

**Analyzing Topics in Classroom of the Elite:**

*How Japanese Light Novels Unveil the Shortcomings of Topic Modelling*

Morine Croguennec

Institution: University of Antwerp

Course: Computational Literary Science

Academic Year: 2021-2022

# Abstract

# This paper will make use of topic models to measure the statistical relationship between topic clusters for each school trimester discussed in the light novel *Classroom of the Elite*. The research found that while some statistically meaningful or unambiguous trends in the generated topic distributions are present, others are hard to interpret. The paper goes on to discuss the limits of the study and those of topic modelling in general before reaching a final conclusion.

# Introduction

“Though equality is a complete lie, we should not accept inequality either” (Syougo, 2015). *Classroom of the Elite* is a Japanese novel that philosophizes about the role education has in enforcing equality as well as inequality. In short, it tells the story of Japanese high school students striving to graduate school as the top class or ‘class a’ of their year. The students that manage to achieve this goal are after all, promised to gain any job or stature they would like after graduation. The story revolves around protagonist Ayanokoji Kiyotaka and his classmates starting their first year of high school in the bottom class or ‘class d’. While facing challenges measuring their physical, academic and mental strength they too, aim to reach the famed ‘class a’- graduation (ibid, 2015). This series of light novels thus tells a story spanning over a three year time period. During this time the characters mature and get to know more people. This paper will aim to reveal if that growth can be derived from a change in topics discussed as the main characters experience each consecutive school trimester. In order to attain sets of comparable topics, a mixed-membership model, better known as a topic model will be utilized as the main tool of analysis (Karsdorp, Riddell and Kestemont, 2021).

# Datasets

The *Classroom of the Elite* series of Japanese light novels counts over 15 volumes. However, when this study was executed only 15 volumes were translated to English. In addition, many of these translations are fan made since the official translations tend to come out rather slowly. Therefore international fans who cannot wait to read the story have collectively asked for fanmade translations. These translations get posted online and are deleted once the official translations come out (Graze et al. , 2021). These fanmade translations were gathered to compose the main data for this study.

The words *main data* refer to a self-made tab separated dataset. It contains fan-made translations of all separate chapters of each volume of *Classroom of the Elite* up to the 15th (Syougo, 2015, Graze et al. , 2021). These translations were copy pasted from the internet and their usage falls under the concept of fair use (Stim, 2021). Each translator was credited by means of their pen name in the *translator* column. The following table (1) exemplifies the first two rows of the dataset:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1 | | | | | | | | | | | |
| ID | Volume | Volume  \_Type | Chapter  \_Type | Chapter  \_Num | Publishing  Date | Title | Text | Translator | Editor | Proof-  reader | Trimester |
| 1\_P | 1 | Main | Prologue | 0 | 2015-05-25 | Title of chapter 1 volume 1 | Text in chapter 1 volume 1 | qbomb | NaN | NaN | 1 |

Each chapter was attributed to a unique ID number referring to the volume number and chapter type. The volume type clarifies whether a chapter belongs to a main or sub volume and the trimester column indicates what school trimester the protagonist finds himself in for each chapter. A last column worth mentioning is the chapter type column which can be divided into the following categories: Prologue, Chapter, Epilogue and Bonus. For this study, each chapter belonging to the Epilogue and Bonus category were separated from the main data and added to the *pretraining data*.

Pretraining data is needed in order to pretrain any topic model. This pretraining data should be large in size and as similar to the main data as possible (Bianchi, Terragni and Hovy, 2021). The topic model will be able to identify a set of topics based on the pretraining data so, later on it can assess whether these identified topics are present in the main data and if so, it detects to what degree they are present (Karsdorp, Riddell and Kestemont, 2021, Newman, Chemudugunta, Smyth and Steyvers, 2006). In this case, the pretraining data contains the aforementioned bonus chapters and epilogues of *Classroom of the Elite*. They do not make part of the main story entailed in the volumes though they are very closely related. Thus they are fit to be used as pretraining data. However, since they only count a number of small number characters, on their own they do not suffice as pretraining data. As a solution to this problem, a selection of data from the Project Gutenberg Corpus was added to the pretraining data as well (Gerlach and Font-Clos, 2020). This selection contains boarding school related books which were copied from the Project Gutenberg Corpus and transformed into a tab separated format (tsv) making it easier to merge with the bonus and epilogue chapters from *Classroom of the Elite*. The resulting tsv file is exemplified in the table (2) below:

|  |  |  |  |
| --- | --- | --- | --- |
| Table 2 | | | |
| Filename | Author | Title | Text |
| PG24025\_tokens | Angela Brazil | Title of the book | Text included in the book |

The definitive pretraining data is eventually restricted to the text information of the bonus chapters, epilogues and books selected from the Project Gutenberg Project. However, each dataset was fully introduced accompanied by its metadata for the sake of completeness.

