LINKÖPINGS UNIVERSITET

732A92 - TEXTMINING

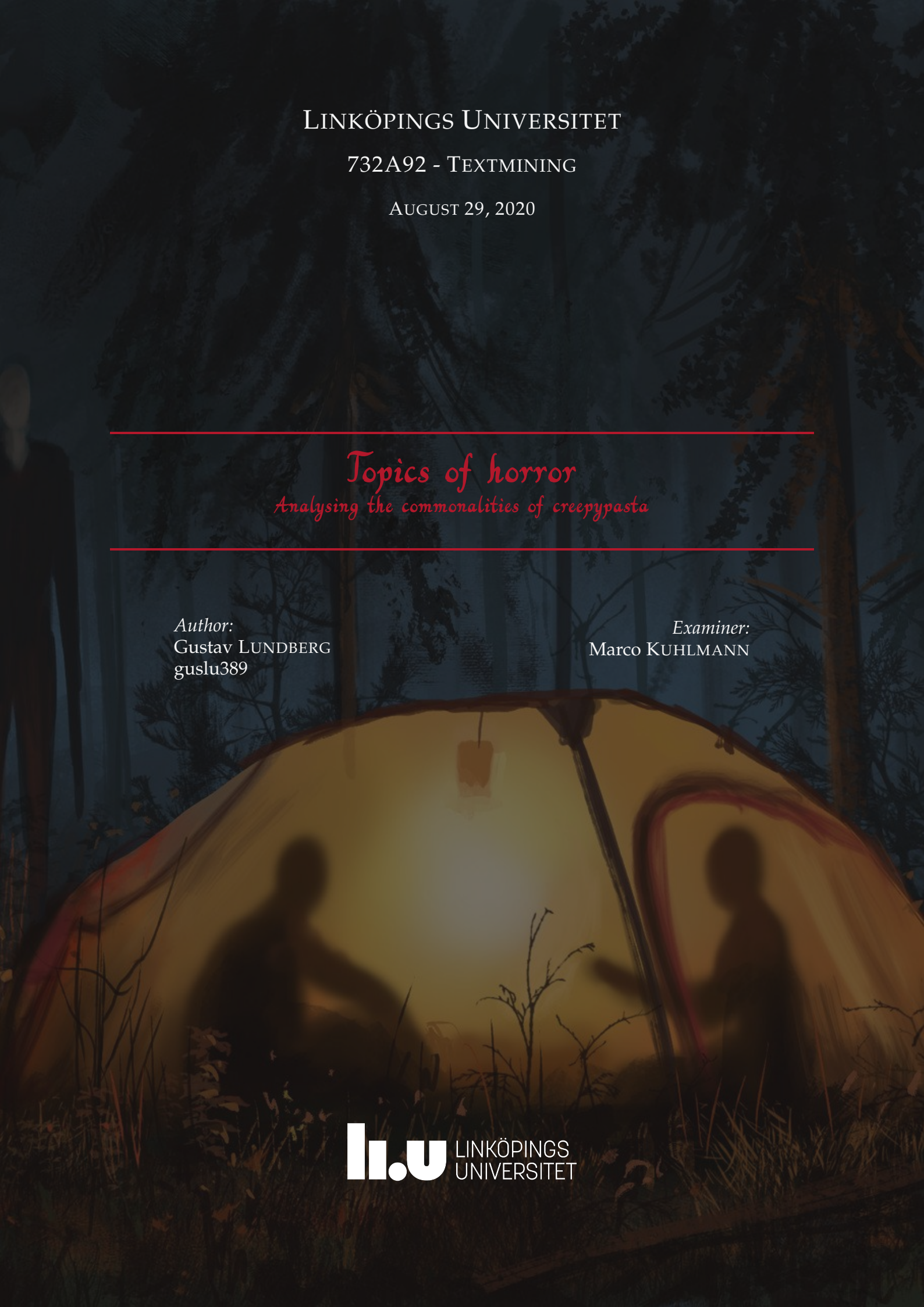
AUGUST 29, 2020

Topics of horror

Analysing the commonalities of creepypasta

Author:
Gustav LUNDBERG
guslu389

Examiner:
Marco KUHLMANN



LINKÖPINGS UNIVERSITET

Abstract

Horror stories tap in to one of humanities basic notions - fear. They are a traditional way of conveying warnings of dangerous areas and objects, but are also told for pure entertainment. It is therefore no wonder that tales of ghosts and monsters have found their place on the internet as creepypasta being shared from site to site. This way, the stories continue to evolve and adapt to the fears of modern humans, generating new characters and settings in conjunction with our inherited fears. This project aims at finding the common building blocks found in the creepypasta. To do this, almost 500 stories found on creepypasta.se are analysed using topic modelling. The resulting topics can be broadly divided into four categories of settings, characters and frightful objects. Stories that have similar distribution of topics are discussed as are stories that appear very different from each other.

Thank you to

A special thanks goes out to Jack Werner for creating his podcast and taking the time to preserve the manuscripts in written form, and also answering my questions on the topic! This project wouldn't have happened without his efforts. The podcast has been a nice companion during my daily commutes to and from Linköping over the past five years, and it deserves the honor of being included in my last (probably) academic project!

I also want to thank Stefan Koidl for his fantastically inspiring art works and letting me use some of it in the report. A picture says a thousand words!

Contents

1 Introduction

1.1 Background

Fear is one of the basic instincts of any advanced organism, acting to keep it away from dangerous situations. It can be an instinctual feeling, like fear of heights or spiders that apply for large parts of the human population, or it can be a learned behaviour, like a child being afraid of its room because its (dysfunctional) parent has told it there's a monster under the bed that will come out if the child doesn't behave. As such, fear can be a tool of control of anyone from pets and children to entire populations, and while it is not always the best method in the long run, it is usually effective in the moment it is used.

Since the notion of fear is so basic, it is no wonder that it also used as entertainment in various ways, and walking the thin red line between arousal and sheer panic can be an adventure. Tales of ghosts and monsters have entertained people for centuries, and like everything else, they have found their place on the internet. Here, folklore sits next to more modern gory stories of body horror and tales of what can be found on the dark parts of the web. Given how these stories are spread by being copied and pasted from one forum to another, they have become known as *Creepypasta*[5]. Jack Werner wrote a book about the phenomena in 2014[5] and later created a podcast[6] mixing stories found online as well as in books with commentary and analysis. These stories are the subject of study in report and I have listen to all of them atleast once over the last couple of years.

