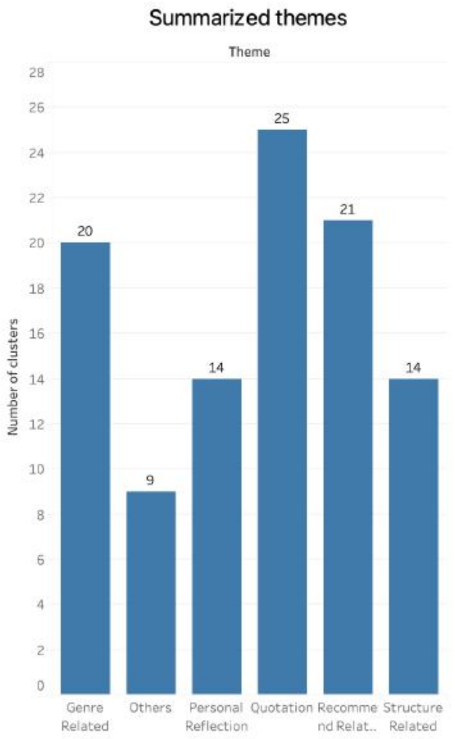


Figure 24: bar chart of summarized themes

The analysis and summarization of the topics reveal a relatively even distribution across the five themes and all themes have unearthed valuable information that warrants further exploration. This thesis will continue use topic modelling method to analyze 'Genre related', 'Personal reflection' and 'Recommend related' themes.

In addition to topic modelling, there are two additional avenues of analytics that inspired by these results. First, the 'Genre Related' theme mentioned other authors who share a similar style, prompting a social network analysis. This analysis leveraged supplementary information from Wikipedia (for instance, whether other writers within the network were mentioned) to establish connections between various authors.

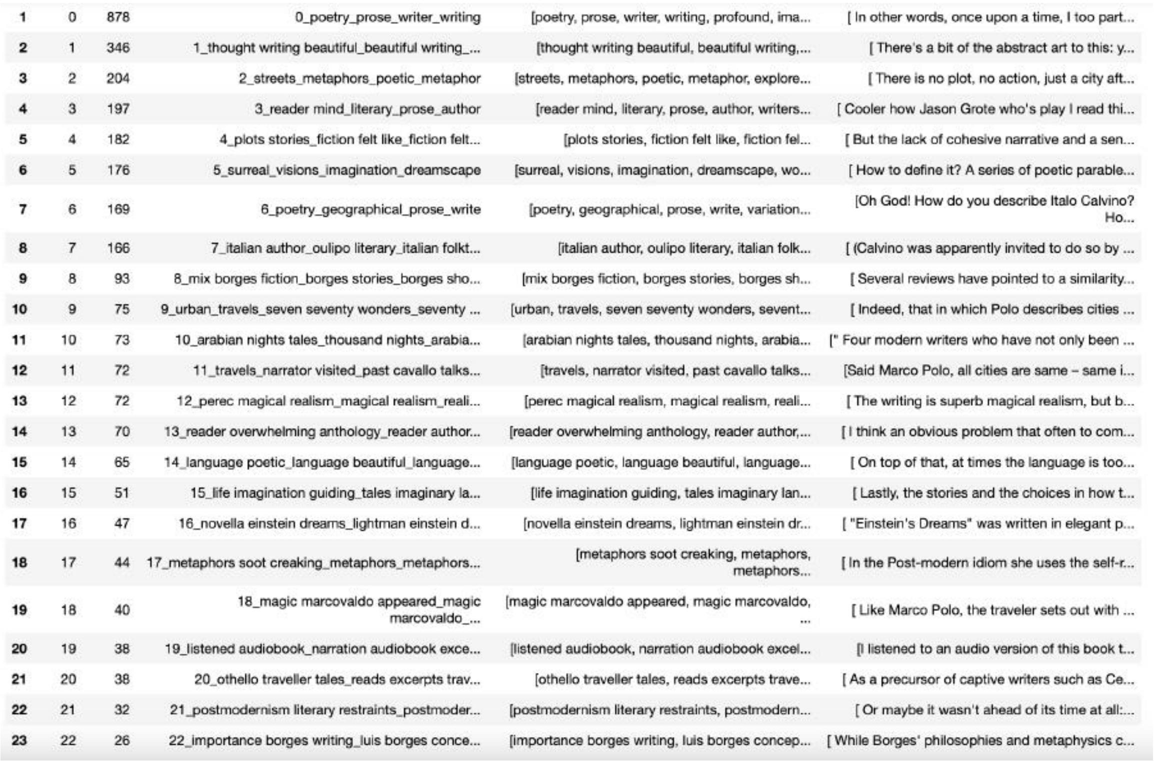
The second avenue is related to quotation. Using the fuzzy approach to process all comments and matching them with the original text, identifying the sentences that were quoted most frequently. These two exploratory paths will be elaborated on in the following sections.

4.3.3 Theme analytics

Through the overall topic modeling, there are five themes are determined, subsequently, this part will do further topic modeling based on concluded sentences to selected themes (taking the same steps with book analytic). Also, include a social network analysis in ‘genre related’ theme and a quotation analysis inspired by ‘Quotation’ theme.

Genre related:

The BERTopic analysis yielded a total of 23 clusters in this theme. After deploying the built-in outlier reduction function, the refined results are depicted in Figure 25. Upon close examination and summarization, seven distinct sub-themes emerged, namely: 'Poetic Compelling', 'Imagination & Metaphors', 'Borges Style', '1001 Arabian Nights', 'Magical Realism', 'Einstein’s Dreams', and 'Postmodernism'.



1	0	878	0_poetry_prose_writer_writing	[poetry, prose, writer, writing, profound, ima...	[In other words, once upon a time, I too part...
2	1	346	1_thought writing beautiful_beautiful writing_...	[thought writing beautiful, beautiful writing,...	[There's a bit of the abstract art to this: y...
3	2	204	2_streets_metaphors_poetic_metaphor	[streets, metaphors, poetic, metaphor, explore...	[There is no plot, no action, just a city aft...
4	3	197	3_reader mind_literary_prose_author	[reader mind, literary, prose, author, writers...	[Cooler how Jason Grote who's play I read thi...
5	4	182	4_plots stories_fiction felt like_fiction felt...	[plots stories, fiction felt like, fiction fel...	[But the lack of cohesive narrative and a sen...
6	5	176	5_surreal_visions_imagination_dreamscape	[surreal, visions, imagination, dreamscape, wo...	[How to define it? A series of poetic parable...
7	6	169	6_poetry_geographical_prose_write	[poetry, geographical, prose, write, variation...	[Oh God! How do you describe Italo Calvino? Ho...
8	7	166	7_italian author_oulipo literary_italian folkt...	[italian author, oulipo literary, italian folk...	[(Calvino was apparently invited to do so by ...
9	8	93	8_mix borges fiction_borges stories_borges sho...	[mix borges fiction, borges stories, borges sh...	[Several reviews have pointed to a similarity...
10	9	75	9_urban_travels_seven seventy wonders_seventy ...	[urban, travels, seven seventy wonders, sevent...	[Indeed, that in which Polo describes cities ...
11	10	73	10_arabian nights tales_thousand nights_arabia...	[arabian nights tales, thousand nights, arabia...	[" Four modern writers who have not only been ...
12	11	72	11_travels_narrator visited_past cavallo talks...	[travels, narrator visited, past cavallo talks...	[Said Marco Polo, all cities are same – same i...
13	12	72	12_perec magical realism_magical realism_reali...	[perec magical realism, magical realism, reali...	[The writing is superb magical realism, but b...
14	13	70	13_reader overwhelming anthology_reader author...	[reader overwhelming anthology, reader author,...	[I think an obvious problem that often to com...
15	14	65	14_language poetic_language beautiful_language...	[language poetic, language beautiful, language...	[On top of that, at times the language is too...
16	15	51	15_life imagination guiding_tales imaginary la...	[life imagination guiding, tales imaginary lan...	[Lastly, the stories and the choices in how t...
17	16	47	16_novella einstein dreams_lightman einstein d...	[novella einstein dreams, lightman einstein dr...	["Einstein's Dreams" was written in elegant p...
18	17	44	17_metaphors soot creaking_metaphors_metaphors...	[metaphors soot creaking, metaphors, metaphors...	[In the Post-modern idiom she uses the self-r...
19	18	40	18_magic marcovaldo appeared_magic marcovaldo_...	[magic marcovaldo appeared, magic marcovaldo, ...	[Like Marco Polo, the traveler sets out with ...
20	19	38	19_listened audiobook_narration audiobook exce...	[listened audiobook, narration audiobook excel...	[I listened to an audio version of this book t...
21	20	38	20_othello traveller tales_reads excerpts trav...	[othello traveller tales, reads excerpts trave...	[As a precursor of captive writers such as Ce...
22	21	32	21_postmodernism literary restraints_postmoder...	[postmodernism literary restraints, postmodern...	[Or maybe it wasn't ahead of its time at all:...
23	22	26	22_importance borges writing_luis borges conce...	[importance borges writing, luis borges concep...	[While Borges' philosophies and metaphysics c...

Figure 25: result of BERTopic on the ‘Genre Related’ theme

Among these topics, ‘Poetic Compelling’ and ‘Imagination & Metaphors’ do reflect author’s writing style. Other topics note more straightforward writing genre, like ‘Magical Realism’ and ‘Postmodernism’ as well.

Furthermore, two specific topics directly reference the author 'Jorge Luis Borges' and the book 'Einstein’s Dreams'. This observation paved the way for a deeper exploration into the network of authors mentioned in the reviews. To achieve this, a combination of Named-entity recognition (NER) models, web scraping tools, and Gephi was employed. The initial step involved using the NER model to identify all the names cited in the reviews. The results were initially broad due to variations in how reviewers referenced the same individual; for instance, both 'Borges' and 'Jorge Luis Borges' indicate the same author. After refining and tallying the mentions, the final list consisted of 26 related authors, as shown in the Figure 26.

Apart from Jorge Luis Borges and Alan Lightman (the author of 'Einstein’s Dreams'), several authors renowned for their stream-of-consciousness writing technique, like Virginia Woolf and Marcel Proust, were frequently mentioned. While "Invisible Cities" might not fit squarely within the stream-of-consciousness genre, its capacity to ignite readers' imaginations and breaks temporal and spatial boundaries resonate with the genre's essence.