# Hypotheses

Before delving into topics that can be derived computationally, this short section aims to explain which topics and topic trends are expected to show up. The following word cloud (Fig. 1) represents a set of key words that were derived from the main data.

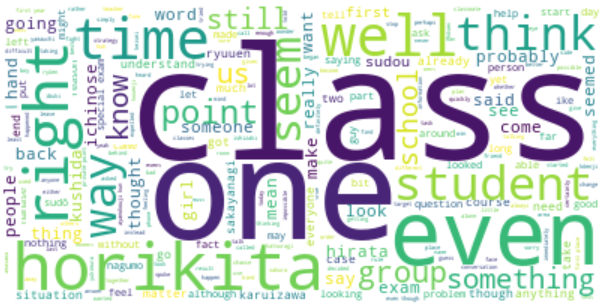


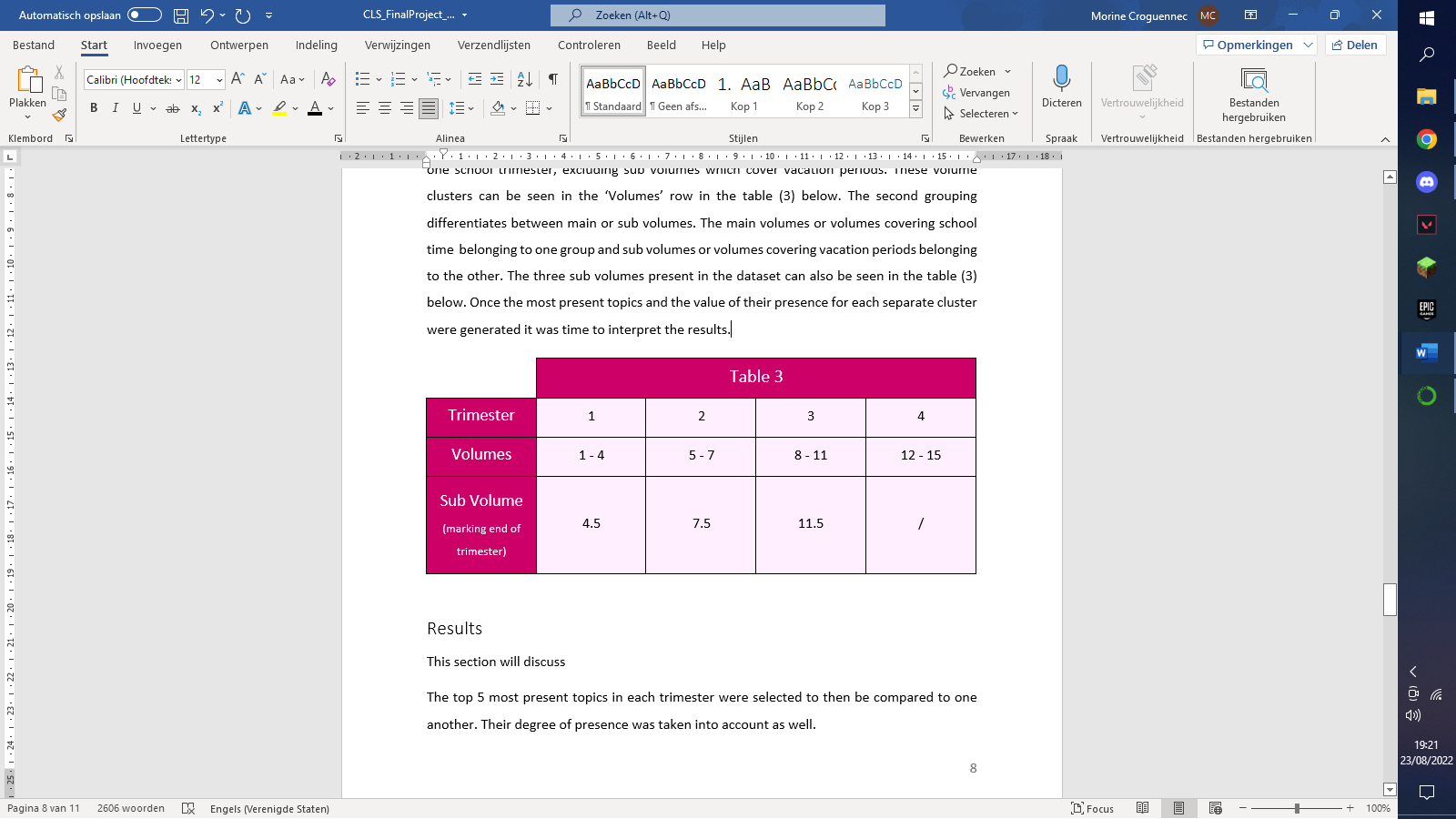
Fig. 1 Word Cloud Representing Key Words in the Classroom of the Elite dataset.