Werner divides the stories into a few broad categories mainly depending on the origin of the story, characters created by the internet, "natural" monsters, rituals and horrors based on real world events to name a few. But origins are not the only things that bind and separate stories, their actual contents and which fears they tap into in order to frighten and entertain the reader can be common across stories with wildly different background. On the contrary, characters like Slenderman (depicted on the title page graphic by Stefan Koidl) and Siren Head are both products of the web, one sneaks around at the edge of perception and the other can be heard for miles.

Research on horror stories in written form seem scarce, but there is some articles and books published that revolves around another medium - horror movies. One such review, by Martin Neil [2], mention evolutionary phobias like arachnophobia[1] and blood-related phobias as common elements of movies. The previous can probably be found to some extent in stories aimed at a Swedish audience, but as Neil also mentions, cultural differences exist in what induces fear and body-horror and gore-related stories are far less common here than in for instance North American literature and movies. There may also be a discrepancy between the two mediums, some things are easier conveyed on film than in writing and vice versa.

1.2 Research questions

This project aims at researching what commonalities exist in horror fiction aimed primarily at a Swedish audience. To do this, the stories presented in Creepypodden will be analysed using Topic Modelling to extract some of the main building blocks of a horror story.

The report aims at answering the following questions:

[1] Fear of spiders

- Can topic modelling find a number of common topics in horror stories?
- What are some of the common building blocks of a horror story?
- Are stories with similar topic distribution alike in the same way that stories with differing distributions are separated?

1.3 Delimitations and biases

Although the main source is a podcast, the content analysed is originally in written form and is thus suited for a text mining analysis. The stories are not transcribed from the tellings in the podcast, but rather make up the original material. The stories are selected by the podcast presenter Jack Werner and the selection is therefore likely biased in some way. Werner is however a well renowned journalist with publications in fact checking, and he aims at creating a diverse set of stories. The project does not handle the varying lengths of the stories (seen in figure 2.2) as this was found to be a potential problem too late in the process.

