Universiteit Antwerpen

# 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 Studies

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 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, promised to gain any job or stature they would like after graduation. The story revolves around protagonist Ayanokouji Kiyotaka and his classmates who start 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* is an ongoing series of Japanese light novels written by Kinugasa Syougo and illustrated by Tomose Shunsaku that started being published in 2015. So far, the series includes 22 volumes. Light novels are a specific type of visual novel originating in Japan, their illustrations commonly cohere to manga art styles. These novels tend to cater to a young adult audience around the age of high school students (Enomoto, 2018). When this study was executed only 15 volumes of the light novel series 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 fan-made translations. These translations get posted online and are deleted once the official translations come out (Graze et al. , 2021).

In this study, the words *main data* will 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 15$^{th}$ (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, Afterword and Bonus. For this study, each sub volume and each chapter belonging to the Afterword 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, in terms of content and vocabulary used, as possible (Bianchi, Terragni and Hovy, 2021). The topic model will be able to identify a set of topics based on the pretraining data. Afterwards, the model 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 afterwords of *Classroom of the Elite*. Though they are not part of the main story entailed in the volumes, they are very closely related in terms of characters and setting. Thus, they are fit to be used as pretraining data. The sub volumes of Classroom of the Elite will also be used as pretraining data since they entail content regarding vacation periods which do not take place during a school trimester. However, their total token count is on the small side (193,177 tokens). Thus, 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 (Gerlach and Font-Clos, 2020). This selection contains 23 boarding school related books (1,402,919 tokens) 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 afterword 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 |

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, mostly revolving around Japanese teens. 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, afterwords and sub volumes used as pretraining data were excellent in terms of comparability to the main data, they were limited to 193,177 tokens. Extra data was needed. However, adding more data was not simple. As mentioned, there is a lack of data similar to the contents of *Classroom of the Elite*. It is not a simple 'school story'. It has a fictive Japanese school setting with unique school rules that lead to more than one survival-like exam staged on an island. When adding more generic data to pretraining data, the topics generated failed to represent topic trends in *Classroom of the Elite*. The only accessible data that could be seen as suitable was a selection of boarding school related data from the Project Gutenberg Corpus (Gerlach and Font-Clos, 2020). While still not the best fit, since they do not take place in a Japanese school system, this data could still provide topics befitting a boarding school setting which is present in the *Classroom of the Elite* series. Merging these two data caused the pretraining data to count 1,596,096 tokens instead. While the amount of pretraining data was largely expanded, questions could still be raised as to whether this is enough data for the topic model to be trained on. However, I opted not to add more data that would not suit the vocabulary or content of the main data.

## 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 *Classroom of the Elite* revolves mostly around how each separate class and its leaders handle obstacles 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 outperforming other classes 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 any more than it already has. Aside from this goal, the protagonist also aims to stay hidden as a 'leader behind the scenes'. At this point in time, he does not have any deep relationships with other students yet. During the second trimester, the protagonist establishes a study group that quickly grows into his first friend group. The reader can see him grow closer to his fellow students. In the third trimester, most first year class leaders have realized that Ayanokouji is the 'shadow' leader of class D. He gets challenged by the leader of class A, amongst others on multiple occasions. The first years start to interact more with each other and with senior students. Soon after, the oldest students or the third 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 first year students and a survival exam including students from all school years that takes place on an island. Ayanokouji also gets negative attention from the school's new acting director (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 (Bird, Klein and Loper, 2009). Markers of politeness typical of the Japanese language such as 'san' and 'kun' were added to this list. Next to that, all text data was lowercased and special characters such as brackets, ampersands and hashtags in addition to punctuation(', !, ", -, ., :, ;, etc.) were removed.

After this process was executed, a Non-Negative Matrix Factorization (NMF) model was trained and fitted on the pretraining data. This model was chosen for its interpretably and effectivity as a topic model (Zhang, 2012). Any word present in the pretraining data that occurred in at least 9 of the 84 documents (10 percent) and appeared in less than 50 percent or 42 of those 84 documents were viable to be included in the vocabulary list for the model. In this case, a document refers to either one of the 23 books in the Gutenberg selection or one of the 61 chapters of the pretraining selection of *Classroom of the Elite*. If a word only occurred in 10 percent of the documents, it was assumed that it would not be a marker of the pretraining data's general content. They were not suited for the formation of topic clusters based on the topics generally present in this data. The 50 percentile was chosen after a process of trial and error, if multiple topics included the same set of words combined differently, the maximum document frequency was lowered. The resulting vocabulary list counted a total of 5,000 words, an amount suitable to form an extensive vocabulary of content words out of the relatively small pretraining dataset.