The figure (1) portrays the main theme: a strive towards becoming the number one class. Most key characters are represented as well. The main character (Ayanokoji) through this image does not appear to be as frequently mentioned as other characters however, this is likely due to spelling variations of the name. These spelling variations of character names should not pose a problem for the main research since character names generally do not make part of topic clusters.

The *Classroom of the Elite* revolves mostly around how each separate class and their leaders handle each obstacle on the path towards becoming or staying ‘the best’. Being ‘the best’ in this case means garnering the largest amount of class points by the time of their graduation ceremony. Gaining class points can been done in multiple ways but the most notable one is by ranking 1st in exams. In this type of school system, class atmospheres and cooperation are of the utmost importance. At the start of the story, the students of class D have yet to realize this. They do not know each other yet and refuse to take the school system seriously. In the first trimester of school the protagonist is mostly concerned with solving internal class issues while ensuring his class does not fall behind anymore than it already is. Aside from this goal, the protagonist also aims to stay hidden as a ‘leader behind the scenes’. He does not have an deep relationships with other students yet at this point. During the second trimester, the protagonist establishes a study group that quickly grows into his first friend group. In the third trimester, most 1st year class leaders have realized that Ajanokoji is the ‘shadow’ leader of class D. He gets challenged by the leader of class A, amongst others on multiple occasions. The 1st years start to interact more with each other and with senior students. Soon after, the oldest students or the 3rd years graduate after taking part in an exam in which all students in the school were involved. The fourth trimester mostly revolves around the introduction of new 1st years and a survival exam including students from all school years that takes place on an island. Ayanokoji also gets negative attention from the new headmaster (Syougo, 2015). Topic clusters related to these events, specific to each trimester are expected to be portrayed within the following research.

# Methodology

This section will concisely describe what steps were taken in the coding stage of this project. First of all, the main data and the pretraining data were preprocessed so they could optimally be interpreted by the topic modelling algorithm. All stop words present in the English nltk stop words list were removed from the text data for both datasets. Next to that, all text data was lowercased and special characters and punctuation were removed. After this process was executed, a Non-Negative Matrix Factorization (NMF) model was trained and fitted on the pre- training data. This model was chosen for its interpretably and effectivity as a topic model (Zhang, 2012). Any word present in the pretraining data the occurred in at least 21 of the 65 documents and appeared in less than 60 percent of those 65 documents were viable to be included in the vocabulary list for the model. This vocabulary list counted a total of 10,000 words. The variables mentioned here were decided on after a thorough optimalization process. It is important to mention that if any of these variables would have been adjusted the results of the research would also be influenced. Therefore this part of the topic modelling process cannot be deemed fully ‘objective’. Therefore, any results concluded from this research should be taken with a grain of salt.

Once the NMF model was trained and fitted, a total of 200 topics had been derived from the training data. Each topic can be seen as a cluster of words that are used in a similar context within the pretraining data (Newman, Chemudugunta, Smyth and Steyvers, 2006). Once these topics were defined, the NMF model was fitted on the main project data. Resulting from this process, a data frame containing a measure of *presence* of each topic in each chapter of *Classroom of the Elite* was devised. Thus, it became possible to measure which topics were most present in each chapter as well as in each set of chapters present in the main *Classroom of the Elite* dataset (Meeks and Weingart, 2012). According to the proposed hypotheses, four separate groupings of chapters were made. They cluster all volumes covering each separate school trimester, excluding sub volumes which cover vacation periods. These volume clusters can be seen in the ‘Volumes’ row in the table (3) below. Once the most present topics and the value of their presence for each separate cluster were generated it became possible to analyze and compare these topics to the contents of the *Classroom of the Elite*.