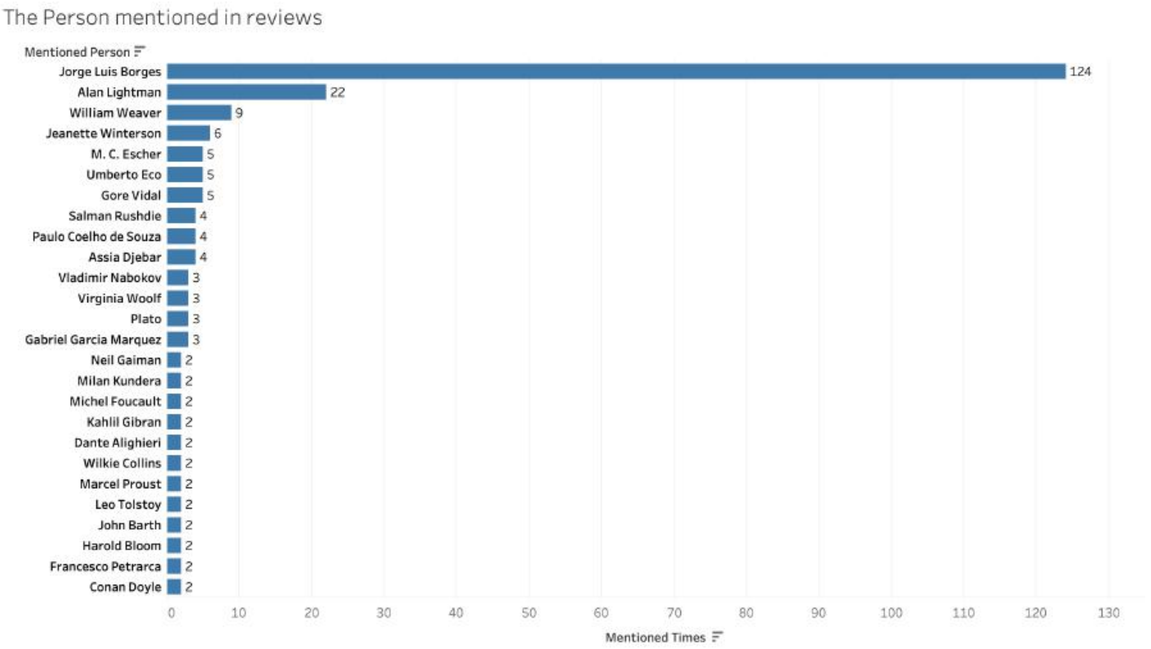
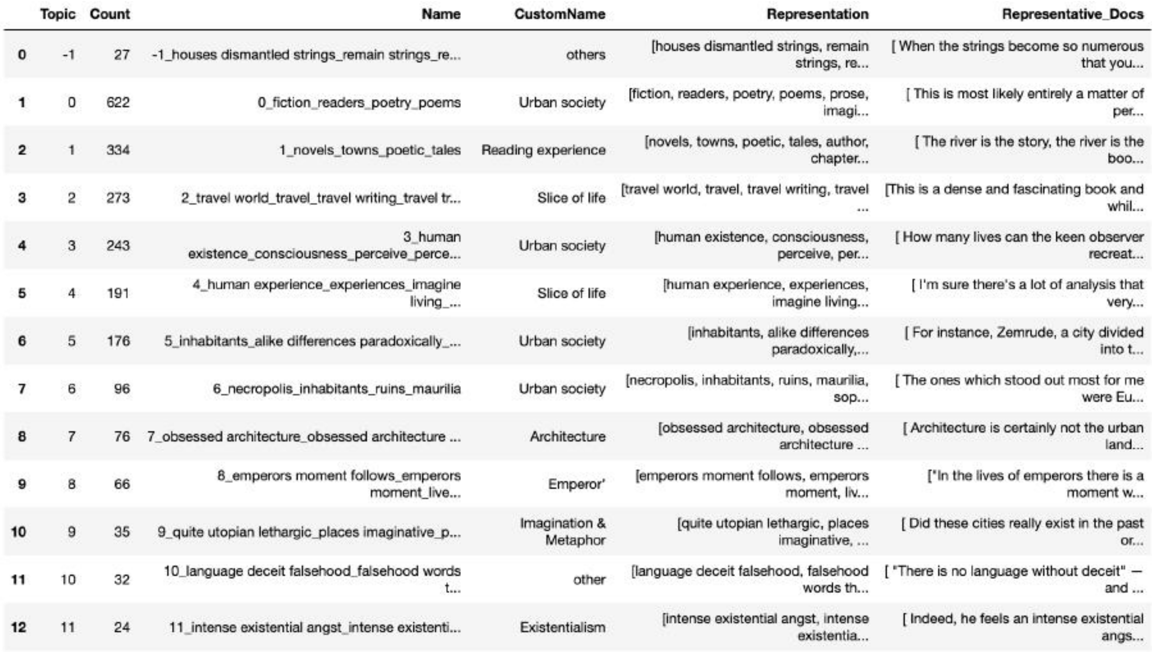


Figure 26: NER result of mentioned person in genre related theme

Personal reflections related:

There are 12 clusters identified in this theme after the topic modeling processing, and eight of them are potential topics (Figure 27). These are: ‘Urban society’, ‘Slice of life’, ‘Reading experience’, ‘Architecture’, ‘Emperor’, ‘Imagination & Metaphor’, and ‘Existentialism’. ‘Urban society’ and ‘Slice of life’ are the majority of topics. They share the similarity, but the main reason that divides them into two topics because sentences in ‘Urban society’ remind urban life and reflect society; sentences in ‘Slice of Life’ are general life reflections, which could include a rethinking of the current way of living but not limited in urban area.



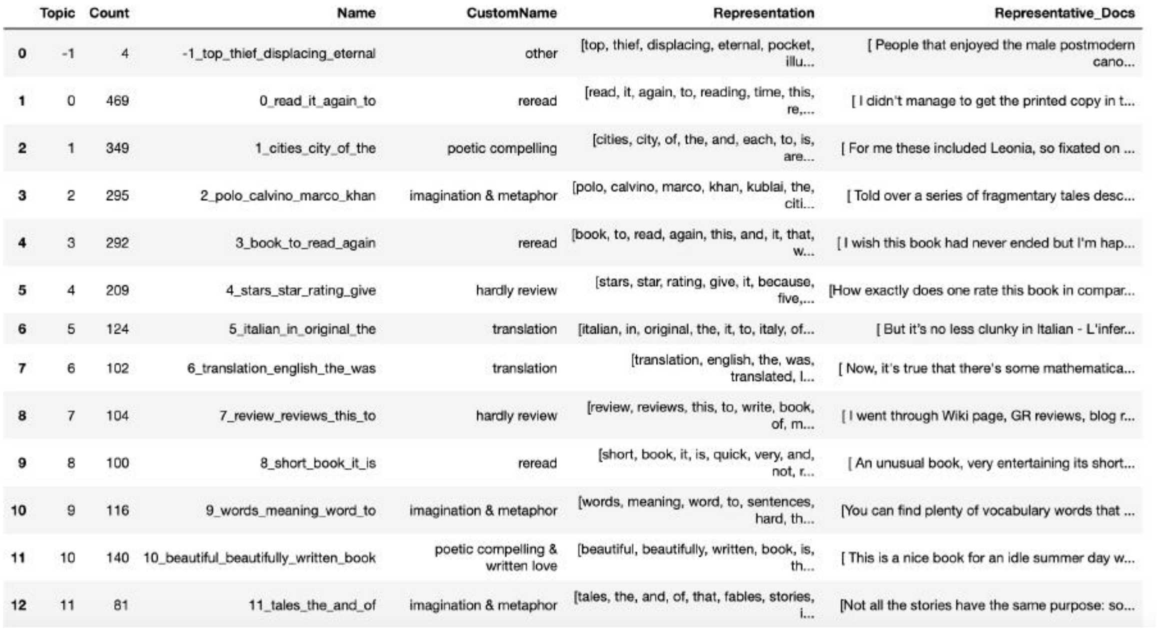
	Topic	Count	Name	CustomName	Representation	Representative_Docs
0	-1	27	-1_houses dismantled strings_remain strings_re...	others	[houses dismantled strings, remain strings, re...	[When the strings become so numerous that you...
1	0	622	0_fiction_readers_poetry_poems	Urban society	[fiction, readers, poetry, poems, prose, imagi...	[This is most likely entirely a matter of per...
2	1	334	1_novels_towns_poetic_tales	Reading experience	[novels, towns, poetic, tales, author, chapter...	[The river is the story, the river is the boo...
3	2	273	2_travel world_travel_travel writing_travel tr...	Slice of life	[travel world, travel, travel writing, travel ...	[This is a dense and fascinating book and whil...
4	3	243	3_human existence_consciousness_perceive_perce...	Urban society	[human existence, consciousness, perceive, per...	[How many lives can the keen observer recreat...
5	4	191	4_human experience_experiences_imagine living_...	Slice of life	[human experience, experiences, imagine living...	[I'm sure there's a lot of analysis that very...
6	5	176	5_inhabitants_alike differences paradoxically_...	Urban society	[inhabitants, alike differences paradoxically,...	[For instance, Zemrude, a city divided into t...
7	6	96	6_necropolis_inhabitants_ruins_maurilia	Urban society	[necropolis, inhabitants, ruins, maurilia, sop...	[The ones which stood out most for me were Eu...
8	7	76	7_obsessed architecture_obsessed architecture ...	Architecture	[obsessed architecture, obsessed architecture ...	[Architecture is certainly not the urban land...
9	8	66	8_emperors moment follows_emperors moment_live...	Emperor'	[emperors moment follows, emperors moment, liv...	["In the lives of emperors there is a moment w...
10	9	35	9_quite utopian lethargic_places imaginative_p...	Imagination & Metaphor	[quite utopian lethargic, places imaginative, ...	[Did these cities really exist in the past or...
11	10	32	10_language deceit falsehood_falsehood words t...	other	[language deceit falsehood, falsehood words th...	["There is no language without deceit" — and ...
12	11	24	11_intense existential angst_intense existenti...	Existentialism	[intense existential angst, intense existentia...	[Indeed, he feels an intense existential angs...

Figure 27: result of BERTopic on the ‘Personal reflections’ theme

This outcome resonates with the Calvino’s intent, emphasizing the importance of individuals recognizing the nuances of their daily lives rather than merely existing within the overarching narratives typical of modernism. During the 1960s, advancements in massive construction techniques within the field of civil engineering catalyzed rapid growth in the real estate industry, leading to transformative changes in urban living. It was during this time that a few forward-thinking scholars began to appreciate the significance of the social dimension in urban planning. They initiated studies exploring the varied ways in which individuals perceive urban environments. Seminal works from this era include Jane Jacobs's "The Death and Life of Great American Cities" (1961) and Kevin Lynch's "The Image of The City" (1960). These groundbreaking publications critiqued the prevailing approach to urban planning that predominantly focused on its physical aspects. Today, they stand as monumental works in the realm of urban planning. Calvino, wearing his editor's hat, was deeply influenced by these transformative texts during that period.

Recommend Related:

This theme yielded 23 clusters through topic modeling (Figure 28), with seven emerging as prominent topics: 'Poetic Compelling', 'Imagination & Metaphor', 'Re-read', 'Difficult to Comment', 'Translation', 'Favorite', and 'Unusual'. Among these, 'Poetic Compelling' and 'Imagination & Metaphor' stand out, acting as defining characteristics of this work that are frequently highlighted in recommendations. 'Re-read' and 'Difficult to Comment' also constitute a significant portion. Modena (2011) points out that Calvino envisioned readers revisiting the book multiple times, each time entering it from a different angle, initiating varied journeys. The 'Difficult to Comment' topic underscores the book's unique non-linear structure and its expansive imaginative scope, which can make it challenging to encapsulate in a straightforward review.



	Topic	Count	Name	CustomName	Representation	Representative_Docs
0	-1	4	-1_top_thief_displacing_eternal	other	[top, thief, displacing, eternal, pocket, illu...	[People that enjoyed the male postmodern cano...
1	0	469	0_read_it_again_to	reread	[read, it, again, to, reading, time, this, re,...	[I didn't manage to get the printed copy in t...
2	1	349	1_cities_city_of_the	poetic compelling	[cities, city, of, the, and, each, to, is, are...	[For me these included Leonia, so fixated on ...
3	2	295	2_polo_calvino_marco_khan	imagination & metaphor	[polo, calvino, marco, khan, kublai, the, citi...	[Told over a series of fragmentary tales desc...
4	3	292	3_book_to_read_again	reread	[book, to, read, again, this, and, it, that, w...	[I wish this book had never ended but I'm hap...
5	4	209	4_stars_star_rating_give	hardly review	[stars, star, rating, give, it, because, five,...	[How exactly does one rate this book in compar...
6	5	124	5_italian_in_original_the	translation	[italian, in, original, the, it, to, italy, of...	[But it’s no less clunky in Italian - L'infer...
7	6	102	6_translation_english_the_was	translation	[translation, english, the, was, translated, l...	[Now, it's true that there's some mathematica...
8	7	104	7_review_reviews_this_to	hardly review	[review, reviews, this, to, write, book, of, m...	[I went through Wiki page, GR reviews, blog r...
9	8	100	8_short_book_it_is	reread	[short, book, it, is, quick, very, and, not, r...	[An unusual book, very entertaining its short...
10	9	116	9_words_meaning_word_to	imagination & metaphor	[words, meaning, word, to, sentences, hard, th...	[You can find plenty of vocabulary words that ...
11	10	140	10_beautiful_beautifully_written_book	poetic compelling & written love	[beautiful, beautifully, written, book, is, th...	[This is a nice book for an idle summer day w...
12	11	81	11_tales_the_and_of	imagination & metaphor	[tales, the, and, of, that, fables, stories, i...	[Not all the stories have the same purpose: so...

Figure 28: result of BERTopic on the ‘Recommend Related’ theme

Quotation:

Inspired by the quotation, instead of doing topic modeling, this part will do fuzzy string matching to find the most quoted sentence in reviews. In this analysis, the whole book is considered, not only the 55 narrations but also the dialogs.