Once the NMF model was trained, a total of 75 topics had been derived from the training data. This parameter as well, was chosen through a process of trial and error. Choosing a larger amount like 100 or 150 topics mostly lead to nonsensical outcomes where the words inside a topic cluster could not be related to one another. Generating a large amount of topics would also not be suitable based on the fact that the pretraining data only counts 1,596,096 tokens. It is important to mention that if any of the mentioned variables would have been adjusted, the results of the research would also be influenced. Subsequently, any results concluded from this research should be taken with a grain of salt.

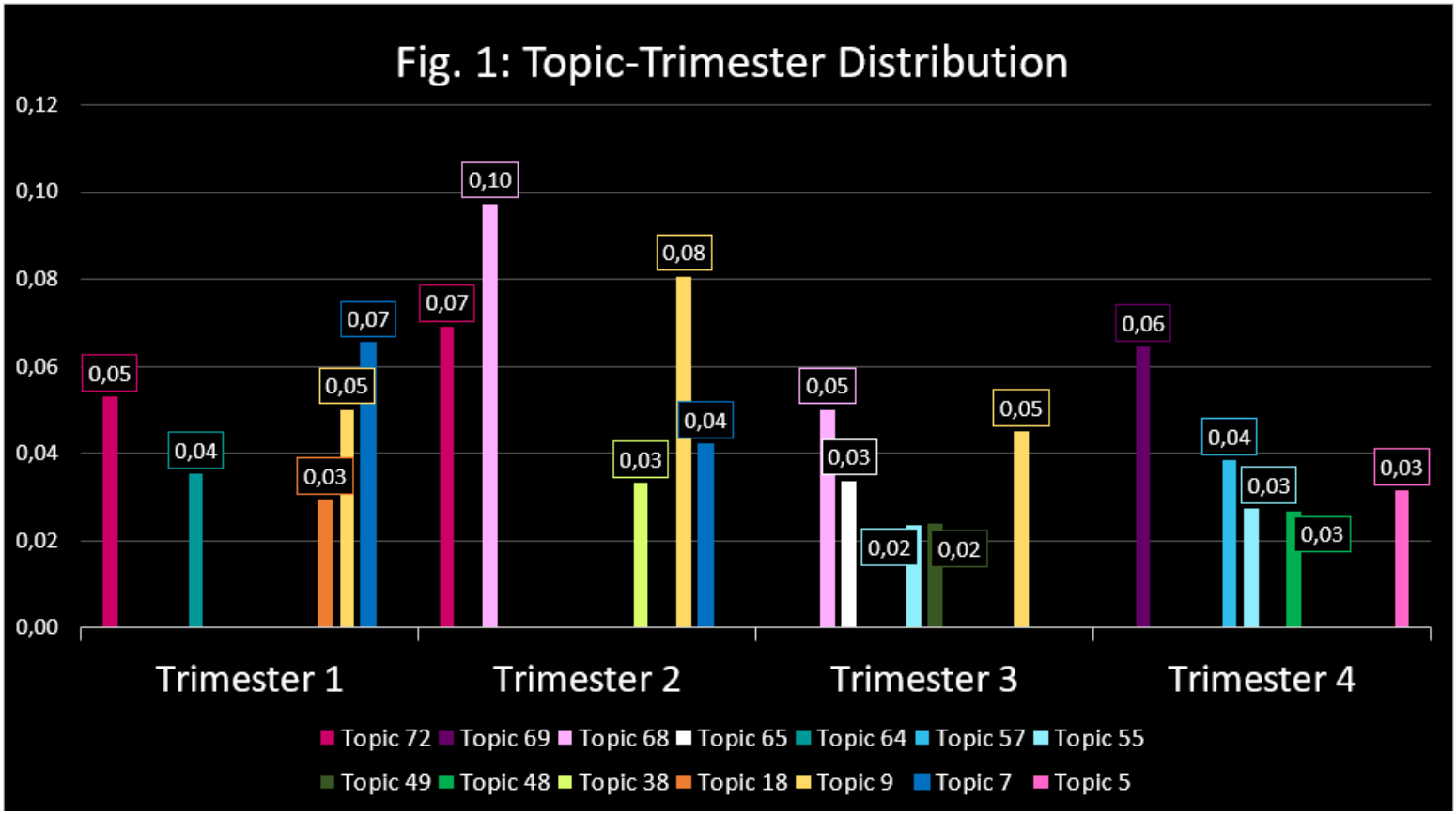
Each topic can be seen as a cluster of words that are used in a similar context within the pretraining data (Newman, Chemudugunta, Smyth and Steyvers, 2006). Once these topics were defined, the NMF model was fitted on the main project data. Resulting from this process, a data frame containing a measure 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 them to the contents of the *Classroom of the Elite*.