## Results

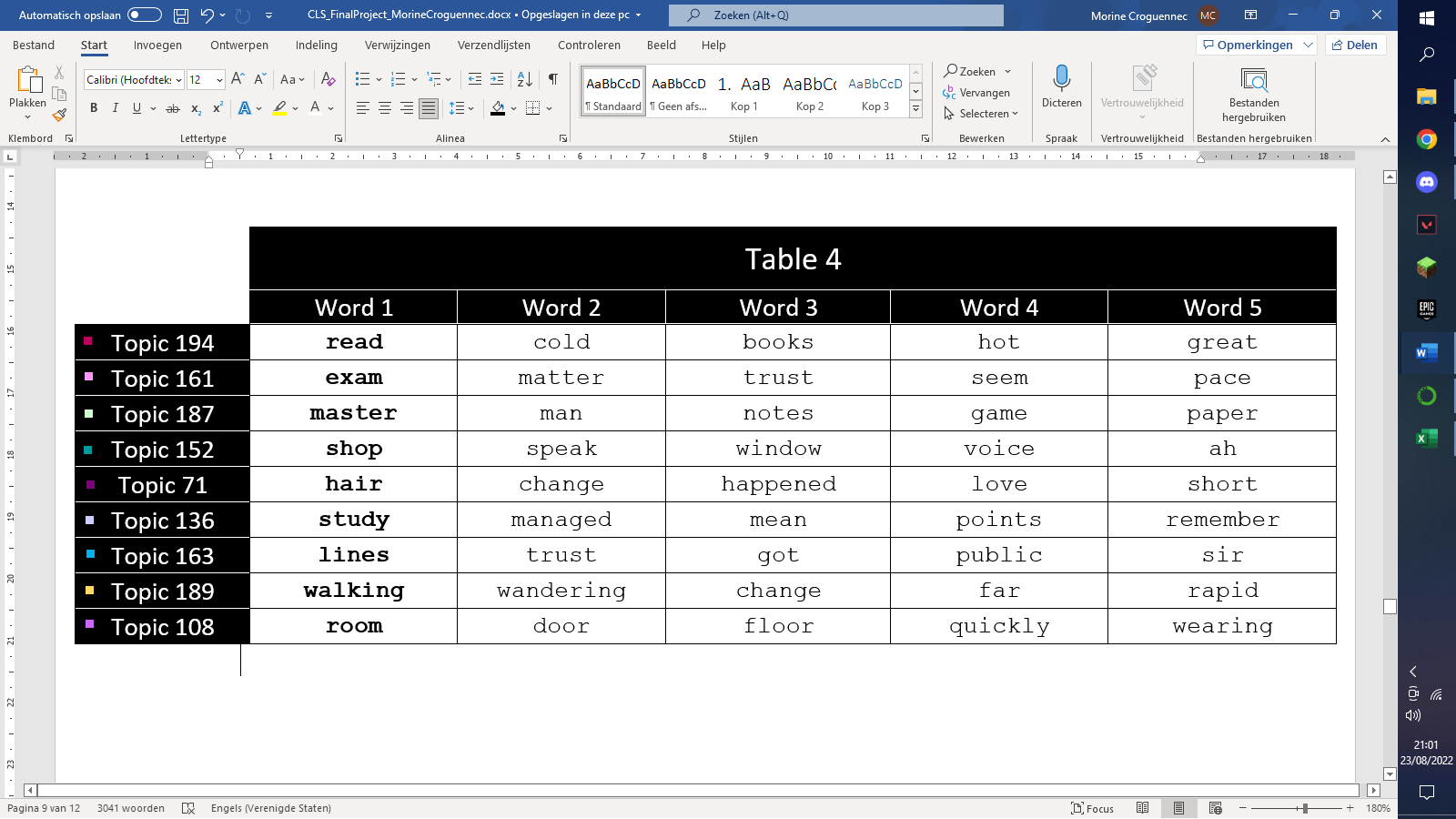
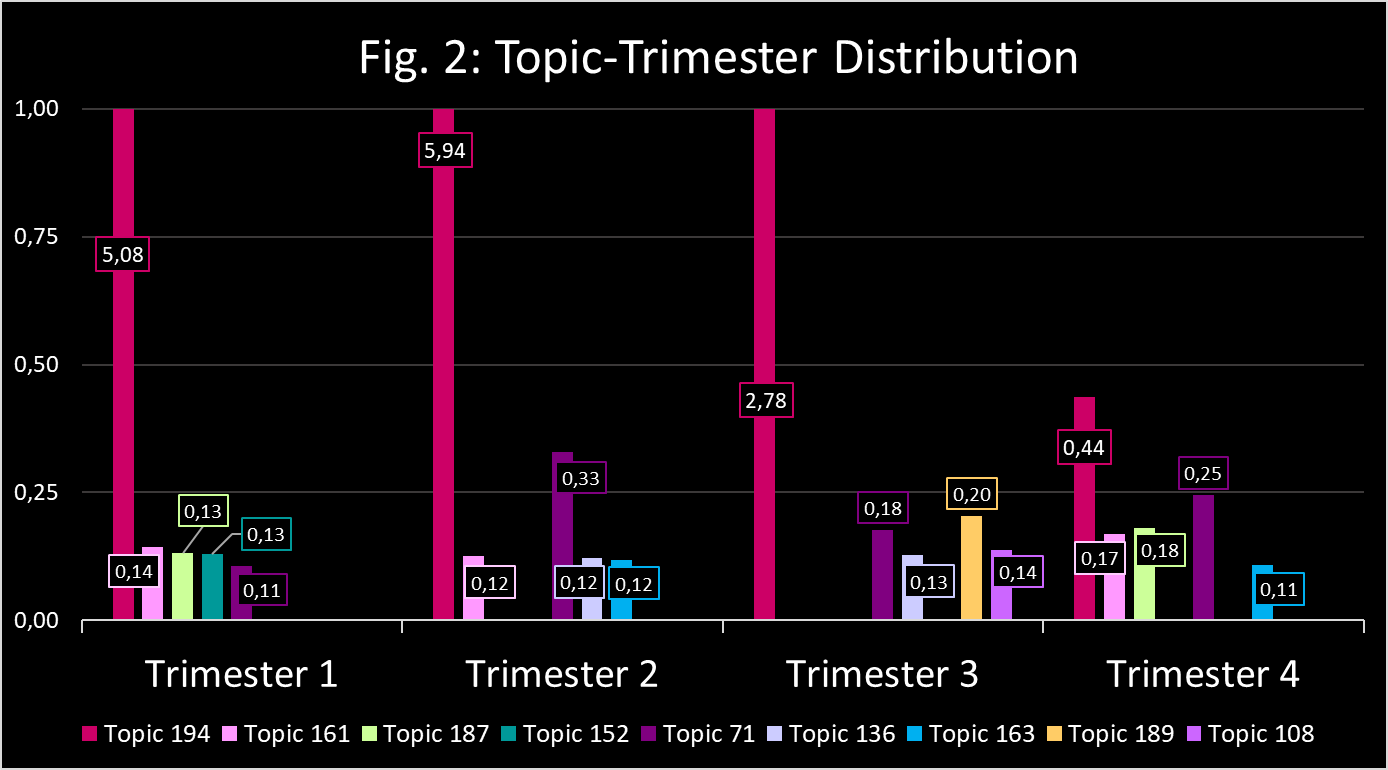
This section will discuss each generated trimester-based topic cluster. A first interesting, observation was an overall increase in topic diversity as time progresses in the story. This increase could lead to the simple conclusion that the story involves more topics as its characters grow older. This makes sense, at the start most time is spent introducing the main characters. Then, as time passes more individuals develop further and narrate different scenes. As a result, more topics naturally arise (Syougo, 2015).