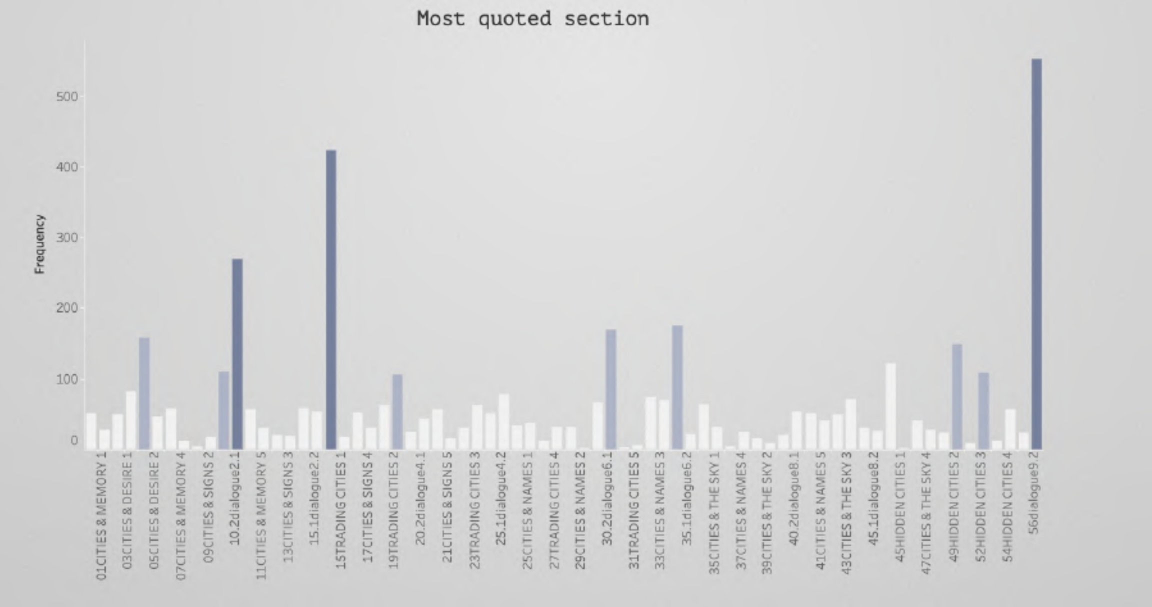


Figure 29: Bar chart of most cited section

The visualization of analytics results, represented through bar graphs, unveiled an interesting trend: dialogues emerged as the most frequently cited sections, surpassing the narratives about cities. Distinctively, three sections where citations exceeded 200 instances stood out: the beginning dialogue of Chapter 2, the beginning dialogue of Chapter 3, and the book's concluding dialogue (Figure 29). These recurrently cited passages, which convey the Calvino’s reflections more transparently than abstract metaphors, offer insights into the book's underlying tenets. An intriguing interpretation could be discerned from this trend: perhaps Marco Polo embodies Calvino, while Khan symbolizes the reader. If so, the nine dialogues can be viewed as Calvino's interactions with the audience of his time. This perspective offers a fresh lens to comprehend the essence and arrangement of these dialogues.

Discussion:

By employing clustering strategies for topic modeling on the review sentences, the users can delve deeper to extract more insightful information, even though this approach requires users to invest additional time and rely on prior knowledge to summarize the clusters. Using BERTopic on the review dataset, the initial analysis identified five themes within the reviews in which the representative documents of clusters helped that. Reading through these documents provided richer insights that traditional modeling based on the Bag of Words (BoW) couldn't offer. In a more granular analysis, combining representative documents and words proved more intuitive in understanding the topics than using representative words alone. For instance, in the analysis related to genre, terms like "Einstein" or "Einstein’s Dreams" frequently appeared. However, without the context provided by the documents, it would be challenging to discern that these refer to a book, leading to potential misunderstandings.

4.4 Visualization of the review's text analytics results

This part will only focus on visualization strategy of the review's text analytics results. Different with the visualization strategy on topic determined from book, this part adopt Sankey Diagram to convey insight to reader. Since the analytics reflect a clear structure of reviews, Sankey Diagram could illustrate it through the flow between the nodes. Also, through changing color to highlight nodes and flow could help narrative to better convey information to reader. It's crucial to emphasize that the numbers associated with each theme and topic in the graph do not directly correspond to specific reviews. Since the analysis is conducted on a sentence-by-sentence basis rather than per review, these figures are derived by calculating the proportion of relevant sentences and then multiplying by the total number of reviews.

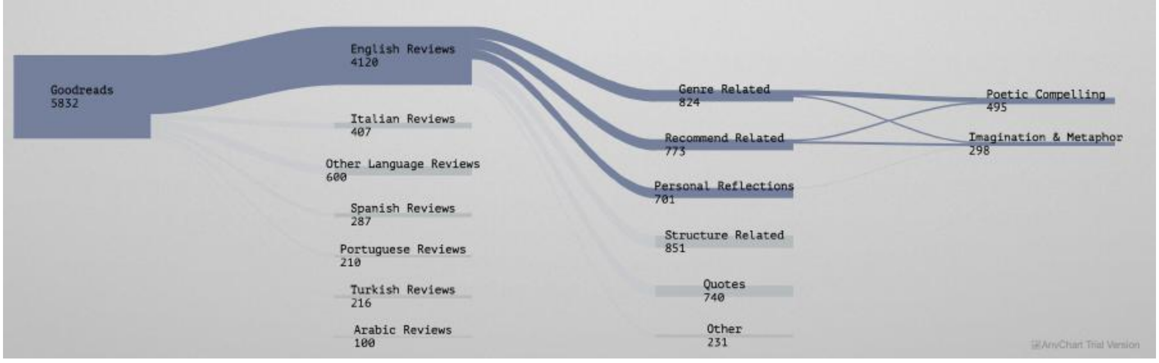


Figure 30: The themes of user reviews based on topic modeling and the key topic to be emphasized in the overview stage

Before introducing the detailed topics from each theme, Figure 30 illustrate the analytics process. Commencing from the left, the diagram displays the entire dataset, narrowing its focus onto the English reviews. From these reviews, five distinct themes have been summarized. Among them, three themes are emphasized for a more detailed analysis. Notably, 'Poetic Compelling' and 'Imagination & Metaphor' emerge as recurrent topics. These two have been prioritized for interpretation due to their dominance and their potential to provide readers with a deeper comprehension of the book.

Figure 31 illustrates the analytics results of 'Genre related' reviews. Beyond the two predominant topics, the remaining topics concern writing style and have been underscored. This analysis subsequently culminates in a network graphic (Figure 32). As depicted in Figure 32, each node represents an individual mentioned in the reviews, with the size of each node indicating the frequency of mentions. In addition to the links between Calvino and those mentioned, links are also established by scraping the Wikipedia pages of all individuals. If there's a mention, a link is created between the two individuals. After acquiring the data, this network graphic is initially generated using Gephi and later transitioned to Tableau. This allows for the incorporation of interactive features such as hover and click, details of which will be expounded upon in the subsequent chapter.

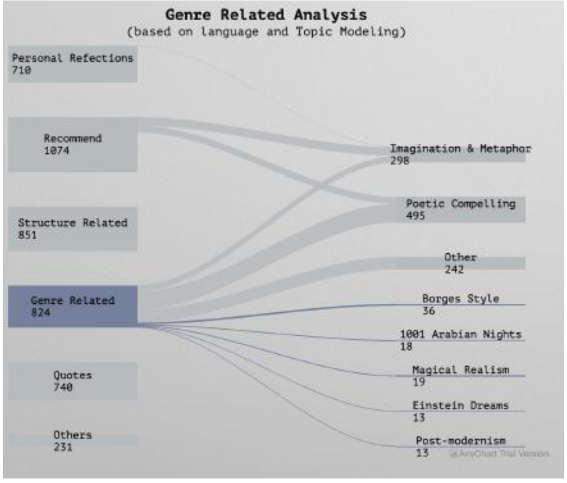


Figure 31: Sankey Diagram of Genre related analytics

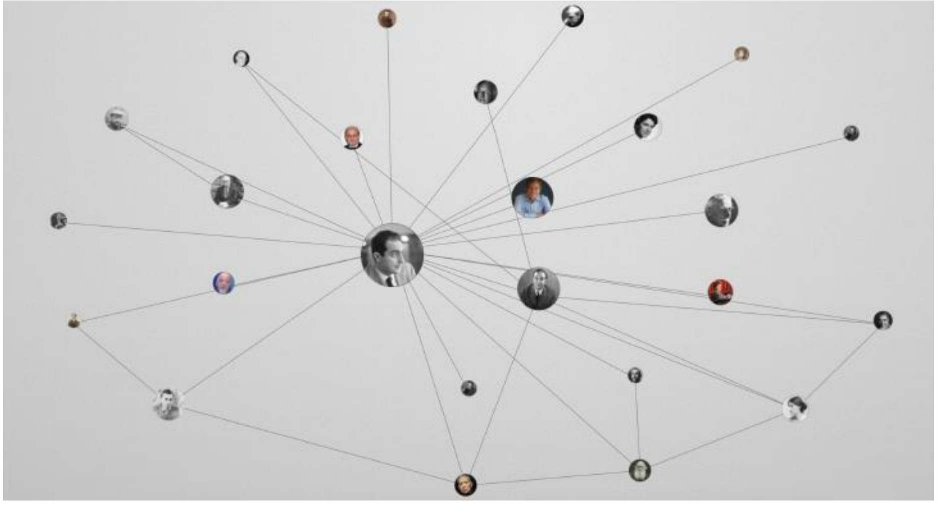


Figure 32: network of mentioned person

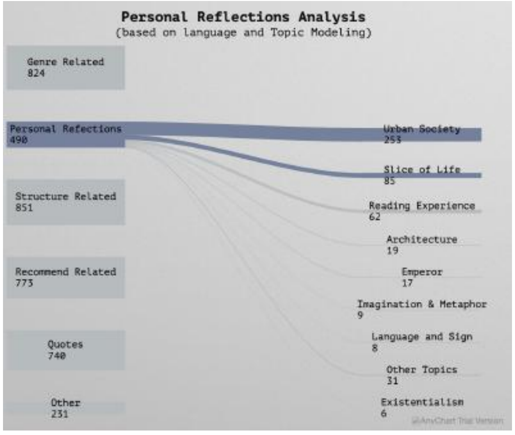


Figure 33: Sankey Diagram of Personal reflections analytics

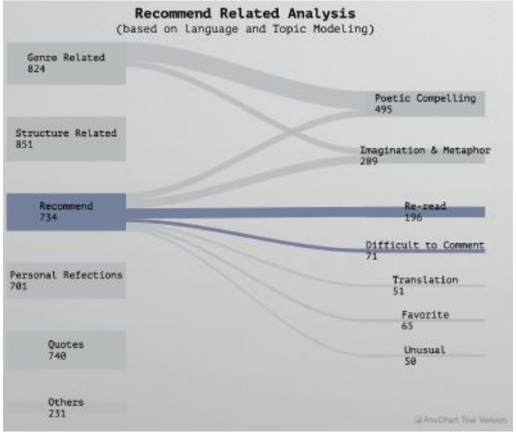


Figure 34: Sankey Diagram of Recommend related analysis

Figure 33 and Figure 34 take the same strategy to present analytics results. In Figure 33, ‘Urban Society’ and ‘Slice of Life’ are highlighted, and ‘Re-read’ and ‘Difficult to Comment’ are highlighted in Figure 34. These highlighted topics are detailed interpreted in the narrative and discussed in the sequent chapter.

For extra findings, like the most quoted section (Figure 29), the histogram was used for visualization. This kind of simple strategy directly shows information, companying with gradual color to reflect hierarchy.

Chapter 5

Communicate with reader

This part will focus on how to do an appropriate conversion of text analytic into graphics and communicating with audience by digital storytelling. The result will be presented on a website combining with images, text and graphics, and target audience is set for funs of this book and Italo Calvino.

This website consists of five parts to form this storytelling, including ‘Introduction’, ‘Life Map’, ‘Book’, ‘Reviews’ and ‘Unexpected Discovery’. As a complete storytelling, in addition to book and review analytics, Introduction and Life Map are also included for extra background information to help user understand Calvino’s relevant experience and the context of the times. ‘Unexpected Discovery’ is designed to present some extra findings during the text mining.

The whole website employs a navigational approach that allows users to directly read different parts. Instead of scroll-down narrative, this kind of structure aims to provide a flexible experience. When users navigate from page to page, the navigation bar located in the upper right corner highlights the current page to guide the reader's location.

5.1 Website design principle

User Experience (UI) and coherence

The website aims to evoke a vintage ambiance, utilizing a dark palette complemented by a silver-gray background and typewriter-inspired fonts. Each button, fashioned after a typewriter key, further enhances this retro feel. In terms of color scheme, the text primarily adopts an ink-black hue, while infographics leverage shades of red and muted blue to spotlight key details, ensuring a cohesive and harmonious overall design aesthetic.