Table 3

| Trimester | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Volumes | 1 - 4 | 5 – 7 | 8 - 11 | 12 - 15 |
| Sub Volume (marking the end of each trimester) | 4.5 | 7.5 | 11.5 | / |

## Results and Discussion

This section will discuss each generated trimester-based topic cluster. In order to investigate these 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. 1) was prepared to visualize the results found. Each topic will be discussed in order of presence per trimester. The order will be set from most present to least present. This discussion will be followed up by a concluding section linking the results found to the above-mentioned hypotheses.




### Trimester 1

Topic 7(sudou, ike, yamauchi, council, birthday, …): Sudou, Ike and Yamauchi are three guys from the protagonist's class that quickly grew into a close friend group. On the first day of classes they form a group of boys to discuss different matters. Amongst other matters, they discuss how stern the student council president looked at the welcoming ceremony for the first years. The word birthday does not occur in the volumes covering the 1st trimester.

Topic 9(karuizawa, nagumo, hirata, date, love, …): At the start of the schoolyear, it quickly becomes clear that Karuizawa and Hirata start dating. They are two of the three most popular

people in class D and end up taking a leading role during class activities and discussions. The name nagumo does not show up in any of the volumes included in the 1st trimester. This name refers to the second year who later becomes the new student council president. His name seems to be included in this topic based on the content of one of the sub volumes. In this particular sub volume (volume 7.5), Nagumo briefly interrupts a double date Karuizawa and Hirata are on. However, in the story Nagumo is unrelated to Karuizawa and Hirata's relationship.

Topic 72(sensei, reception, ayanokouji, sakayanagi, permission, …): This topic is hard to make sense of since the words included do not seem related to one another in a clear way. The Japanese word 'sensei' refers to a teacher. However, teachers were not involved in the club receptions at the start of the year. Of course, sometimes students such as the protagonist Ayanokouji or Sakayanagi(the leader of class A) ask teachers for permission in order to do something but there is no such specific moment to refer to within the contents of the 1st trimester.

Topic 64(changing, ike, car, peeping, pool, …): The boys in class D, led by Ike, show interest in what the girls will look like after changing into their bathing suits for a swimming class. Ike even contemplates going into their changing room.

Topic 18(hair, image, kushida, sudou, love, …): This topic could refer to Kushida, the most popular girl in class D and her secret personality. While she has the image of a goody-two-shoes who tries to get along with everyone, secretly she curses and complains about how much everyone bothers her. She often tends to twirl her hair when exposing this secret side of hers to others. However, the words Sudou and love do not seem to have a connection with Kushida's secret personality.

Overall, the topics discussed here seem to correlate to the content present in the volumes covering the 1st trimester. As predicted in the hypothesis section, the first trimester mostly covered topics depicting the internal relations within class D (the protagonist's class). The protagonist, Ayanokouji has more of an observing function in this trimester. As a consequence, the most present topics revolve more around his classmates' actions rather than his own.

## Trimester 2

Topic 68(ryuuen, nagumo, council, older, points, …): Ryuuen, class C's leader takes a leading role in the second semester as he tries to identify the mastermind behind class D; Ayanokouji. Ryuuen's style of leadership was compared with those of the former and current student council presidents'. At the end of the second trimester, Ayanokouji comes into contact with the now 'former' student council president Horikita Manabu. The former president is concerned about the new president; Nagumo. He fears Nagumo intends to change the school too extremely. He hopes to entrust Ayanokouji with the task of revoking Nagumo from his position. While Ayanokouji is not particularly interested  he owes Horikita Manabu a favor since he ensured Ayanokouji's victory against Ryuuen.

Topic 9(karuizawa, nagumo, hirata, date, love, …): At the end of the first trimester, it is revealed to the protagonist that Karuizawa and Hirata are in a fake relationship. Karuizawa was bullied in middle school and wanted to protect herself from being looked down on by dating the most popular guy in class; Hirata. Hirata was kind enough to help her out even though he was not interested in a relationship. Throughout the second trimester, Ayanokouji points out how proficient the two are at pretending to be a couple multiple times. Though again, it is hard to link Nagumo to this topic cluster.

Topic 72(sensei, reception, ayanokouji, sakayanagi, permission, …): As mentioned earlier the words in this topic do not have a clear connection to one another, thus this cluster is hard to contextualize.

Topic 7(sudou, ike, yamauchi, council, birthday, …): Sudou, Ike and Yamauchi remain a close group of friends throughout the second trimester. They often appear as a trio. This trimester, however, they did not discuss anything related to the student council. The word birthday, again does not occur in the volumes covering the second trimester either.

Topic 38(months, importantly, reliable, kinugasa, refrain, …): This topic was likely generated from the epilogues written by the author of classroom of the elite, Kinugasa Syougo. It does not seem to tie in with the story.