In order to investigate the generated trimester based topic clusters in a clear, interpretable way, the top 5 most present topics in each trimester were selected to then be compared to one another. Their degree of presence was taken into account as well. The following figure (Fig. 2) was prepared to visualize the results found. First off, topic 194 (read, cold, books, …) was present in every trimester. Though, as affirmed by the Kendall’s Tau correlation measurement (-0.241) its presence slightly decreases over time. This trend is likely directly proportional to the increase of topic diversity over time. Reading is the most relevant word for this topic as can be seen in table 4. It is also an overall present theme in *Classroom of the Elite,* since the students read the instructions for each exam and have to read or study for academic exams. The peak that occurs in the reading related topic happens in the second semester. During this semester, Ayanokoji meets a girl in the library who is very excited to meet someone with a similar interest in books. She often carries around books just so she could lend them to anyone that shows interest. This meeting and the introduction of her character can serve as an explanation for the peak of presence for topic 194 (read, cold, books, …).

Topic 71(hair, change, happened,…) is the only other topic that ranks in the top 5 for each trimester. With the introduction of characters often comes a discussion of appearance, one character is referred to as ‘red hair’ for a large part of volume 1 for example. However, this topic relates not only to hair but also to an occurring change. In trimester 2 a girl who was bullied in the past finds herself again, she contemplates if she will ever change while here hair is dripping wet from a bucket of water that was thrown on her. In this scene Ayanokoji threatens to reveal what *happened* to the school. It is a very intense scene that ties most of the keywords of topic 71(hair, change, happened,…) together. This scene could have partially caused this topic to be more present in the second trimester than in others. A similar incident that ‘happened’ after something changed occurs in volume 15 at the end of the fourth trimester. One person twirls their hair during this incident while another brushes their hair during this incident. It is clear why this topic is deemed very present in the story by the model. While hair and change are not commonly associated, the characters in *Classroom of the Elite* tend to fiddle with their hair or think about their hair while incidents and change ‘happen’. According to the Kendall’s Tau statistical method, no significant decrease or increase is present for topic 71 or any of the following topics.

The next most present topic is topic 161(exam, matter, trust, …). The presence of this topic is logical. The students are constantly discussing the next or the previous exam after all. Each exam also relies on trust between classmates. The word matter here could refer to course matter to be studied for each exam as well as to how much each exam matters when it comes to the class rankings. The topic is present in every trimester but the third. While all of the words related to this topic do appear in this trimester, the 9th volume unlike others does not focus on a specific exam. Instead, it focusses more on internal relationships as secrets get exposed and romances start to develop around Valentine’s day. This likely resulted in the topic getting overtaken by others.