Interaction and Sequence

Most of the content on the website is not presented in a scroll-down manner but is narrated progressively through simple interactive buttons. These straightforward interactions include buttons, drop-down options, hover effects, and clicks that redirect to relevant Wikipedia pages. Additionally, some visualizations are created using Tableau and embedded into the website. As a result, Tableau's inherent hover and click interactions are also harnessed to convey information.

Narrative

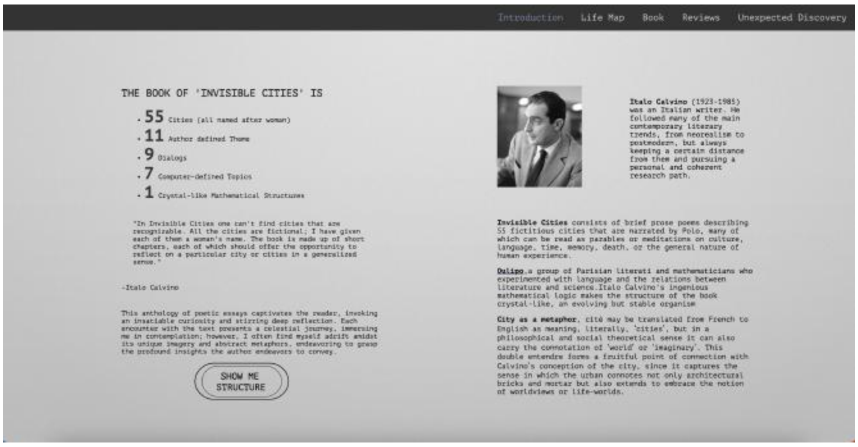
The narration of this website is achieved through a combination of visualization and text. Visualization primarily conveys insights to readers, while text provides more detailed explanations. This approach aims to effectively communicate information visually. Additionally, through thoughtful typography, the text highlights key points and offers external Wikipedia links for specific terms or background knowledge.

5.2 Website

Each page adopts distinct narrative styles and interaction to suit the specific features.

Introduction

On the "Introduction" page, background details of the book and Calvino are presented, paired with visual representations of the analytical results to captivate the user (Figure 35, Figure 36). The design is bifurcated into two columns (Figure 35). The left column offers an overview of the book, encapsulating a mathematical interpretation, Italo Calvino's viewpoint, and the author's perspective. A button anchored at the bottom reveals the mathematical structure of the book, graphically delineating the relationship between the sequence and length of the 55 narratives (Figure 36). Conversely, the right column furnishes details about the author and a succinct elucidation of the thematic essence of the book. Given that "Oulipo" might be an arcane term to some users, an interactive clickable link is provided, redirecting to its Wikipedia page for a comprehensive understanding.



THE BOOK OF 'INVISIBLE CITIES' IS

- 55 Cities (all named after woman)
- 11 Author defined Theme
- 9 Dialogs
- 7 Computer-defined Topics
- 1 Crystal-like Mathematical Structures

"In Invisible Cities one can't find cities that are recognizable. All the cities are fictional; I have given each of them a woman's name. The book is made up of short chapters, each of which should offer the opportunity to reflect on a particular city or cities in a generalized sense."

-Italo Calvino

This anthology of poetic essays captivates the reader, invoking an insatiable curiosity and stirring deep reflection. Each encounter with the text presents a celestial journey, immersing me in contemplation; however, I often find myself adrift amidst its unique imagery and abstract metaphors, endeavoring to grasp the profound insights the author endeavors to convey.

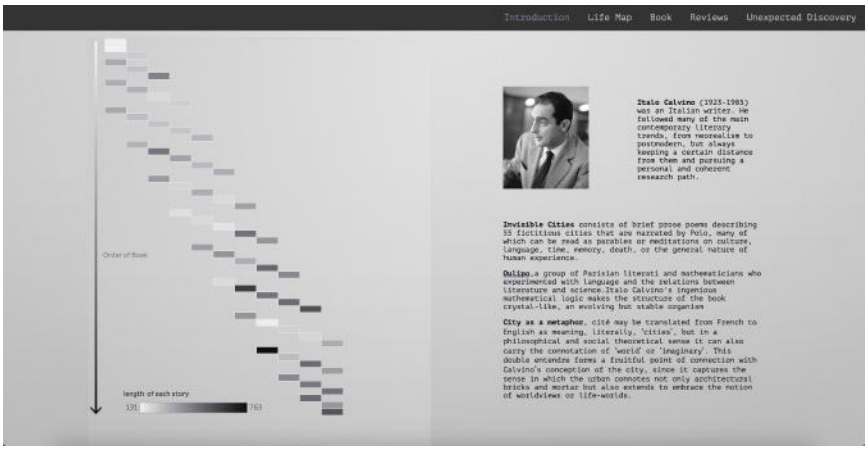
Italo Calvino (1923-1985) was an Italian writer. He followed many of the main contemporary literary trends, from neorealism to postmodern, but always keeping a certain distance from them and pursuing a personal and coherent research path.

Invisible Cities consists of brief prose poems describing 55 fictitious cities that are narrated by Polo, many of which can be read as parables or meditations on culture, language, time, memory, death, or the general nature of human experience.

Oulipo,a group of Parisian literati and mathematicians who experimented with language and the relations between literature and science.Italo Calvino's ingenious mathematical logic makes the structure of the book crystal-like, an evolving but stable organism

City as a metaphor, cité may be translated from French to English as meaning, literally, 'cities', but in a philosophical and social theoretical sense it can also carry the connotation of 'world' or 'imaginary'. This double entendre forms a fruitful point of connection with Calvino's conception of the city, since it captures the sense in which the urban connotes not only architectural bricks and mortar but also extends to embrace the notion of worldviews or life-worlds.

Figure 35: Introduction page (default status)



Italo Calvino (1923-1985) was an Italian writer. He followed many of the main contemporary literary trends, from neorealism to postmodern, but always keeping a certain distance from them and pursuing a personal and coherent research path.

Invisible Cities consists of brief prose poems describing 55 fictitious cities that are narrated by Polo, many of which can be read as parables or meditations on culture, language, time, memory, death, or the general nature of human experience.

Oulipo,a group of Parisian literati and mathematicians who experimented with language and the relations between literature and science.Italo Calvino's ingenious mathematical logic makes the structure of the book crystal-like, an evolving but stable organism

City as a metaphor, cité may be translated from French to English as meaning, literally, 'cities', but in a philosophical and social theoretical sense it can also carry the connotation of 'world' or 'imaginary'. This double entendre forms a fruitful point of connection with Calvino's conception of the city, since it captures the sense in which the urban connotes not only architectural bricks and mortar but also extends to embrace the notion of worldviews or life-worlds.

Figure 36: Introduction page (after click ‘show me structure’ button)

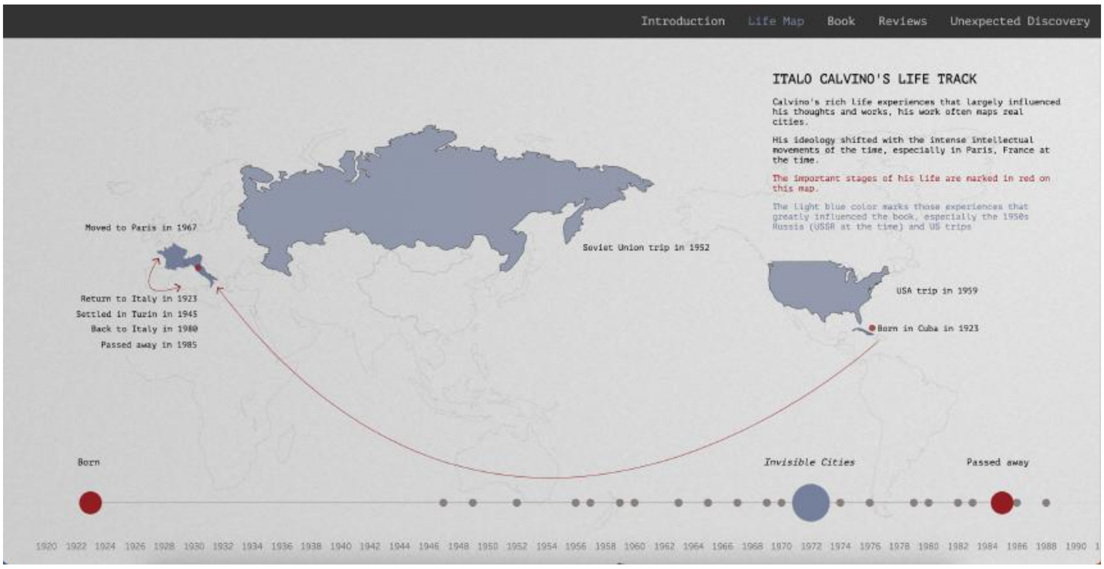
Life Map

The "Life Map" page showcases a world map and timeline that chronicles the pivotal life events and experiences of Italo Calvino, which profoundly influenced the creation of "Invisible Cities". In the top right corner, a brief introduction is provided, accompanied by a color-coded legend; significant life stages are highlighted in red, while experiences that heavily impacted the book are marked in muted blue.

The map pinpoints Calvino's birthplace, his final resting place, and his four relocations. Moreover, Russia and the USA are emphasized on the map, given that his travels to these nations in the early and late 1950s respectively, played a critical role in shaping his literary perspectives. And considering Calvino's image-centric writing style, architectures noted in his diaries are also presented with pictures (this information need to click map and then show up).

At the bottom of the page, spanning across the timeline are milestones that mark the release dates of Calvino's major works and significant events in his life. Notably, the publication date of "Invisible Cities" stands out prominently.

The entire page is meticulously crafted to serve as a cohesive unit. Upon accessing the page, users are immediately drawn to the red lines, which delineate the trajectory of the author's life journey (Figure 37). Following that, a muted blue intertwines France, Italy, the USA, and Russia with "Invisible Cities". As users hover their cursor over the USA or Russia, an interactive hover effect transforms the country's hue to red, signaling its clickability (Figures 38 & 39). Clicking on either of these countries triggers the display of excerpts from Calvino's diary, accompanied by historical images from the corresponding time frame (Figure 40). The objective of this interactive feature is to create a palpable link between Calvino's written narratives and the real-world experiences that inspired them.



ITALO CALVINO'S LIFE TRACK

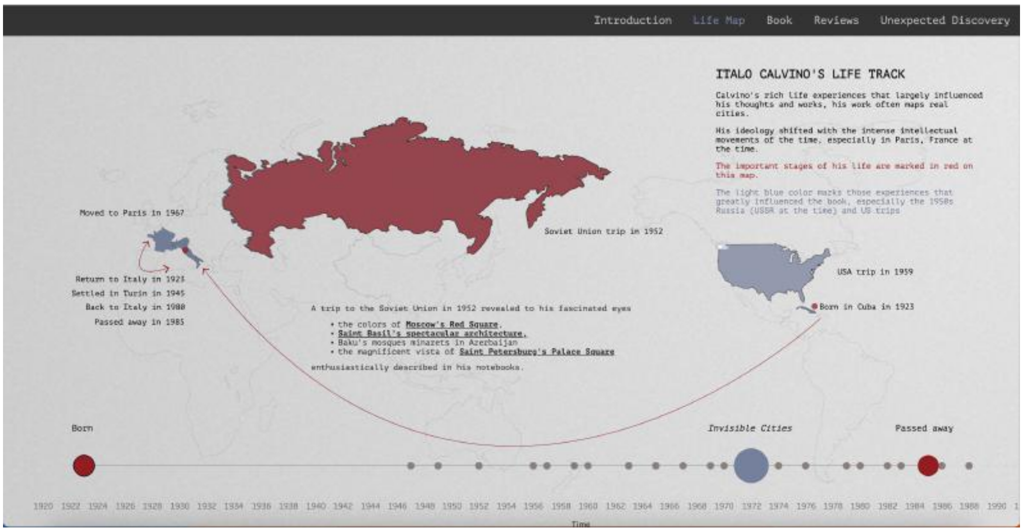
Calvino's rich life experiences that largely influenced his thoughts and works, his work often maps real cities.

His ideology shifted with the intense intellectual movements of the time, especially in Paris, France at the time.

The important stages of his life are marked in red on this map.

The light blue color marks those experiences that greatly influenced the book, especially the 1950s Russia (USSR at the time) and US trips

Figure 37: Life Map page (default status)



ITALO CALVINO'S LIFE TRACK

Calvino's rich life experiences that largely influenced his thoughts and works, his work often maps real cities.

His ideology shifted with the intense intellectual movements of the time, especially in Paris, France at the time.

The important stages of his life are marked in red on this map.