The most 'present' topics generated for the second trimester are similar to those of the first. Only two differ, of which one is difficult to interpret in relation to the story. While topic 68 does indicate that the protagonist is starting to make interpersonal connections, it does not

prove the hypothesis provided for this trimester. A topic regarding the creation of the protagonist's new friend group was expected to show up but it did not.

### Trimester 3

Topic 68(ryuuen, nagumo, council, older, points, …): Ayanokoji encounters president Nagumo and former president Horikita during an exam involving all school years. He observes their interactions and secures information about the older students. Ryuuen however is not tied in with this occurrence. He does not have much contact with the senior students at all during the third trimester.

Topic 9(karuizawa, nagumo, hirata, date, president, …): Karuizawa becomes stable enough to live independently from Hirata. In volume 9, he decides to break up with him for both of their sakes. This information becomes shocking gossip to their fellow students who were unaware of the fake nature of their relationship. Again, Nagumo has no significant connection to their relationship in the story.

Topic 65(sakayanagi, ichinose, ryuuen, watch, theater, …): At this point in time; Ayanokouji has been recognized as the mastermind of class D by the leaders of all other classes in his year; Sakayanagi(class A), Ichinose(class B) and Ryuuen(class C). The word theater is not present in any of the volumes included in the third trimester.

Topic 55(council, nagumo, brother, join, president, …): Ajanokouji aims to interfere with president Nagumo through his classmate Horikita Suzune. However, she does not seem like she would willingly cooperate. To convince her, he lies that her brother (Horikita Manabu), the former president wants her to join the student council. Once she joins the council, his goql is to extract information from her.

Topic 49(ayanokouji, boy, amazing, cup, formidable, …): During the first exam of the third trimester, Ayanokouji is grouped together with his classmate Kouenji, amongst other students. Koenji calls him 'Ayanokouji boy'.

The generated topics for this trimester tie in well with the predictions made in the hypothesis section. Topic 65 relates to the intensifying contest between the first years' class leaders. From topics 68 and 55, it can also be derived that the first years have more contact fist senior students. The third years' graduation is also subtly reflected in these topics. Horikita Manabu

wants Ayanokouji to defeat Nagumo, since he himself would be unable as he has too little time before he graduates.

## Trimester 4

Topic 69(senpai, understood, detail, human, Ayanokouji, …): This topic is on the vague side. It may refer to a set of occasions that occur during the uninhabited island exam which takes place in volume 14. During this exam, Ayanokoiji spends most of his time travelling with a junior student of his. Since they are originally from different groups, Ayanokouji decides not to reveal his thought processes in detail. The junior student (Nanase), naturally understood this.

Topic 57(senpai, ayanokouji, answering, partners, okay, …): Ayanokouji, has now become a second year faces and thus a 'senpai' or senior student. He quickly faces a new academic exam. However, for this exam, second years need to pair up with the newly introduced first years. All second years are looking for a partner, and after a long process of consideration, Ayanokouji settles on a partner for the exam as well.

Topic 55(council, nagumo, brother, join, president, …): Horikita Suzune becomes determined and is eventually allowed to join the student council.

Topic 48(senpai, climbing, catch, definitely, Ayanokouji, …): As events during the uninhabited island exam unravel, it becomes clear that a very large bounty (20 million points) has been placed on his head by the acting director. All first years are aware of this bounty and either as separate groups or as individuals, they try to catch Ayanokouji off guard and take the bounty for themselves. The largest confrontation between the first years and the protagonist takes place after Ayanokouji has climbed a mountain to avoid being followed.

Topic 5(senpai, ayanokouji, beach, win, flag, …): During the uninhabited island exam, students come across multiple events they can choose to participate in to win points. One of these events is called: 'Beach Flags Showdown'. During the exam, Ayanokouji wants to participate in this event. However, a senpai notices him and takes up the last spot left to participate, though Ayanokouji still ends up spectating the event.

As speculated in the hypothesis, the topics generated for the fourth trimester mostly revolve around the introduction of new 1$^{st}$ year students (topic 57) and the survival exam including

students from all school years (topics 69, 48 And 5). Lastly, topic 48 could be related to the fact that the new acting director's goal of getting Ayanokouji expelled.

## 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).

Topic modelling is not without its downsides. Aside from the need for a large set of pretraining data comparable to the main data mentioned in the *dataset* section of this paper, 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 tempted 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)*.

Bird, S., Klein, E. & Loper, E. (2009). *Natural language processing with python*. Sebastopol, California: O'Reilly Media, Inc.

Enomoto, A. (2008). Raitonoberu bungaku-ron [Light Novel Criticism] (in Japanese). Tokyo, Japan: NTT Shuppan.

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.