2 Data

The main data for this project is the set of stories told in the podcast *Creepypodden* [6]. The manuscripts for all episodes are found on the pod's manuscripts website and include both the actual stories and the interlude. These stories are extracted and annotated before applying the main analysis methods of the project.

For the data retrieval, the R-package `rvest` [7] was used. Data wrangling was performed using functionality from the `tidyverse`-packages [8] while data annotation and POS-tagging was done using the package `Udpipe` [9].

2.1 Data retrieval

The main page of the manuscripts is retrieved and all links to episodes are extracted. Each episode link is followed and the HTML-page downloaded. The actual stories are all contained within `blockquote`-tags and are extracting each of these results in a list of story texts for the episode. The episode name and number is also extracted from the episode page.

Figure 2.1: An example of the page structure, the final and starting paragraphs of two stories with some interlude in between on the left and the corresponding blockquote elements in the page source on the right.

2.1.1 A pleasant side effect

During the data retrieval phase, some episode-links caused errors in the scripts. It turned out that the site had had some issues with a plugin that had inserted some suspicious links to adult and casino sites. The site owner was made aware of this and now has a site with fewer surprises on it!

2.2 Dataset compilation

An episode of the podcast may include multiple stories, and one special anniversary story (*HOIN*, episode 100) spans multiple (5) episodes. This story will be considered as multiple stories in the analysis, however the chapters in each episode will be concatenated into one. No other stories span multiple episodes.

A few stories are divided into multiple `blockquote`-tags and are hence joined into one coherent story[2]. This results in a total of 498 observations that make up the final data set, a sample of which is found in table 2.1.

Table 2.1: Example of episode data prior to annotation. Only the first 200 characters of the raw_text-field are shown

story_id	story_words	raw_text
E19_S4	193	*Många klassiska skräckikoner, som till exempel surrealisten H. R. Gigers xenomorpher, Pyramid Head i Silent Hill-spelen och andra liknande skrämmande varelser, delar några centrala drag i sina utseend...*
E41_S1	545	*En tidig höstmorgon vid 06.30-tiden åkte jag och en granne ut för att jaga i Värmlandsskogarna. Det var disigt och dimmigt. Jag medförde en kortbent jakthund. Vi skulle jaga rådjur och vi stannade bil...*
E47_S1	656	*Jag hade en kompis till mig som vi kan kalla för Daniel. Daniel arbetade som nattvakt i Stockholm på ett litet större vaktbolag. I regel trivdes han ganska bra med sitt jobb. Så länge det inte hände n...*
E128_S9	1991	*Gården där min morfar var född och uppvuxen ligger i en liten by omgiven av tät granskog, berg och svarta tjärnar i mörkaste Dalarna. Den ligger där än idag, nu när jag sitter och skriver detta, preci...*
E28_S3	4031	*Jag uppfattar inte mig själv som en snokade person. Iallafall inte mer snokande än någon annan. Jag är bara vad min mamma skulle kalla ohälsosamt nyfiken. Jag var den ungen som klättrade upp i toppen ...*

The stories retrieved are of varying length, from just a few hundred words up to over 10 000 words. A distribution of story length is found in figure 2.2. This may pose a problem in the modelling as implied by Sbalachero and Eder [4], but that issue was discovered too late in the process to be handled. This may be a potential future project.