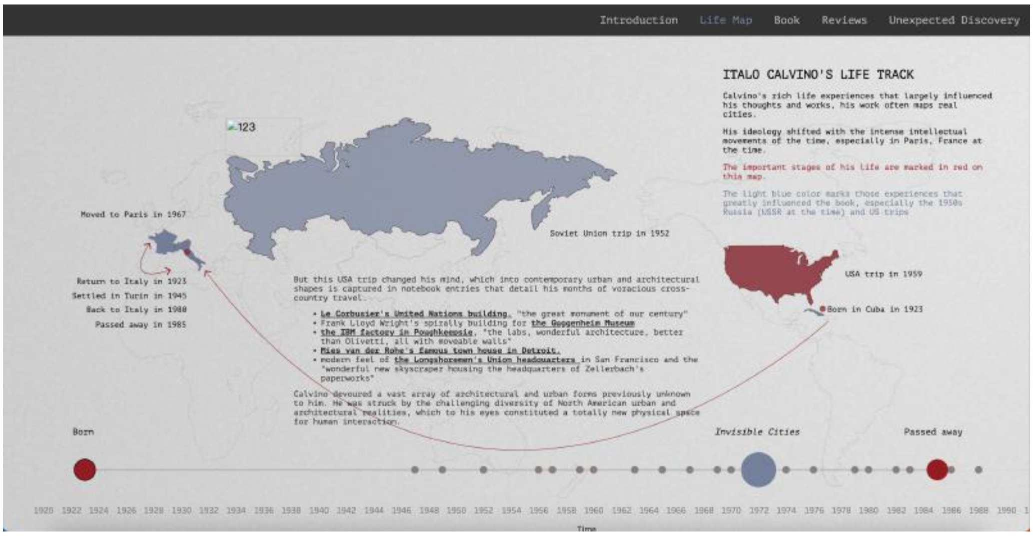
The light blue color marks those experiences that greatly influenced the book, especially the 1950s Russia (USSR at the time) and US trips

A trip to the Soviet Union in 1952 revealed to his fascinated eyes

- the colors of **Moscow's Red Square**,
- **Saint Basil's spectacular architecture**,
- Baku's mosques minarets in Azerbaijan
- the magnificent vista of **Saint Petersburg's Palace Square**

enthusiastically described in his notebooks.

Figure 38: Life Map page (hover Russia and click Russia)



ITALO CALVINO'S LIFE TRACK

Calvino's rich life experiences that largely influenced his thoughts and works, his work often maps real cities.

His ideology shifted with the intense intellectual movements of the time, especially in Paris, France at the time.

The important stages of his life are marked in red on this map.

The light blue color marks those experiences that greatly influenced the book, especially the 1950s Russia (USSR at the time) and US trips

But this USA trip changed his mind, which into contemporary urban and architectural shapes is captured in notebook entries that detail his months of voracious cross-country travel.

- **Le Corbusier's United Nations building**, "the great monument of our century"
- Frank Lloyd Wright's spirally building for **the Guggenheim Museum**
- **the IBM factory in Poughkeepsie**, "the labs, wonderful architecture, better than Olivetti, all with moveable walls"
- **Mies van der Rohe's famous town house in Detroit.**
- modern feel of **the Longshoremen's Union headquarters** in San Francisco and the "wonderful new skyscraper housing the headquarters of Zellerbach's paperworks"

Calvino devoured a vast array of architectural and urban forms previously unknown to him. He was struck by the challenging diversity of North American urban and architectural realities, which to his eyes constituted a totally new physical space for human interaction.

Figure 39: Life Map page (hover US and click US)

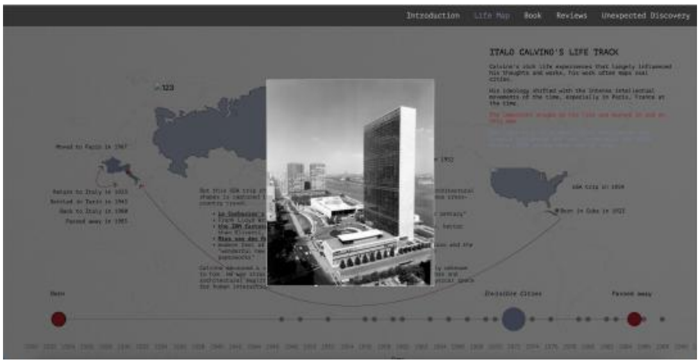


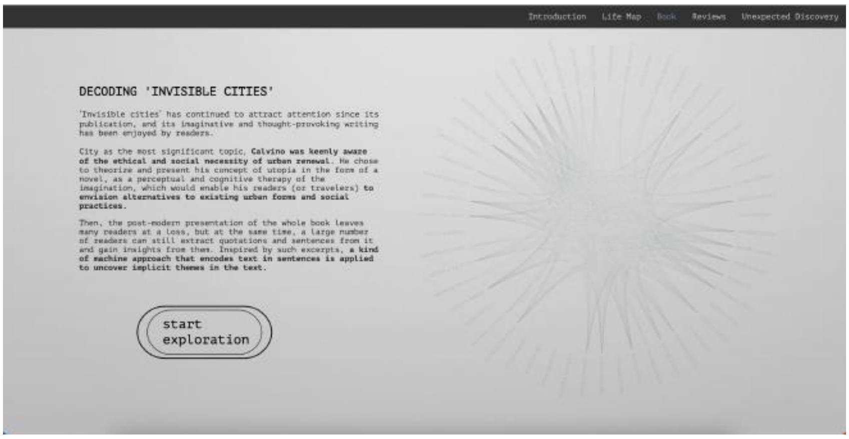
Figure 40: Life Map page (click text then show image of architecture)

Book

The "Book" page is bifurcated into two segments: text and visualization, synergizing to deliver an interactive narrative experience. Beyond the visualization tactics delineated earlier, this section leans heavily into the strategy of user engagement.

Upon landing on the default page (Figure 41), the left segment offers a succinct introduction to the book's predominant theme and the reason why this digital interpretation. At this juncture, the visualization is deliberately subdued, serving a dual purpose: Firstly, it directs users' attention primarily to the text, setting the stage for deeper exploration. Secondly, even if it is set with a kind of degree of transparency, the visualization subtly hints at a complex, interwoven network, piquing the reader's curiosity and kindling their desire to delve deeper.

Activating the button transitions the textual content to display summarized topics and a brief explanation of topic modeling (Figure 17). Accompanied by a dropdown list (Figure 42), users can navigate through various topics. Upon selecting a specific topic, the visualization dynamically adjusts to spotlight the corresponding narratives (Figure 18). Additionally, within this list, options to revert to the default page or access an overview are provided, ensuring users can revisit prior information when needed.



DECODING 'INVISIBLE CITIES'

'Invisible cities' has continued to attract attention since its publication, and its imaginative and thought-provoking writing has been enjoyed by readers.

City as the most significant topic, **Calvino was keenly aware of the ethical and social necessity of urban renewal**. He chose to theorize and present his concept of utopia in the form of a novel, as a perceptual and cognitive therapy of the imagination, which would enable his readers (or travelers) **to envision alternatives to existing urban forms and social practices.**

Then, the post-modern presentation of the whole book leaves many readers at a loss, but at the same time, a large number of readers can still extract quotations and sentences from it and gain insights from them. Inspired by such excerpts, **a kind of machine approach that encodes text in sentences is applied to uncover implicit themes in the text.**

Figure 41: Book page (default status)

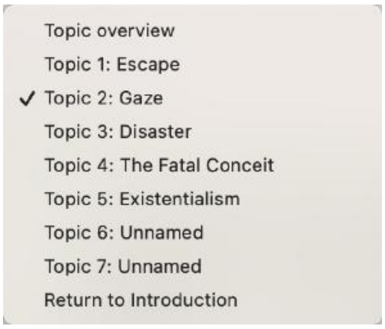


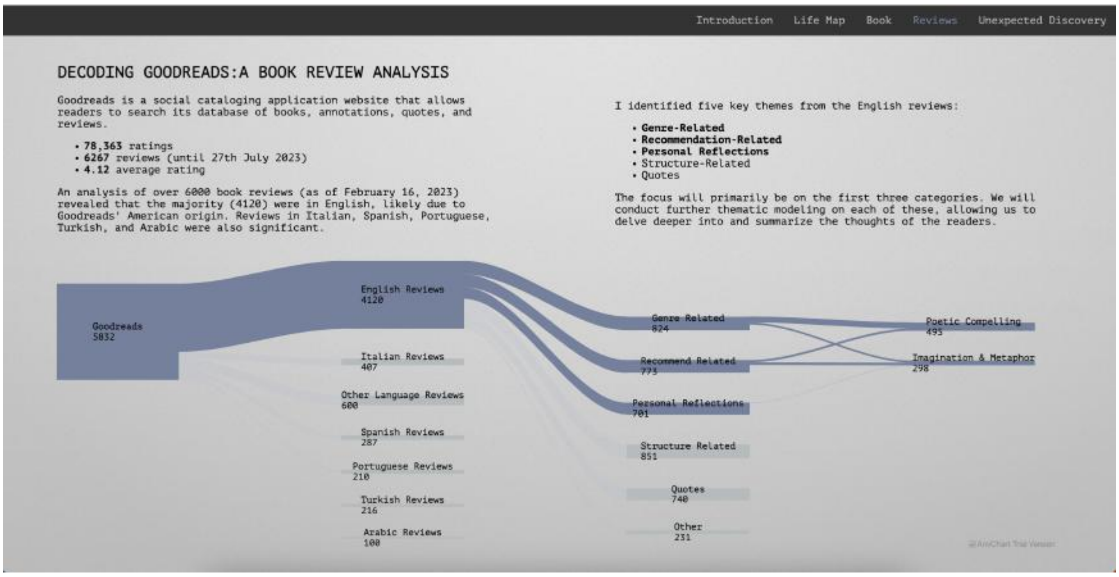
Figure 42: dropdown list

Review:

On the "Review" page, to convey the results of topic modeling more effectively to readers, a scroll-down narrative style has been adopted, steering the storytelling from a broad overview down to specific details, with the Sankey diagram serving as the central narrative thread.

Instead of opting for an interactive chart that allows users to explore, a decision was made to use static visuals. This choice emphasizes the overarching narrative coherence. Through the highlighted flows in the Sankey diagram, the data analysis process and key insights are conveyed to the readers. While presenting the most intriguing findings, other details not explicitly emphasized by the author are also visualized for those readers inclined towards deeper exploration.

The accompanying text elucidates only those parts of the Sankey diagram that are spotlighted, ensuring clarity and focus. Additionally, to aid comprehension, external Wikipedia links are provided for specific terminologies or concepts that readers might be unfamiliar with. This ensures that readers are not left in the dark due to a reluctance to read or unfamiliarity with specific terms.



DECODING GOODREADS:A BOOK REVIEW ANALYSIS

Goodreads is a social cataloging application website that allows readers to search its database of books, annotations, quotes, and reviews.

- **78,363** ratings
- **6267** reviews (until 27th July 2023)
- **4.12** average rating

An analysis of over 6000 book reviews (as of February 16, 2023) revealed that the majority (4120) were in English, likely due to Goodreads' American origin. Reviews in Italian, Spanish, Portuguese, Turkish, and Arabic were also significant.

I identified five key themes from the English reviews:

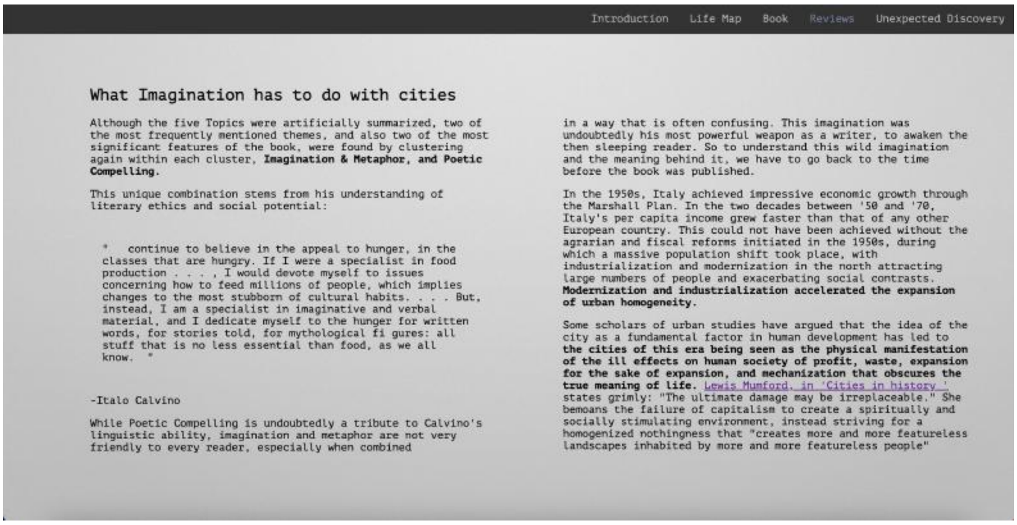
- **Genre-Related**
- **Recommendation-Related**
- **Personal Reflections**
- Structure-Related
- Quotes

The focus will primarily be on the first three categories. We will conduct further thematic modeling on each of these, allowing us to delve deeper into and summarize the thoughts of the readers.

