Information retrieval and text mining on the book series Red Rising by Pierce Brown

Benedikt Hornig (i6294146)

ABSTRACT

For this project the first two books of the book series ”Red Rising” by Pierce Brown were analyzed. For this analysis the progression of the main topics of the story were extracted and visualized by a topic river. Additionally, named entities were extracted and their relationships in the books were visualized with network graphs.
First, all the used and applied methods, models and metrics will be specified. Thereafter, the analysis process will be described and conclusions will be drawn. Finally, improvements regarding this project will be presented.

INTRODUCTION

For this project the first two books of the book series ”Red Rising” by Pierce Brown were analyzed. It is a futuristic science-fiction novel where mankind has colonized the planets in our solar system. To preserve the wealth of the ruling class, they have established a color-based social hierarchy. At the top the physically superior Golds reign above the other colors. The main character is born as a Red, the lowest class, which is mining minerals to sustain the ongoing terraforming of the planets. During the books he tries to infiltrate and overthrow this system and fight for the freedom of all humans again.

As the first part of the analysis the progression of the main topics of the story will be extracted and visualized. Subsequently, the named entities of the book will be extracted and their relationships between each other visualized. Thereafter, conclusions will be drawn and improvement regarding this project will be presented,

METHODS AND MATERIALS

This section lists all the methods which were applied during the project and which tools were used for each application.

Text Corpus

For this project the text from the first two books of the book series ”Red Rising” written by Pierce Brown was retrieved. Each document represents a book in the resulting text corpus.

Stop words

The stop word list from the python package scikit-learn (Pedregosa et al., 2011) was used and expanded. The full list of stop words can be inspected in the scikit-learn Github repository[1]

Tokenizers

One Tokenizer was explicitly used for this project for the topic modeling. For the other tasks the Tokenizers are directly included in the used package when using a specific model. They are described in the sections of these models.

For the topic modeling the word tokenizer from the nltk package (Bird et al., 2009) was used. It divides the the input string into a list of substrings by separating the detected words and punctuation.

[1] https://github.com/scikit-learn/scikit-learn/blob/main/sklearn/feature_extraction/_stop_words.py

TF-IDF

For calculating a matrix of TF-IDF features the TfidfVectorizer from the python package scikit-learn was used. In this instance, the inverse document frequency is calculated as follows:

$$Idf(t) = log(\frac{n}{1+df(f)})$$

Topic Coherence Normalized Pointwise Mutual Information

Topic Coherence Normalized Pointwise Mutual Information (TC-NPMI) (O'Callaghan et al., 2015) was implemented and utilized:

$$\text{TC-NPMI} = \frac{1}{\binom{N}{2}} \sum_{j=2}^{N} \sum_{i=1}^{j-1} \frac{\log \frac{P(w_j,w_i)+\varepsilon}{P(w_j)P(w_i)}}{-\log P(w_j,w_i)+\varepsilon}$$

Non-Negative Matrix Factorization

During this project Non-Negative Matrix Factorization (NMF) was applied. Here, the NMF class of the python package scikit-learn was used.

Levenshtein distance

Named entity normalization was done during this project by calculating the Levenshtein distance. For this, the python package python-Levenshtein (Haapala et al., 2021) was used.

Named Entity Recognition

For the Named Entity Recognition (NER) in the project the NER-pipeline from Stanza/StanfordNLP (Qi et al., 2020) was used. The pipeline includes a Tokenizer and a BERT model which was trained on the OntoNotes dataset (Weischedel et al., 2013). The Tokenizer splits the input document into sentences then words and punctuation.

Co-reference resolution

To resolve co-references SanBERT from the AllenNLP package (Gardner et al., 2017) was applied to the documents. SpanBERT is trained to better recognize spans of text by extending the training method of BERT by "(1) masking contiguous random spans, rather than random tokens, and (2) training the span boundary representations to predict the entire content of the masked span" (Joshi et al., 2019). This package uses the Tokenizer from the python package spaCy (Honnibal and Montani, 2017) which is splitting the input document into words and punctuation.

Visualizations

Visualizations in this project have been done with Matplotlib (Hunter, 2007), Seaborn (Waskom, 2021) and Gephi: (Bastian et al., 2009). Whereas Gephi was used for the network plots and Seaborn and Mapltolib for the rest.

TOPIC MODELING

Data preprocessing

To extract and visualize the story progression the text of the books was first split into documents, each representing a chapter, and then preprocessed by deleting unwanted characters which do not represent words or spaces.

Afterwards, english stop words were eliminated which were given by scikit-learn. This list was expanded by the following words: *know, like, dont, want, say, au, did, think, says, just, youre, man, eyes, im*, because they often ranked high in the topics extracted by NMF and did not yield much benefit regarding the meaning of the topic. Here, it is worth to mention that in the books the word *au* is being used as a connector of Forenames and the family names, e.g.: Darrow *au* Augustus. Whereas, Darrow is the main characters Forename and Augustus is the family name of his patron in the second book. Hence, it is a common used word in the books.

Topic extraction

Extracting the topics was done by first translating the preprocessed documents into a vector representation. For this, the text from the documents were tokenized by the word tokenizer of the nltk python package and then formed into document-term matrix which was then forwarded to the NMF algorithm. The words which ranked highest in the results of the NMF were then selected as a topic and their quality measured by TC-NMPI.

To try different configurations the documents, each representing a chapter, were also divided into equally big amount of sentences. Ultimately, it was better to keep the chapters together rather than splitting them up into smaller pieces, because the TC-NPMI score for the complete chapters were significantly higher than when the chapters were split as seen in figure 1. There, the TC-NMPI scores for each number of topics chosen for NMF are visualized for each book. For each chosen number of topics and each book the figure shows three scores for each of the chapter splitting methods.