Topics 187, 136 and 163 each appear in two trimesters. Topic 187(man, master, notes, game…) appears in the top 5 topics for the first and fourth trimester. The word ‘master’ does not appear regularly in the story, it is most present in the fourth trimester while it is barely present in the volumes for the first and third trimester. Unlike some other topics the top most frequent words are all of similar importance to the topic cluster as a whole. Consequentially, the fact that the word man seemingly occurs more frequently in the volumes included in trimesters one and four than in others could have been a determining factor leading to this result. It is noteworthy that trimester one includes athletic exams such as basketball and tug of war which could be described as a ‘game’. The word ‘game’ also gets mentioned the most in the volumes belonging to the first trimester. This combination of frequencies likely lead to the rise of this topic in the first and fourth trimester. While these frequencies explain the occurrence of the topic, it is hard to find a conclusive narrative link that would cause this topic to be more present in these trimesters than in, for example the third trimester which also contains athletic games and a chess game.

Studying is of importance in each trimester of *Classroom of the Elite*, however, trimester two and three contain more content based on important academic exams such as the ‘paper shuffle’ which took place in the second trimester (volume 6). This exam led to the creation of study groups that will remain intact throughout the story. Volume 9, which is part of the third trimester, contains the end of year academic exam. Some academic tests such as a math test also took place in volume 11, the last volume pertaining to the third trimester. It seems natural that topic 136(study, manage, mean, …) took a place in the top 5 for these trimesters.

Topic 163(lines, trust, got, public) is a hard one to intuitively make sense of. In addition, it includes the word ‘got’ which does not hold a lot of meaning. However, this topic still became one of the most pertinent topics for trimester two and four. It is hard to explain this result by use of textual or conceptual evidence from the novel.

Overall, trimester three seems to be most of an outlier, since two topics, namely topic 108(room, door, floor, …) and topic 189(walking, wandering, change, …) are unique to it. As mentioned before, this trimester contains one volume which focusses mostly on personal and relational developments as Valentine’s day approaches. The presence of topics 108(room, door, floor, …) and 189(walking, wandering, change, …) is easily explained by all the secret rendezvous that take place in the dorms during this volume as well as by similar events which take place in volume 11.

Lastly, topic 152(shop, speak, window, …) appears only in the first trimester. In the first trimester, the students of class D are not yet aware of how important their private points are and how hard it will become to obtain more of them. They squander their points in the on-campus shops and soon run out. The first volume also describes many scenes where Ayanokoji looks out the window and overhears his classmates speaking. In fact, conversations of tend to take place before, after or while someone is looking outside a window in this volume.

## Discussion

This section will relate the results found to the earlier proposed hypotheses. Generally, topics discussed are related to their cohering trimesters. Only two topics were hard to connect to the contents of the *Classroom of the Elite*. The first trimester, as hypothesized contains a unique topic 152(shop, speak, window, …) which can be related to the students getting to know the school system as well as the protagonist’s apparent trouble to find friends to talk to in class. Just like any other trimester, reading books (topic 194) and figuring out how to pass exams together (topic 161) are important topics. Topic 71(hair, change, happened, …) relates change to hair, these two concepts seem to be related in trimester one as well. After all the introduction of new characters (description of hair) is often correlated to change. Though it can be said this is one of the less intuitive topics in the list. Similarly, while the words in topic 187(master, man notes, game, …) are present in the texts belonging to this trimester this topic is hard to interpret and thus hard to link to any specific concepts.

The second trimester, again includes reading books (topic 194) and figuring out how to pass exams together (topic 161). The peak in the presence of topic 194(read, cold, books, …) in this trimester can be explained by the introduction of a character that is very fond of reading books. Topic 71(hair, change, happened, …) is most present in this section, this, due to a certain bullying scene where the victim reflects on what has happened, if anything will ever change as her hair is dripping wet from the buckets of water that have been thrown onto her. Since this trimester includes the most academic exams, it is clear why topic 136(study, managed, mean, …) is related to it by the model. Lastly, topic 163(lines, trust, got) while appointed by the model as a more ‘present’ topic does not seem to have a clear link with the content of *Classroom of the Elite* in general.

According to the NMF model, the third trimester contains the most unique set of topics. Topics such as 194(read, cold, books, …) and 136(study, managed, mean, …) do reveal a more common pattern of reading and studying for academic exams. Topic 71(hair, change, happened, …) is present again as well. However, other topics discuss completely new trends. These allude to sneaking around to have secret meetings in the dorms as exemplified by topic 108(room, door, floor, …). Topic 189(walking, wandering, change) can refer to sneaking around the dorms as well as to the psychological changes that occur in certain characters. An example would be the leader of class B, who gets publicly accused of being a criminal. This rumor was spread by the student council president, who aimed to break her down mentally as a punishment for not accepting his request to start a romantic relationship. She ends up wandering around on her own in thought. The second trimester is the only trimester that contains a volume (9) which does not center around an exam, this explains why topic 161(exam, matter, trust, …) seems to have been overtaken by different topics.