Figure 43: Reviews page (first section: overview)

At the outset, the screen is bifurcated into two sections: text and the Sankey diagram. This overview using the Sankey diagram emphasizes the overall analytic framework and underscores two pivotal summarized topics from the three selected themes (Figure 43). The textual content serves as a brief introduction, complementing the Sankey diagram by accentuating the three themes set for discussion. Additionally, it spotlights the rating overview of the book on the Goodreads platform.

When the user scrolls down, the next screen will provide a detailed explanation about these two highlighted topics (Figure 44). This explanation focuses on how Calvino explains imagination and what kind of social issue he wants the public to be aware of through this imagination. In this part, some key words are highlighted to avoid people losing interest and missing information.



What Imagination has to do with cities

Although the five Topics were artificially summarized, two of the most frequently mentioned themes, and also two of the most significant features of the book, were found by clustering again within each cluster, **Imagination & Metaphor, and Poetic Compelling.**

This unique combination stems from his understanding of literary ethics and social potential:

> " continue to believe in the appeal to hunger, in the classes that are hungry. If I were a specialist in food production . . . , I would devote myself to issues concerning how to feed millions of people, which implies changes to the most stubborn of cultural habits. . . . But, instead, I am a specialist in imaginative and verbal material, and I dedicate myself to the hunger for written words, for stories told, for mythological figures: all stuff that is no less essential than food, as we all know. "

-Italo Calvino

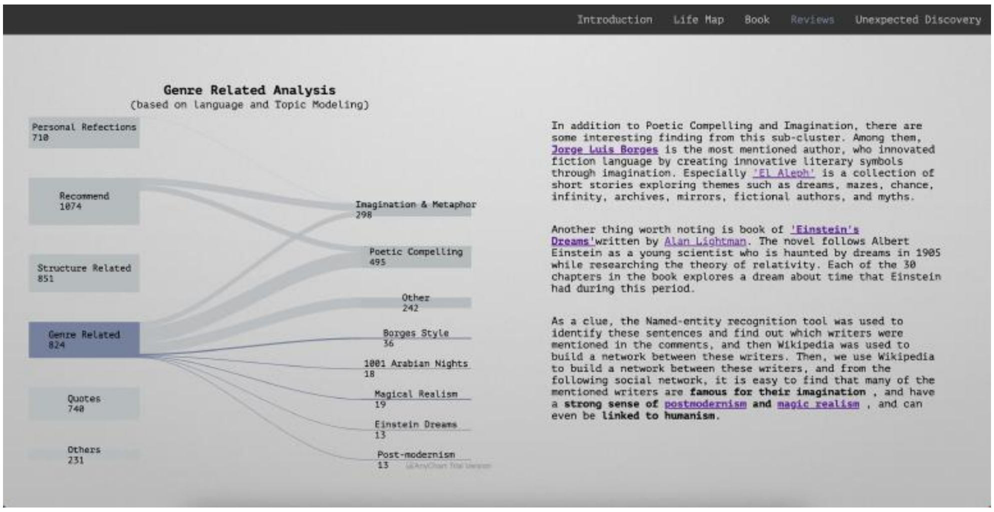
While Poetic Compelling is undoubtedly a tribute to Calvino's linguistic ability, imagination and metaphor are not very friendly to every reader, especially when combined in a way that is often confusing. This imagination was undoubtedly his most powerful weapon as a writer, to awaken the then sleeping reader. So to understand this wild imagination and the meaning behind it, we have to go back to the time before the book was published.

In the 1950s, Italy achieved impressive economic growth through the Marshall Plan. In the two decades between '50 and '70, Italy's per capita income grew faster than that of any other European country. This could not have been achieved without the agrarian and fiscal reforms initiated in the 1950s, during which a massive population shift took place, with industrialization and modernization in the north attracting large numbers of people and exacerbating social contrasts. **Modernization and industrialization accelerated the expansion of urban homogeneity.**

Some scholars of urban studies have argued that the idea of the city as a fundamental factor in human development has led to **the cities of this era being seen as the physical manifestation of the ill effects on human society of profit, waste, expansion for the sake of expansion, and mechanization that obscures the true meaning of life.** Lewis Mumford, in 'Cities in history ' states grimly: "The ultimate damage may be irreplaceable." She bemoans the failure of capitalism to create a spiritually and socially stimulating environment, instead striving for a homogenized nothingness that "creates more and more featureless landscapes inhabited by more and more featureless people"

Figure 44: Reviews page (Scroll down to see the second part: What imagination has to do with cities)

Moving on to the 'Genre-related' theme, the Sankey diagram accentuates five writing styles (Figure 45). The text continues to adhere to the principle of providing in-depth explanations corresponding to the highlighted segments within the diagram. Furthermore, inspired by this, the social network analysis and visualization are positioned just below, available upon scrolling. Users can hover over the avatars to view the names of the individuals (Figure 46). For a more comprehensive understanding, users can click on a particular person, which triggers a pop-up link to their Wikipedia page (Figure 47).



In addition to Poetic Compelling and Imagination, there are some interesting finding from this sub-cluster. Among them, Jorge Luis Borges is the most mentioned author, who innovated fiction language by creating innovative literary symbols through imagination. Especially 'El Aleph' is a collection of short stories exploring themes such as dreams, mazes, chance, infinity, archives, mirrors, fictional authors, and myths.

Another thing worth noting is book of 'Einstein's Dreams'written by Alan Lightman. The novel follows Albert Einstein as a young scientist who is haunted by dreams in 1905 while researching the theory of relativity. Each of the 30 chapters in the book explores a dream about time that Einstein had during this period.

As a clue, the Named-entity recognition tool was used to identify these sentences and find out which writers were mentioned in the comments, and then Wikipedia was used to build a network between these writers. Then, we use Wikipedia to build a network between these writers, and from the following social network, it is easy to find that many of the mentioned writers are **famous for their imagination** , and have a **strong sense of** postmodernism **and** magic realism , and can even be **linked to humanism**.

Figure 45: Reviews page (Scroll down to see the third part: Genre Related Analysis)

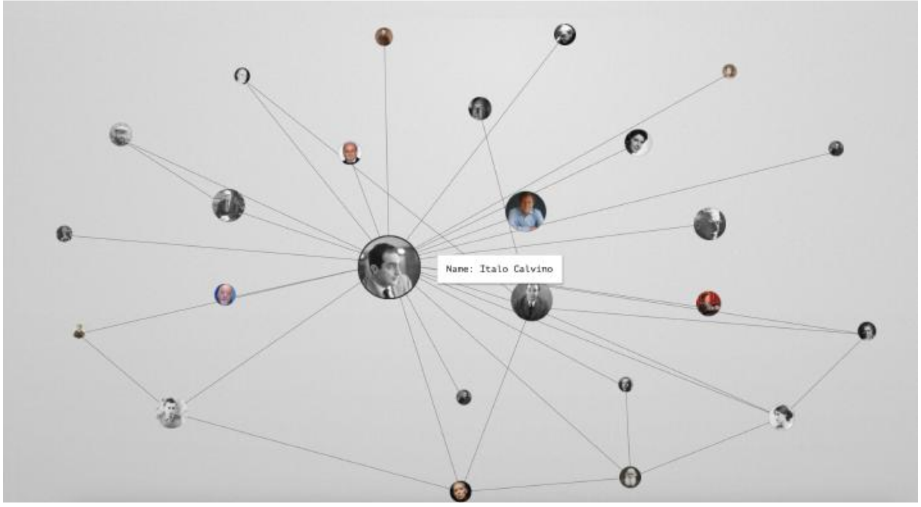


Figure 46: social network (Scroll down to see the fourth part: social network)

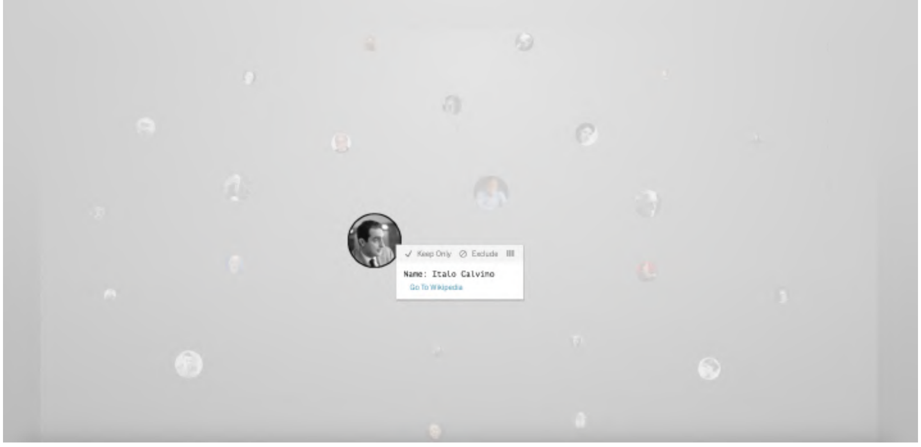
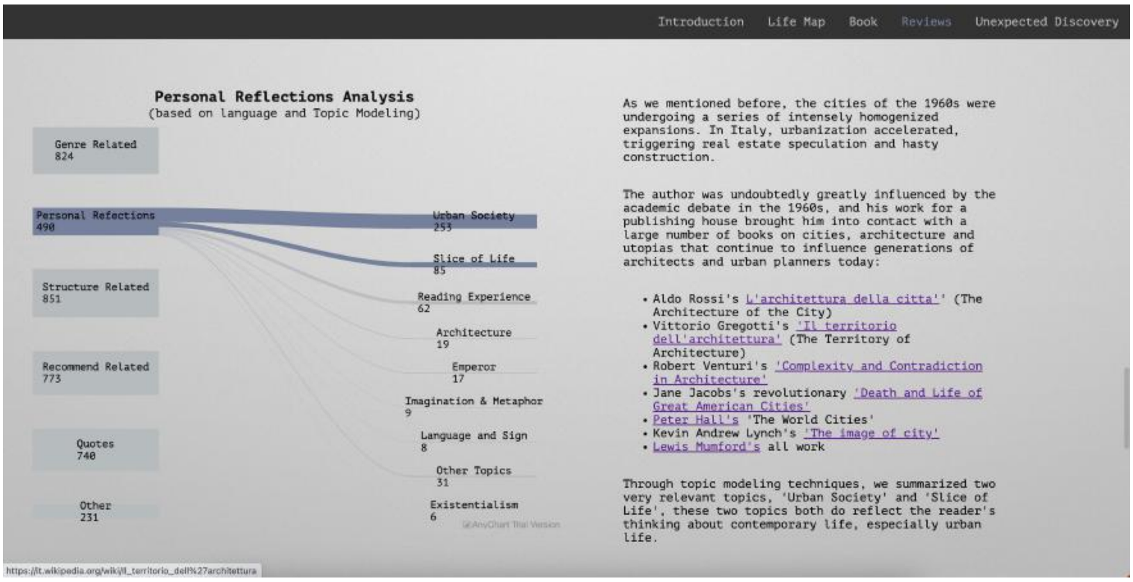


Figure 47: social network (Click on the image node then the Wikipedia link appears)

As user continue scrolling, they will reach the 'Personal Reflection' analysis section (Figure 48). Here, the emphasis is placed on two themes: 'Urban Society' and 'Slice of Life'. Given that the interpretation of this section delves into urban planning knowledge and the prevailing paradigms of urban development of that time, the textual content provides a brief introduction. Moreover, due to Calvino's editorial work, books from related fields had a significant influence on his contemplation of urban spaces and societal structures. Notable books that profoundly impacted him are highlighted here, accompanied by their respective Wikipedia links for further exploration.



As we mentioned before, the cities of the 1960s were undergoing a series of intensely homogenized expansions. In Italy, urbanization accelerated, triggering real estate speculation and hasty construction.