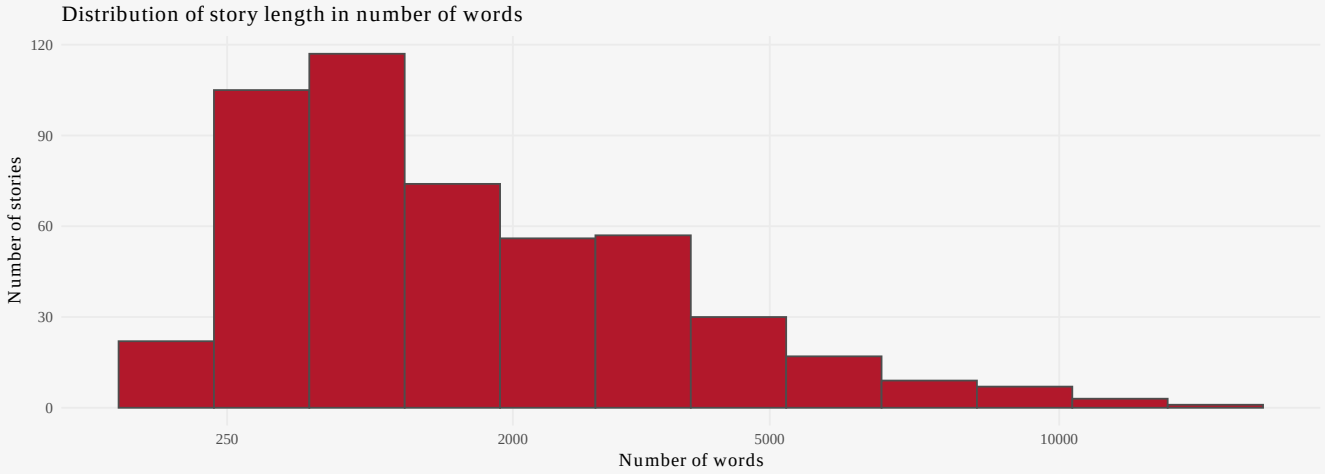


Figure 2.2: Distribution of story lengths by number of words, note that the x-axis is non-linear.

[2]See the data handling code for details

The texts are converted to normalized UTF-8[3] and then passed to `udpipe_annotate()` which returns a data frame with a row for each token, its POS-tag, lemma and various identifiers. The model used for annotation is *Swedish Talbanken, v2.4*. The main reason for selecting the `Udpipe`-package is that it does lemmatization of Swedish texts with acceptable results. The lemmatization done by the language agnostic model in `spaCy` looked more like stemming [4] while `NLTK` only does stemming in Swedish.

[3]The first attempts at modelling the texts resulted in topics containing only stopwords with umlaut characters. This issue was resolved after much headache by normalizing the texts to Unicode composed normal form, allowing the POS-tagging to perform correctly.

[4]Stemming is simply removing the last few letters of each word according to a set of rules.

3 Method

3.1 Topic Modelling

As mentioned; topic modelling is used to try an find a soft clustering of the stories. Since this method was used in the course, it will not be presented here as well.

Instead of just running the topic modelling on all words in the stories, it is run using sets of POS-tags. The POS-tags considered and some motivations are found in table 3.2. This makes it possible to not just interpret the set of POS-tags that create the most coherent topics, but also enables interpretation of models with different kinds of words.

Table 3.2: POS-tag sets and examples considered in the topic modelling

POS-tag	Motivation	Example
Nouns	Topics made from only things mentioned in the stories	'man', 'skog', 'röst'
Nouns, adjectives	Topics representing things and properties of things	'hus', 'gammal', 'ljus'
Nouns, verbs	Topics formed by objects mentioned and actions performed in the stories	'sova', 'soldat', 'bil'
Nouns, adjectives, verbs	Topics formed by all three word types as separate words	'kall', 'lyssna', 'grotta'
Nouns, bigrams of nouns and adjectives	Topics made from things mentioned and things along with descriptions as separate tokens	'familj', 'gammal dam', 'vit ansikte'

To find the optimal number of topics to use, all even number of topics in $[4, 28]$ are considered and models for each combination of word types and topic count are fitted. For each model, the average[5] topic coherence is calculated and the model with the highest average coherence is considered the best model and is interpreted and evaluated.

Evaluation is done by coherence - a measurement of how associated the words in a topic is - and prevalence - the frequency with which the topic appears in all documents. The actual topics are also evaluated by the terms included in them. Furthermore, some of the stories whose topic distribution are the most and least similar are commented.

The topic modelling is done using word frequency document-term matricies with a minimum occurrence of 5 needed for a term and 3 for a bigram to be included in the model.

3.2 Results clustering

Finally, the similarity of the stories as well as topics will be visualised using agglomerative clustering to create a dendogram. The topics will be clustered using the distribution of topics within them, while the stories will be clustered based on their topic distribution. In both cases, the distance between observations are calculated using cosine similarity.