The last trimester’s generated topics still include 194(read, cold, books, …) as its most ‘present’ topic, though its presence has reduced significantly. This reduction is inversely proportional to the increase in topic diversity. As the story progresses and the amount of characters discussed increases, the amount of topics seems to go up as well. Topic 161(exam, matter, trust, …) reappears, since each volume, again focusses on at least one exam. This volume also has a clear example of an incident being caused, while the perpetrator is seen fiddling with their hair. This incident could explain the presence of Topic 71(hair, change, happened, …). The last two topics, 187(master, man, notes, game, …) and 163(lines, trust, got, …) are hard to relate to story contents though it is apparent that the word ‘master’, specifically when referring to the head master occurs most frequently during this trimester.

It was expected that in general, topics related to exams, reading and studying would be present in each trimester. The topics unique to each trimester also reveal hypothesized tendencies. For example, topic 152(shop, speak, window) can, albeit vaguely be related to the protagonist’s inability to socialize after entering high school. Whereas topic 108(room, door, floor, …) seems to refer to a deepening of bonds between students during trimester 3, as they have secret meetings. Topic 189(walking, wandering, change) might pertain to personal development with student interactions as its instigating factor. If so, the appearance of this topic also adheres to the hypotheses. The appearance of topic 136(study, matter, trust) in trimester 2 coincides with the initiation of the protagonist’s study group which was an expected main subject for this specific trimester. The last trimester contains topics which least adhere to the made hypotheses. While it contains a topic for which ‘master’ is the most relevant word which could potentially be a hint at Ayanokoji’s qualms with the head master, the topic as a whole is hard to make sense of. The topic 163(lines, trust, got) which appears here as well is also hard to relate to the story.

# Limitations

Topic modelling is a unsupervised method for labeling texts that solely relies on computer algorithms to decide which content words could be grouped together and labelled as one topic (Karsdorp, Riddell and Kestemont, 2021, Newman, Chemudugunta, Smyth and Steyvers, 2006). However, this does not mean no human intervention is present when running a mixed-membership or topic model. The human running the model decides what data to use during the pretraining stage, how many words can be included, and how many topics will be generated amongst other variables. Therefore, topic modelling is never without bias and it is important as a researcher to remain aware of this fact (Haverals and Geybels, 2021, Meeks and Weingart, 2012).

As mentioned, a topic model needs to be pretrained on a large set of data similar to the data the study will be performed on. In this case, the best pretraining data would have been a large dataset containing Japanese light novels, preferably set in a school context. This dataset however, does not exist in a publicly accessible environment, if it even exists at all. Time- and work limitations prevented me from making this type of dataset myself so an attempt was made to find the next-best alternative. While the bonus chapters and epilogues used as pretraining data were excellent in terms of comparability to the main data, they were limited to 104,346 characters. To enlarge the pretraining data, they were supplemented by a selection of boarding school related data from the Project Gutenberg Corpus (Gerlach and Font-Clos, 2020). This merge of data caused the pretraining data to count 390,661 characters instead. While the amount pretraining data was almost quadrupled, questions could still be raised as to whether this is enough data for the topic model to be trained on.

Topic modelling is not without its downsides. Aside from this need for a large set of pretraining data comparable to the main data, researchers have pointed out a multitude of weaknesses the research method has. Meeks and Weingart in their article “The Digital Humanities Contribution to Topic Modeling” summarize some of these weaknesses. They refer to topic modelling as:

*distant reading in the most pure sense: focused on corpora and not individual texts, treating the works themselves as unceremonious “buckets of words” and providing seductive but obscure results in the forms of easily interpreted (and manipulated) “topics” (2012).*

This quote reveals different critiques that have come up in regards to topic modelling. First of all, when applying a topic model algorithm, the narrative context of a written work gets disregarded. A text indeed, gets treated as a set of different words each occurring with a different frequency. The practice has been therefore also been labelled as ‘simply counting words’ (Da, 2019). It is also argued that topic modelling distracts literary scholars from their true purpose of interpreting language, instead they are tempted to simply interpret topics, simple clusters of out-of-context words (Schmidt, 2012). It is true that topic modelling leads to interpreting the topics the model provides, it is important to link the interpretations of those topics to interpretations of each text as a whole. Where the topic model discards context to provide topic clusters, the literary scholar should interpret those topics by relaying them back to that very context the algorithm discarded. Interpreting topics cannot only lead to a loss of literary interpretations but to ‘manipulating’ topics as well. This is another problem which Meeks and Weingart hint at in their article (2012). As shortly addressed in the methodology section of this paper, a topic model can reveal different results based on the variables the coder choses to include. Comparing the same models performed on the same texts, only varying in terms a variables can lead to largely differentiating results (Goldstone and Underwood, 2012). A researcher could easily be temped to choose variables which provide results that best adhere to their hypotheses. Not all of these problems can be prevented though they can be reduced to a minimum as long as the researcher stays aware of them. When choosing variables I attempted to have the topics make sense when related back to the story as a whole before looking at the clustered results for each hypothesis made. The aim was to minimize result manipulation while still generating topic clusters that made sense. The main condition for a topic to be sensible was for model’s vocabulary list to contain mostly content words, thus reducing the presence of other words. As the Merriam-Webster online dictionary states: a topic is “the subject of a discourse or of a section of a discourse” (2022), topics contain subject matter which contain meaning it would be hard to argue that function words should make part of a topic. To prevent the other problems Meeks and Weingart point out from appearing within this study, all topics were interpreted in context of the whole story as much as possible and if no specific relation could be found, it was explicitly mentioned. As McCarty states, it is important to understand what we are doing while modeling topics since that awareness is the key to unlocking the true potential of the practice or in her words: “No one doubts the usefulness of the practice. Rather it's the *intellection of praxis* to which the next stage in the argument I have begun here must turn”(2014).