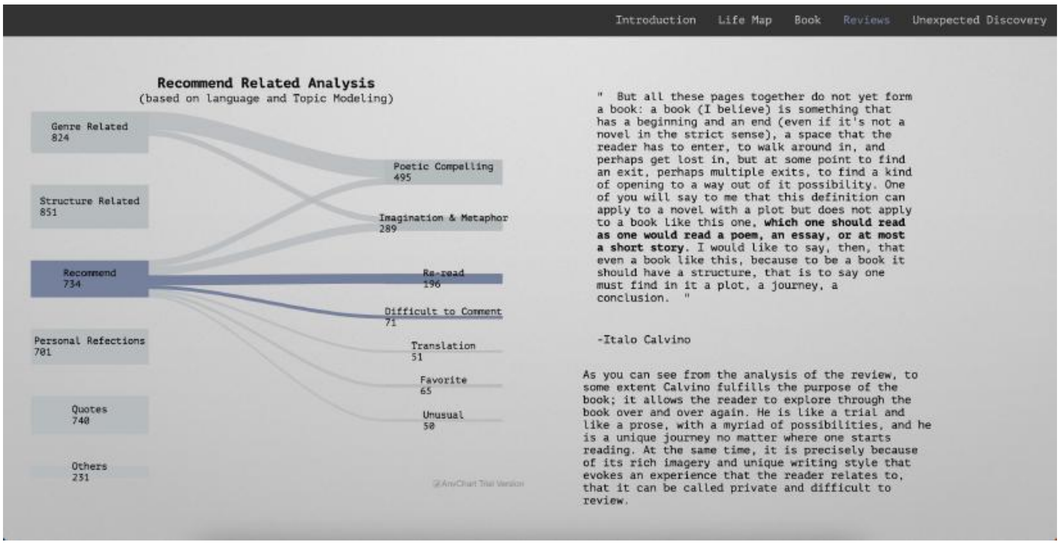
The author was undoubtedly greatly influenced by the academic debate in the 1960s, and his work for a publishing house brought him into contact with a large number of books on cities, architecture and utopias that continue to influence generations of architects and urban planners today:

- Aldo Rossi's L'architettura della citta'' (The Architecture of the City)
- Vittorio Gregotti's 'Il territorio dell'architettura' (The Territory of Architecture)
- Robert Venturi's 'Complexity and Contradiction in Architecture'
- Jane Jacobs's revolutionary 'Death and Life of Great American Cities'
- Peter Hall's 'The World Cities'
- Kevin Andrew Lynch's 'The image of city'
- Lewis Mumford's all work

Through topic modeling techniques, we summarized two very relevant topics, 'Urban Society' and 'Slice of Life', these two topics both do reflect the reader's thinking about contemporary life, especially urban life.

Figure 48: Reviews page (Scroll down to see the fifth part: Personal Reflections Analysis)

Scrolling further to the bottom, you'll find the 'Recommend Related' analysis (Figure 49). Here, 'Re-read' and 'Difficult to Comment' are the focal points. Both these themes resonate with the distinct characteristics of the book, with the 'Re-read' aspect being a deliberate intention of Calvino. The textual content cites Calvino's own interpretation of this book, emphasizing that it shouldn't be read in the same manner as a traditional novel.



" But all these pages together do not yet form a book: a book (I believe) is something that has a beginning and an end (even if it's not a novel in the strict sense), a space that the reader has to enter, to walk around in, and perhaps get lost in, but at some point to find an exit, perhaps multiple exits, to find a kind of opening to a way out of it possibility. One of you will say to me that this definition can apply to a novel with a plot but does not apply to a book like this one, **which one should read as one would read a poem, an essay, or at most a short story**. I would like to say, then, that even a book like this, because to be a book it should have a structure, that is to say one must find in it a plot, a journey, a conclusion. "

-Italo Calvino

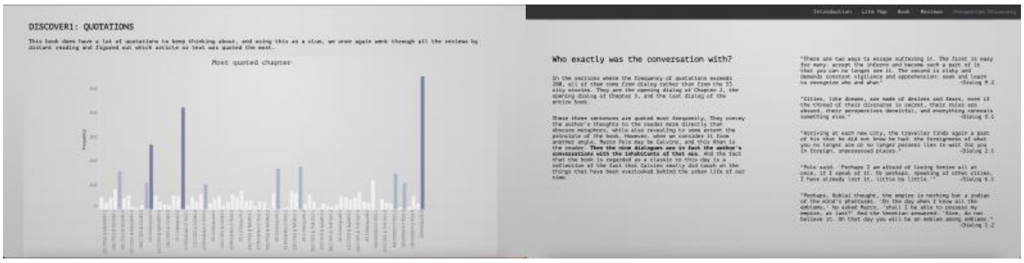
As you can see from the analysis of the review, to some extent Calvino fulfills the purpose of the book; it allows the reader to explore through the book over and over again. He is like a trial and like a prose, with a myriad of possibilities, and he is a unique journey no matter where one starts reading. At the same time, it is precisely because of its rich imagery and unique writing style that evokes an experience that the reader relates to, that it can be called private and difficult to review.

Figure 49: Reviews page (Scroll down to see the sixth part: Recommend Related Analysis)

Unexpected Discovery:

The "Unexpected Discovery" page presents two intriguing observations derived from the analytics: the most quoted section (Figure 50) and reader behavior (Figure 51). While the most frequently cited section has been discussed in the preceding text analysis segment, the patterns of reader behavior were discerned by analyzing the timestamps of the reviews. This behavior manifests in two noticeable dimensions. Firstly, there's a significant deviation in the comment patterns during March, April, and May of 2020 compared to the average levels of other years (Figure 21 above). One could speculate that this surge in reading and commenting might be due to the global lockdown triggered by the COVID-19 pandemic, providing many with the time to read and engage. Secondly, it appears that reviewers show a preference for posting reviews in the evening, especially around midnight (Figure 21 under).

As these findings aren't interconnected and emerged as additional insights during the research, they have been allocated a dedicated page for readers' convenience.



DISCOVER1: QUOTATIONS

This book does have a lot of quotations to keep thinking about, and using this as a clue, we once again went through all the reviews by distant reading and figured out which article or text was quoted the most.

Who exactly was the conversation with?

In the sections where the frequency of quotations exceeds 200, all of them come from dialog rather than from the 55 city stories. They are the opening dialog of Chapter 3, the opening dialog of Chapter 5, and the last dialog of the entire book.

These three sentences are quoted most frequently. They convey the author's thoughts to the reader more directly than obscure metaphors, while also revealing to some extent the principle of the book. However, when we consider it from another angle, Marco Polo may be Calvino, and this Khan is the reader. **Then the nine dialogues are in fact the author's conversations with the inhabitants of that era.** And the fact that the book is regarded as a classic to this day is a reflection of the fact that Calvino really did touch on the things that have been overlooked behind the urban life of our time.

"There are two ways to escape suffering it. The first is easy for many: accept the inferno and become such a part of it that you can no longer see it. The second is risky and demands constant vigilance and apprehension: seek and learn to recognize who and what" -Dialog 9.2

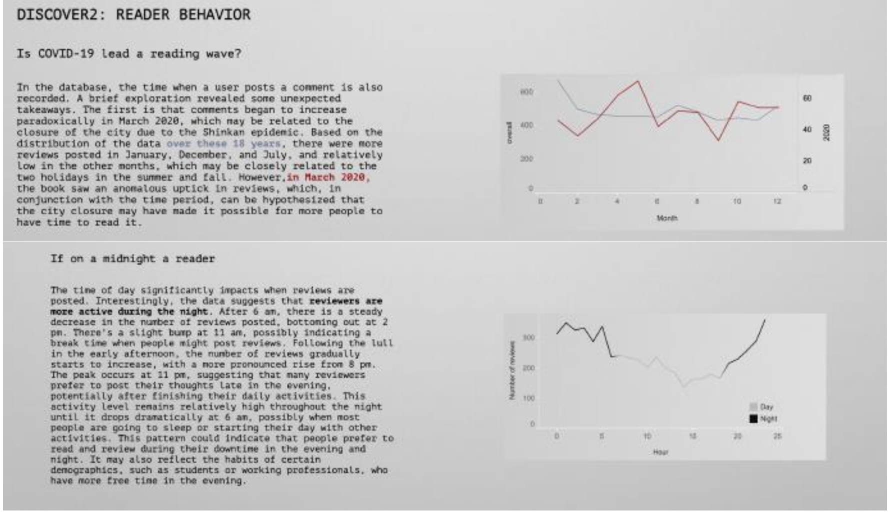
"Cities, like dreams, are made of desires and fears, even if the thread of their discourse is secret, their rules are absurd, their perspectives deceitful, and everything conceals something else." -Dialog 3.1

"Arriving at each new city, the traveller finds again a past of his that he did not know he had: the foreignness of what you no longer are or no longer possess lies in wait for you in foreign, unpossessed places." -Dialog 2.1

"Polo said: 'Perhaps I am afraid of losing Venice all at once, if I speak of it. Or perhaps, speaking of other cities, I have already lost it, little by little.'" -Dialog 6.1

"Perhaps, Kublai thought, the empire is nothing but a zodiac of the mind's phantasms. 'On the day when I know all the emblems,' he asked Marco, 'shall I be able to possess my empire, at last?' And the Venetian answered: 'Sire, do not believe it. On that day you will be an emblem among emblems.'" -Dialog 7.2

Figure 50: Discover 1: most quoted section and related discover



DISCOVER2: READER BEHAVIOR

Is COVID-19 lead a reading wave?

In the database, the time when a user posts a comment is also recorded. A brief exploration revealed some unexpected takeaways. The first is that comments began to increase paradoxically in March 2020, which may be related to the closure of the city due to the Shinkan epidemic. Based on the distribution of the data over these 18 years, there were more reviews posted in January, December, and July, and relatively low in the other months, which may be closely related to the two holidays in the summer and fall. However,in March 2020, the book saw an anomalous uptick in reviews, which, in conjunction with the time period, can be hypothesized that the city closure may have made it possible for more people to have time to read it.

If on a midnight a reader

The time of day significantly impacts when reviews are posted. Interestingly, the data suggests that **reviewers are more active during the night**. After 6 am, there is a steady decrease in the number of reviews posted, bottoming out at 2 pm. There's a slight bump at 11 am, possibly indicating a break time when people might post reviews. Following the lull in the early afternoon, the number of reviews gradually starts to increase, with a more pronounced rise from 8 pm. The peak occurs at 11 pm, suggesting that many reviewers prefer to post their thoughts late in the evening, potentially after finishing their daily activities. This activity level remains relatively high throughout the night until it drops dramatically at 6 am, possibly when most people are going to sleep or starting their day with other activities. This pattern could indicate that people prefer to read and review during their downtime in the evening and night. It may also reflect the habits of certain demographics, such as students or working professionals, who have more free time in the evening.

Figure 51: Discover 2: reader behavior

Chapter 6

Conclusion

The cross between the text analytics and the visual realms through which we comprehend information has been the focal point of this thesis. In an era dominated by an overload of data, this research emphasized the importance of extracting valuable information and presenting it in a manner that is both comprehensible and engaging.

The thesis began with a deep dive into text analytics, where techniques and strategies for extracting insights from textual datasets were discussed. Also discussed is how this kind of approach provides an alternative perspective to reading literature work. The role of natural language processing, as the backbone of text analytics, was highlighted, demonstrating how computers have become proficient in understanding and processing human language.

The choice of Italo Calvino's "Invisible Cities" and its public reviews as the central topic for text analytics and visualization brought forth an intersection of literature and data science. This case showcased how abstract literary and public sentiments could be processed, analyzed, and visualized to narrate a compelling data-driven story.

BERTopic stands out in the text analytics segment of the thesis. This method harnesses the computational strengths of language models, enhancing the accuracy of clustering. Although its application in this research necessitated additional time spent on reading cluster representative documents, BERTopic's ability to integrate context leads to a more nuanced interpretation of clusters, facilitating effective topic summarization.

About visualization, the emphasis was on converting intricate text-based insights into visual narratives, leveraging the innate human capability to understand and process visual information swiftly. The myriad techniques discussed in this segment emphasized the importance of knowing one's audience, as visualization strategies are often tailored based on who the end recipient of the information is.

The website developed in this research signifies the harmonious blend of text mining and visualization techniques. It offers a platform that not only displays the outcomes of detailed data analysis but does so in an engaging, interactive, and user-friendly manner. The design elements, narrative approach, and interactive components have been chosen thoughtfully to ensure that insights from "Invisible Cities" and its reviews are both accessible and interesting to a wider audience.