[5]The arithmetic mean is used as suggested by Röder et al. [3]

4 Result

4.1 Number of topics

In total, 75 LDA-models were fitted to the data and the average topic coherence for each is displayed in figure 4.3. The fitting process took roughly 1 hour 40 minutes using four cores on a Intel i5 3570 CPU.

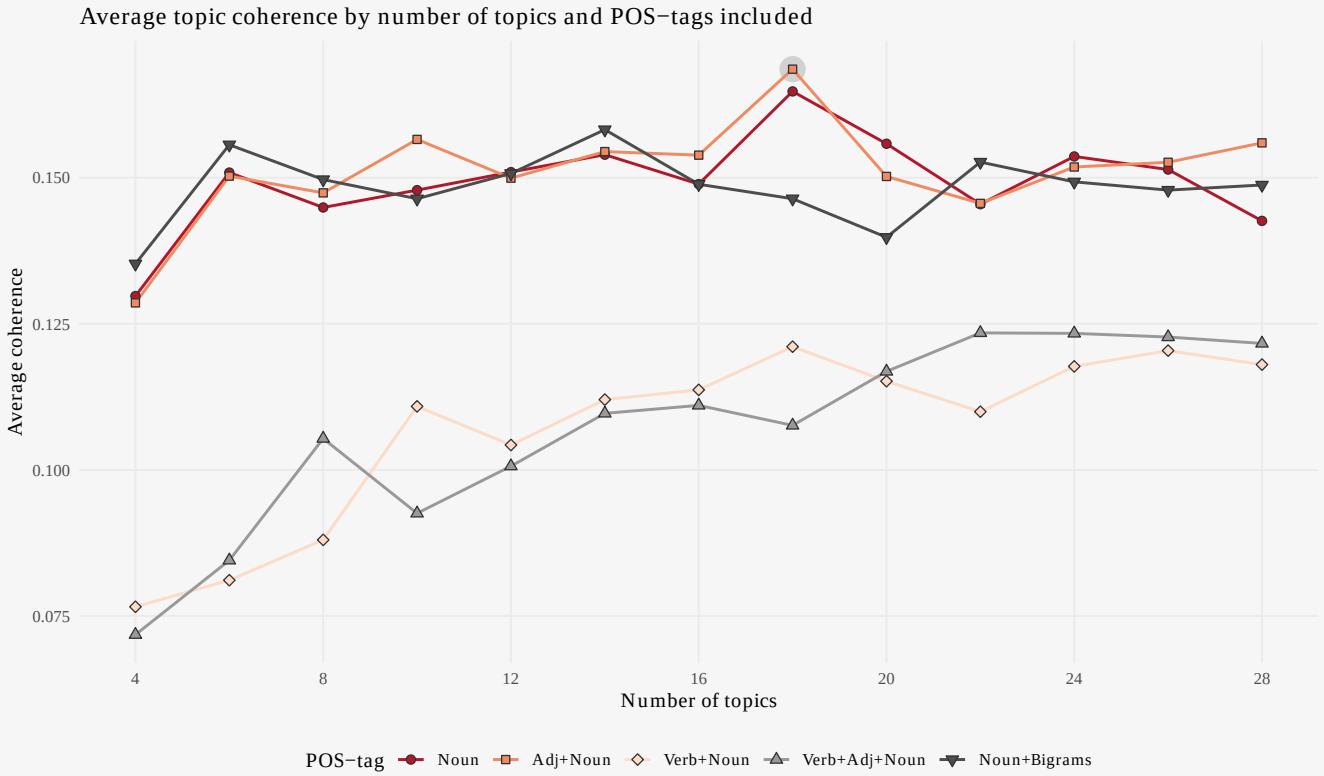


Figure 4.3: Average topic coherence by number of topics and types of words included. The model with highest average coherence is marked by a grey circle.

The models that build on verbs produce less coherent topics than the models without verbs. The remaining models are roughly equally coherent with the highest value produced by the nouns-and-adjectives-model with 18 topics.