# Conclusion

Topic models are a helpful aid when it comes to literary analysis though they should not be utilized mindlessly. Research based on topic modelling is useful only when possible caveats are understood and taken into account. This case study exemplifies that topic models need to be pretrained well for them to create understandable topic clusters. Some topic clusters generated during this project were not as intuitively comprehensible as others. However, considering the scarcity of pretraining data suited for Japanese visual novel research the topic model did a fairly good job at interpreting the given input. Most returned topics were easily related to the content of the *Classroom of the Elite*.

# References

Schmidt, B.M. (2012). *Words alone: Dismantling topic models in the humanities*. [Online]. Journal of Digital Humanities. Available from: <http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/>. [Accessed: 23 August 2022].

Bianchi, F. , Terragni, S. and Hovy, D. (2021). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

Da, N. Z. (2019). The computational case against Computational Literary Studies. *Critical Inquiry*. 45 (3). pp. 601–639.

Gerlach, M. and Font-Clos, F. (2020). A standardized project gutenberg corpus for statistical analysis of Natural Language and Quantitative Linguistics. *Entropy*. 22 (1). pp. 126.

Goldstone, A. and Underwood, T. (2012). What can topic models of PMLA teach us about the history of literary scholarship? [Online]. Journal of Digital Humanities. Available from: <http://journalofdigitalhumanities.org/2-1/what-can-topic-models-of-pmla-teach-us-by-ted-underwood-and-andrew-goldstone/>. [Accessed: 23 August 2022].

Graze et al. (2021). *Youkoso jitsuryoku*. [Online]. 22 December 2021. Confused Translations. Available from: <https://confusedtls.wordpress.com/youkoso-jitsuroku/>. [Accessed: 19 August 2022].

Haverals, W. and Geybels, L. (2021). ‘Putting the Sorting Hat on J.K. Rowling’s Reader: A digital inquiry into the age of the implied readership of the Harry Potter series’, *Journal of Cultural Analytics*. 6 (1). pp. 255–284.

KADOKAWA CORPORATION (n.d.). MF文庫J『ようこそ実力至上主義の教室へ』公式サイト. [Online]. ようこそ実力至上主義の教室へ. Available from: http://youkosozitsuryoku .com/products/. [Accessed: 19 August 2022].

Karsdorp, F. , Riddell, A. and Kestemont, M. (2021). Humanities Data Analysis: Case Studies with Python. Princeton University Press.

McCarty, W. (2014). Modeling: A study in words and meanings. *A Companion to Digital Humanities*. pp. 254–270.

Meeks, E. and Weingart, S. B. (2012). *The Digital Humanities contribution to topic modeling*. [Online]. Journal of Digital Humanities. Available from: <http://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling/>. [Accessed: 23 August 2022].

Newman, D. , Chemudugunta, C. , Smyth, P. and Steyvers, M. (2006). Analyzing entities and topics in news articles using statistical topic models. *Intelligence and Security Informatics*. pp. 93–104.

Stim, R. (2021). *What is fair use?* [Online]. 25 November 2021. Stanford Copyright and Fair Use Center. Available from: <https://fairuse.stanford.edu/overview/fair-use/what-is-fair-use/#:~:text=In%20its%20most%20general%20sense,permission%20from%20the%20copyright%20owner>. [Accessed: 24 August 2022].

Syougo, K. (2015). Classroom of the Elite. Tokyo, Japan: Kadokawa corporation.

*Topic*. [Online]. Merriam-Webster. Available from: <https://www.merriam-webster.com/dictionary/topic>. [Accessed: 23 August 2022].

Zhang, Z. -Y. (2012). Nonnegative matrix factorization: Models, algorithms and applications. *Intelligent Systems Reference Library*. pp. 99–134.