In wrapping up, this thesis showcases the rich possibilities at the intersection of text mining and visualization. By harnessing these two fields, we have the capability to convert extensive textual data into captivating narratives that inform and resonate. As the volume of data continues to expand, the methodologies and strategies touched upon in this thesis will undoubtedly serve as valuable references, aiding future scholars and enthusiasts in navigating the expansive realm of digital text.

Bibliography

Angelov, D., 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), pp.993-1022.

Botta, A., 1997. Calvino and the Oulipo: An Italian Ghost in the Combinatory Machine? *MLN*, 112(1), pp.81-89. JSTOR.

Bostock, M., 2012. *Uberdata*. Available at: https://bost.ocks.org/mike/uberdata/ (Accessed: 25th May 2023).

Breiner, L.A., 1988. Italic Calvino: The Place of the Emperor in "Invisible Cities". *Modern Fiction Studies*, 34(4), pp.559-573. JSTOR.

Case, P. and Gaggiotti, H., 2016. Italo Calvino and the organizational imagination: Reading social organization through urban metaphors. *Culture and Organization*, 22(2), pp.178-198. Taylor & Francis.

Calvino, I., 1974. *Invisible Cities*. Translated by W. Weaver. 1st English edn. New York: Harcourt Brace Jovanovich. (Original work published 1972).

Cao, N. and Cui, W., 2016. Introduction to text visualization, Vol. 1. Springer.

Coles, K. and Lein, J.G., 2013. Solitary Mind, Collaborative Mind: Close Reading and Interdisciplinary Research. In *International Conference on Digital Health*. [Online] Available at: https://api.semanticscholar.org/CorpusID:51875515.

Grootendorst, M., 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Harris, Z.S., 1954. Distributional structure. *Word*, 10(2-3), pp.146-162. Taylor & Francis.

Hullman, J., Drucker, S., Riche, N.H., Lee, B., Fisher, D. and Adar, E., 2013. A deeper understanding of sequence in narrative visualization. *IEEE Transactions on visualization and computer graphics*, 19(12), pp.2406-2415. IEEE.

Hayek, F., 1988. *The Fatal Conceit: The Errors of Socialism*. University of Chicago Press (US); Routledge Press (UK).

Iezzi, D.F., Mayaffre, D., Misuraca, M. and others, 2020. Text Analytics. Springer.

Jagarlamudi, J., Daumé III, H. and Udupa, R., 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp.204-213.

Jasinski, J., 2001. Sourcebook on rhetoric, Vol. 4. Sage Publications.

Jänicke, S., Franzini, G., Cheema, M.F. and Scheuermann, G., 2015. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. *EuroVis (STARs)*, 2015, pp.83-103.

Jockers, M.L., 2013. Macroanalysis: Digital methods and literary history. University of Illinois Press.

Jacobs, J., 1961. *The Death and Life of Great American Cities*. New York: Random House.

Kosara, R. & Mackinlay, J., 2013. Storytelling: The Next Step for Visualization. Computer, 46(5), pp.44-50.

Lee, B., Riche, N.H., Isenberg, P. and Carpendale, S., 2015. More than telling a story: Transforming data into visually shared stories. *IEEE computer graphics and applications*, 35(5), pp.84-90. IEEE.

Le, Q. and Mikolov, T., 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pp.1188-1196. PMLR.

Licata, F.L., 2021. Italian Resistance Literature: Presence of Trauma in Non-Trauma Fiction. An analysis of the narrative techniques in the works of Italo Calvino, Cesare Pavese and Beppe Fenoglio.

Liu, S., Wang, X., Collins, C., Dou, W., Ouyang, F., El-Assady, M., Jiang, L. and Keim, D.A., 2018. Bridging text visualization and mining: A task-driven survey. *IEEE transactions on visualization and computer graphics*, 25(7), pp.2482-2504. IEEE.

Lynch, K. 1960. *The Image of the City*. The MIT Press.

Marello, L., 1986. Form and Formula in Calvino's Invisible Cities. *Review of Contemporary Fiction*, 6(2), pp.95-101.

Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Modena, L., 2011. Italo Calvino's architecture of lightness: the Utopian imagination in an age of urban crisis, Vol. 17. Routledge.

Moretti, F., 2005. Graphs, maps, trees: abstract models for a literary history. Verso.

Oyebode, O., Ndulue, C., Mulchandani, D., Suruliraj, B., Adib, A., Orji, F.A., Milios, E., Matwin, S. and Orji, R., 2022. COVID-19 pandemic: identifying key issues using social media and natural language processing. *Journal of Healthcare Informatics Research*, 6(2), pp.174-207. Springer.

Rakib, M.R.H., Zeh, N. and Milios, E., 2021. Efficient clustering of short text streams using online-offline clustering. In *Proceedings of the 21st ACM Symposium on Document Engineering*, pp.1-10.

Ricci, F., 2001. Painting with words, writing with pictures: word and image in the work of Italo Calvino. University of Toronto Press.

Saunders, D., 2010. Arrival city: How the largest migration in history is reshaping our world. Random House.

Salton, G., Wong, A. and Yang, C.S., 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11), pp.613-620. ACM New York, NY, USA.

Seyser, D. and Zeiller, M., 2018. Scrollytelling--an analysis of visual storytelling in online journalism. In *2018 22nd international conference information visualisation (IV)*, pp.401-406. IEEE.

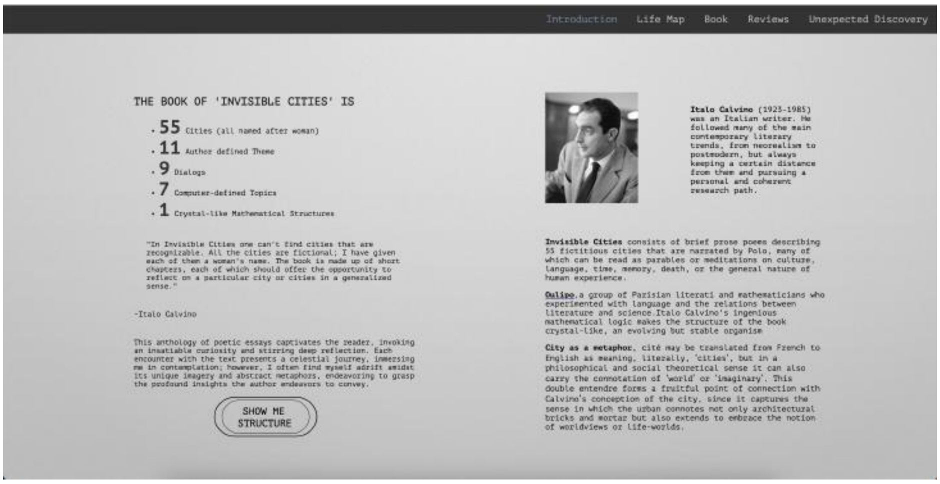
Ware, C., 2019. Information visualization: perception for design. Morgan Kaufmann.

Yamunathangam, D., Priya, C.B., Shobana, G. and Latha, L., 2021. An Overview of Topic Representation and Topic Modelling Methods for Short Texts and Long Corpus. In *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, pp.

Yang, S., Huang, G. and Cai, B., 2019. Discovering topic representative terms for short text clustering. *IEEE Access*, 7, pp.92037-92047.

Appendices

Link of final Website:
http://www.invisiblecities.me/homepage/homepage.html



THE BOOK OF 'INVISIBLE CITIES' IS

- 55 Cities (all named after woman)
- 11 Author defined Theme
- 9 Dialogs
- 7 Computer-defined Topics
- 1 Crystal-like Mathematical Structures

"In Invisible Cities one can't find cities that are recognizable. All the cities are fictional; I have given each of them a woman's name. The book is made up of short chapters, each of which should offer the opportunity to reflect on a particular city or cities in a generalized sense."

-Italo Calvino

This anthology of poetic essays captivates the reader, invoking an insatiable curiosity and stirring deep reflection. Each encounter with the text presents a celestial journey, immersing me in contemplation; however, I often find myself adrift amidst its unique imagery and abstract metaphors, endeavoring to grasp the profound insights the author endeavors to convey.

Italo Calvino (1923-1985) was an Italian writer. He followed many of the main contemporary literary trends, from neorealism to postmodern, but always keeping a certain distance from them and pursuing a personal and coherent research path.

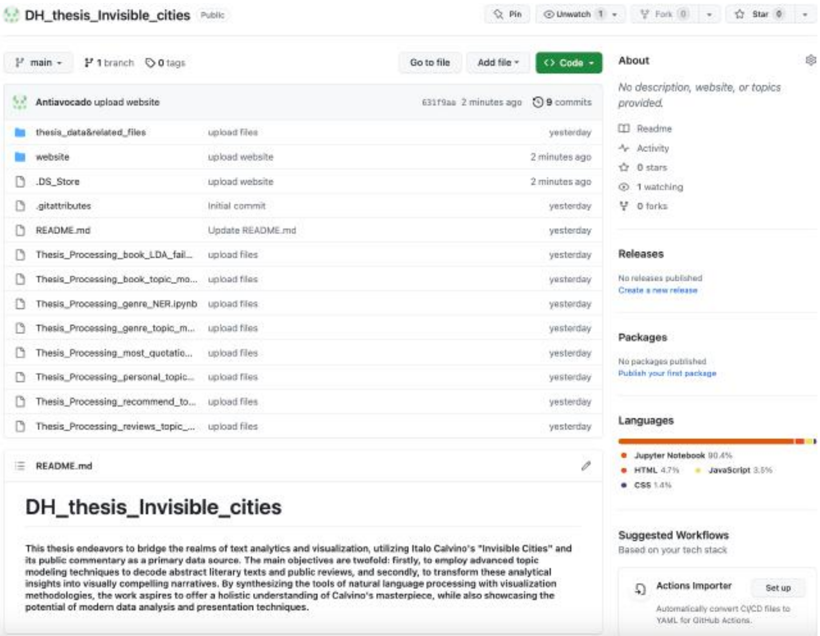
Invisible Cities consists of brief prose poems describing 55 fictitious cities that are narrated by Polo, many of which can be read as parables or meditations on culture, language, time, memory, death, or the general nature of human experience.

Oulipo,a group of Parisian literati and mathematicians who experimented with language and the relations between literature and science.Italo Calvino's ingenious mathematical logic makes the structure of the book crystal-like, an evolving but stable organism

City as a metaphor, cité may be translated from French to English as meaning, literally, 'cities', but in a philosophical and social theoretical sense it can also carry the connotation of 'world' or 'imaginary'. This double entendre forms a fruitful point of connection with Calvino's conception of the city, since it captures the sense in which the urban connotes not only architectural bricks and mortar but also extends to embrace the notion of worldviews or life-worlds.

Files of analytics process code and website:

https://github.com/Antiavocado/DH_thesis_Invisible_cities



Antiavocado upload website 631f9aa 2 minutes ago 9 commits

thesis_data&related_files	upload files	yesterday
website	upload website	2 minutes ago
.DS_Store	upload website	2 minutes ago
.gitattributes	Initial commit	yesterday
README.md	Update README.md	yesterday
Thesis_Processing_book_LDA_fail...	upload files	yesterday
Thesis_Processing_book_topic_mo...	upload files	yesterday
Thesis_Processing_genre_NER.ipynb	upload files	yesterday
Thesis_Processing_genre_topic_m...	upload files	yesterday
Thesis_Processing_most_quotatio...	upload files	yesterday
Thesis_Processing_personal_topic...	upload files	yesterday
Thesis_Processing_recommend_to...	upload files	yesterday
Thesis_Processing_reviews_topic_...	upload files	yesterday

README.md

DH_thesis_Invisible_cities

This thesis endeavors to bridge the realms of text analytics and visualization, utilizing Italo Calvino's "Invisible Cities" and its public commentary as a primary data source. The main objectives are twofold: firstly, to employ advanced topic modeling techniques to decode abstract literary texts and public reviews, and secondly, to transform these analytical insights into visually compelling narratives. By synthesizing the tools of natural language processing with visualization methodologies, the work aspires to offer a holistic understanding of Calvino's masterpiece, while also showcasing the potential of modern data analysis and presentation techniques.

No description, website, or topics provided.

No releases published
Create a new release

No packages published
Publish your first package

- Jupyter Notebook 90.4%
- HTML 4.7%
- JavaScript 3.5%
- CSS 1.4%

Based on your tech stack

Actions Importer
Automatically convert CI/CD files to YAML for GitHub Actions.

AFDELING
Straat nr bus 0000
3000 LEUVEN, BELGIË
tel. + 32 16 00 00 00
fax + 32 16 00 00 00
www.kuleuven.be