4.2 Best model evaluation

The top-six words of each topic for this model, along with their coherence and prevalence as well as the result of the clustering of the topics, is presented in figure 4.4. Although different number of clusters are acquired depending on the height at which the dendogram is cut, a height of 1.5 gives four clusters whose topics are subjectively similar in terms of their top words. The right hand side of the coloured boxes in figure 4.4 is placed at this height over the leaf nodes.

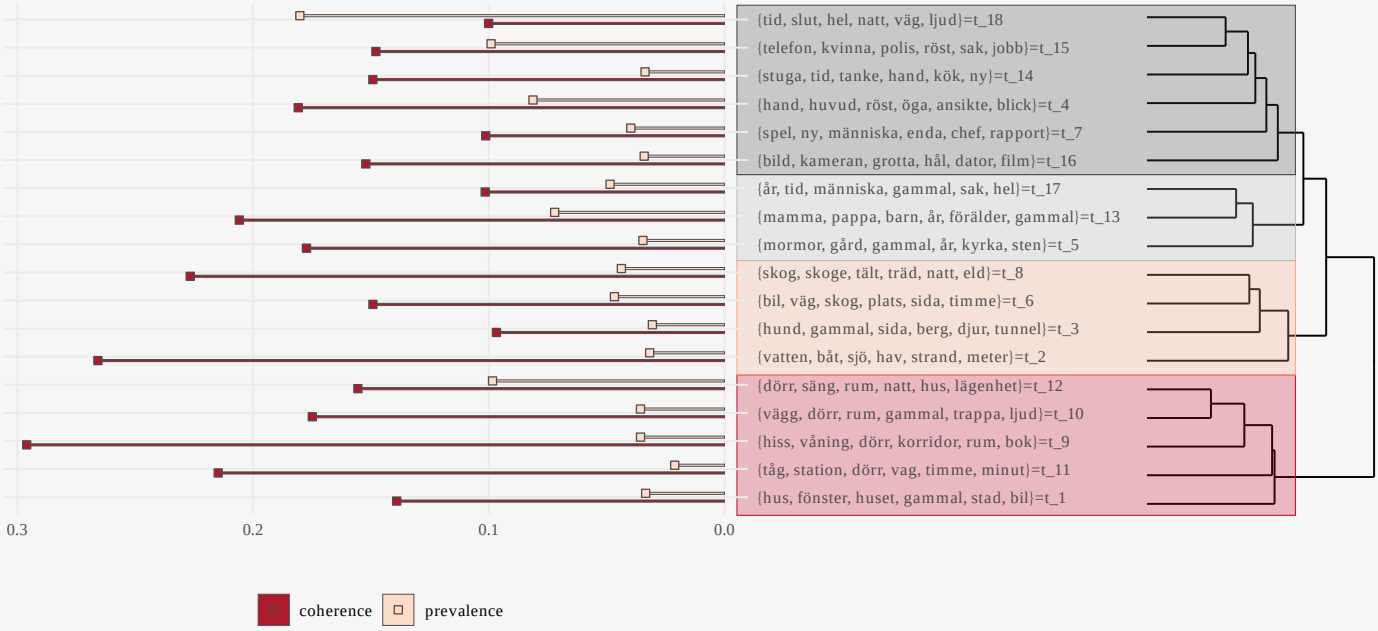


Figure 4.4: Top terms, coherence and prevalence for the best topic model.

The bottom cluster in figure 4.4 consists of topics that deals with homes and buildings in various ways. The most coherent topic (topic 9) is found in this cluster and also the second most prevalent (topic 12). The remaining topics have rather low prevalence, something that may have to do with them being rather similar and thus a single story belongs mostly to one topic rather than several.

The second cluster from the bottom is in a way the opposite of the previous, its topics deals with nature settings and travel. The topics' prevalence values are low, indicating that there are plenty of stories that don't belong to them to any large extent.

The second cluster from the top contains the topics that represent the characters of a story, while the top cluster may represent the actual horror elements of a story, the objects and feelings that actually serves to frighten the reader. These topics naturally score high on prevalence and are consistent in their coherence values.

4.3 Story similarities

Measuring the cosine similarity of stories using the document-topic-matrix, the ten most and least similar pairs of stories are found in table 4.3.

Table 4.3: Table of the most (right) and least (left) similar pairs of stories.

stories		similarity
E24_S4	E81_S3	0.9972
E11_S5	E75_S5	0.9955
E35_S3	E75_S4	0.9955
E67_S10	E125_S1	0.9949
E51_S4	E51_S6	0.9943
E13_S3	E26_S6	0.9940
E69_S5	E112_S2	0.9933
E100:1_S1	E100:2_S1	0.9902
E112_S11	E112_S13	0.9894
E36_S1	E123_S1	0.9894

stories		similarity
E20_S4	E43_S2	0.0016
E20_S4	E59_S4	0.0021
E11_S3	E43_S2	0.0022
E22_S2	E115_S1	0.0023
E20_S4	E53_S1	0.0026
E20_S4	E118_S4	0.0026
E11_S3	E61_S2	0.0027
E11_S3	E13_S1	0.0027
E9_S2	E11_S3	0.0028
E13_S1	E29_S1	0.0028

The most similar pair of stories are both tales of people in the backseat of a car, with a touch of dark roads and friendly truckdrivers. The next pair are both stories about weird encounters on a road trip, and the third couple deals with nightly visitors. Both of these contain stories from episode 75, an episode where the listeners told stories of events that reminded them of previous stories in the podcast. A bit further down are two parts of the special multi-episode story *HOIN* that are expected to be fairly similar to each other, as are the stories from episode 51 that are also part of the same overall story.

The list of most dissimilar stories contains numerous mentions of the fourth story from episode 20, a story told in the form of an SMS-conversation between a person experiencing a burglary and her friend. This makes it stand out from most other stories in terms of style rather than actual content. The same goes for the second story of episode 43, which is a tale of a haunted stage play. This is a translation of a story from the SCP-wiki (which could be a source for a future project in and of itself) and is one of the more unique tales on that site as well. Story three of episode 11 is different in such a way that there is an implied link between the events told and a missing person while the frst story from episode 13 stands out as being a straight up guide for a ritual.

5 Discussion

5.1 Data retrieval and preprocessing

The actual retrieval of the data is rather straight forward and simplified by the fact that all the stories are within `blockquote`-tags. There are some parts of the interlude that could pass for stories on their own, but separating those parts from the discussions of the topic is not simple. One thought would be to use the actual stories to create a model of stories and then apply it to each paragraph of the interlude to see if further stories could be extracted.

The comparatively small set of long documents may play a role in what number of topics are found. Consecutive runs returned varying results on the number of topics with the highest average coherence, but the number was always around 16-20 with most runs resulting in 18 as the optimal number of topics. On a casual inspection, the different lengths of the stories do not seem to have affected the topics found as they all contain a mix of short and long stories. As noted in the article by Sbalachero and Eder [4], it would perhaps be an idea to divide the texts into chunks of fixed word counts before the modelling.

The main challenge of the project was in finding a way to lemmatize Swedish texts. Udpipe was ultimately used as it provided a way of achieving this in a programmatic fashion that could be part of a data pipeline. It is not without error though, as revealed, for instance, by the fact that topic 8 contains both *skog* and *skoge*. It is however much better than the stemming procedures tested as these simply removed a number of characters from the end of words. The University of Gothemburg (GU) does provide their *SALDO*-lexicon as an XML-resource [1], but utilising this would have required building a custom solution with programming skills I did not consider myself having. Maybe it could be part of a future research project. GU also provides a webpage[6] that claims to do lemmatization, but this site did not work during the project time.

5.2 Topic modelling

One way to interpret the result of the topic models is that it is not what happens in the stories that bind them together, as the models including verbs tended to score a lower average coherence. Rather, it is the objects and characters - the nouns - as well as their properties and the feeling - the adjectives - that are common denominators. None of the main fear inducing topics of movies presented in Neil's review [2] are found by the modelling, indicating that movies and written horror may not have entirely common grounds. This discrepancy can also be caused by cultural differences or the curation of stories in the podcast.

While the topics in the top cluster may not seem similar to each other on the surface, the serve the same purpose in a horror story in that they contain the actual scary elements. This is an interesting thing to find as it goes deeper than simply finding words that appear together, the model finds part of the essence of a scary story. There is of course an argument to be made that since only horror stories were analysed, scary topics will be present and possible to find, and this would serve as an interesting future research: would the scary things be found if and equal amount of, say, love stories were added to the documents? The two categories can be suspected to share topics to a large extent, although topics 9 and 16 are perhaps a bit more than just 'love', but the basic premise of the categories are very different.

[6] http://spraakdata.gu.se/svedk/lem.html

One could argue that the height at which the dendogram is cut in figure 4.4 is somewhat arbitrary and a cut slightly higher would result in three clusters. While this is true from a pure clustering perspective, the top words of the topics in this upper cluster would subjectively be less similar. Of course, the character-cluster of topics can contain the actual scary elements as well, as in the story about a boy who's mother behaves very strange after a visit to the laundry room, but they are more than just scary building blocks, they are the the elements that the reader personally identifies with.

5.3 Story similarities

It is interesting to note that measuring the similarities between stories via their topic distribution actually finds stories who share the same atmosphere and scary elements rather than just being about the exact same things. The top most similar stories absolutely have their differences that make them into unique stories, but at the same time they tell the tales of women who encounters a potential horror while on the road and a professional driver with good intentions. This commonality of story content continue throughout the list of similar stories and it is a fascinating way of enjoying the podcast stories that I highly recommend.

Analysing the least similar stories reveals that the style of writing is what set stories apart. There are likely stories that score equally high on one or two topics, while being completely opposite on the remaining, indicating that they would be about the same characters but in different settings for example.

6 Conclusion

This project set out to answer the following questions:

- Can topic modelling find a number of common topics in horror stories?
- What are some of the common building blocks of a horror story?
- Are stories with similar topic distribution alike in the same way that stories with differing distributions are separated?

Topic modelling found a number of topics that can be further clustered into four main building blocks for horror stories, two that deal with the placement of the story, one with the characters in the story and one cluster of topics that are the actual scary elements.

The stories that are found to be similar to each other are so mainly by way of places and events in them. Dissimilar stories are different more in how they are presented and written. Ritualistic manuals differ from pure conversations for instance.

References

[1] Lars Borin et al. "The open lexical infrastructure of Språkbanken". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul: ELRA, 2012, pp. 3598–3602. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/249_Paper.pdf.

[2] G. Neil Martin. "(Why) Do You Like Scary Movies? A Review of the Empirical Research on Psychological Responses to Horror Films". In: *Frontiers in Psychology* 10 (2019), p. 2298. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2019.02298. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2019.02298.

[3] Michael Röder, Andreas Both, and Alexander Hinneburg. "Exploring the Space of Topic Coherence Measures". In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. WSDM '15. Shanghai, China: Association for Computing Machinery, 2015, pp. 399–408. ISBN: 9781450333177. DOI: 10.1145/2684822.2685324. URL: https://doi.org/10.1145/2684822.2685324.

[4] S. (1) Sbalchiero and M. (2) Eder. "Topic modeling, long texts and the best number of topics. Some Problems and solutions." In: *Quality and Quantity* 54.4 (2020), pp. 1095–1108. ISSN: 15737845. URL: https://login.e.bibl.liu.se/login?url=https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=edselc&AN=edselc.2-52.0-85079875858&lang=sv&site=eds-live&scope=site.

[5] Jack Werner. *Creepypasta.* B. Wahlströms, 2020. ISBN: 9789132211423. URL: https://login.e.bibl.liu.se/login?url=https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=cat00115a&AN=lkp.1078194&lang=sv&site=eds-live&scope=site.

[6] Jack Werner. *Creepypodden*. Sveriges Radio. Aug. 2015. URL: https://www.creepypasta.se/creepypodden/.

[7] Hadley Wickham. *rvest: Easily Harvest (Scrape) Web Pages*. R package version 0.3.5. 2019. URL: https://CRAN.R-project.org/package=rvest.

[8] Hadley Wickham et al. "Welcome to the tidyverse". In: *Journal of Open Source Software* 4.43 (2019), p. 1686. DOI: 10.21105/joss.01686.

[9] Jan Wijffels. *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit*. R package version 0.8.3. 2019. URL: https://CRAN.R-project.org/package=udpipe.