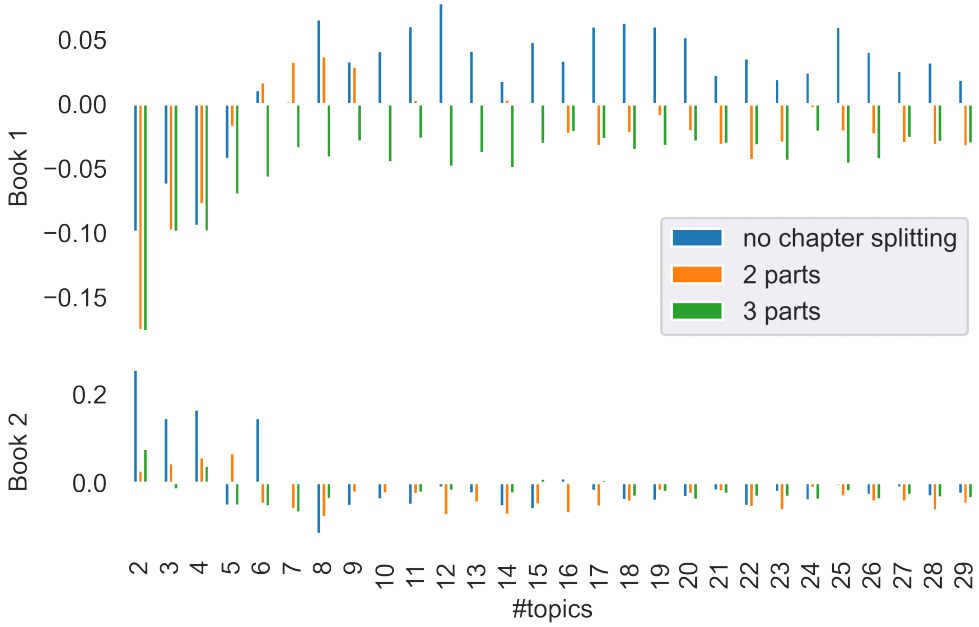

Figure 1. TC-NMPI score comparison between divided chapters and keeping chapters together

Since the goal of the topic modeling is to visualize the progression of the topics, the chosen number of topics was 8 for the first book and 6 for the second book. They both achieved a high score compared to the others and would still guarantee good structure, visibility and enough information for the visualization.

Nevertheless, it is worth noting that the second book achieved very high scores on the lower amount of topics compared to the first book. Even with knowing the books it is not clear why that is the case. Maybe it is due to fewer characters appearing more often in different parts of the second book rather than more characters appearing sequentially like in the first book. In combination, NMF is mostly considering characters of the books as important resulting in higher scores for fewer topics.

Before visualizing the topics in a topic river, the topics were analyzed by checking the highest ranking words found by the NMF algorithm. An excerpt of this list is shown in table 1 which lists the five highest ranking words for the three most occuring topics in the first book. The full table for both books can be viewed in the appendix in tables 2 and 3.

Knowing the books, one can map these topics to the events and characters in the story. Cassius for example is an ally and friend of the main character. They live and fight in house Mars together with Titus and Roque. Titus appears twice in the list of topic 3, because at the start all non-word and non-space characters where eliminated. This includes the character ’. Hence, Titus and Titus’s are viewed as two separate tokens. This could be improved by fine-tuning the preprocessing of the documents.

Topic 3		Topic 1		Topic 2	
word	NMF score	word	NMF score	word	NMF score
cassius	0.387	mustang	0.583	eo	0.214
titus	0.358	pax	0.296	laurel	0.178
roque	0.206	army	0.262	wife	0.173
house	0.153	sevro	0.242	old	0.166
tituss	0.142	jackal	0.238	dance	0.148

Table 1. 5 highest ranked words of the three most occurring topics in order extracted by NMF of the first book

Mustang is the girl the main character establishes a romantic relationship with and Eo is the name of his wife who dies at the start of the book. At start of the book the main character and his clan can win a laurel if they mine the most materials in their area.

Visualization

In summary, these topics can be easily connected to events in the book. These topics can also be visualized to strengthen this connection even further. Therefore, a topic river was created for each of the books as seen in figures 2 and 3. In these figure the topics are represented by the highest ranking word and the time line represents each chapter in each book.

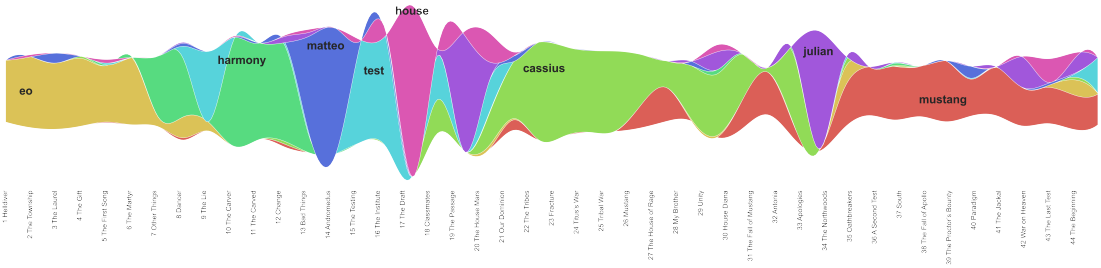

Figure 2. Topic river visualizing the events of the first book

In the first topic river (figure 2) the most prominent topic at the start of the book is the main characters wife Eo. Here, the main motive of the main character gets established: He wants to undermine the social system, because his wife was killed by the most powerful in this system. Afterwards, he joins an organization with the same goal as him named Sons of Ares. He gets recruited by Harmony, transformed by Mickey and then trained by Matteo. Thereafter, he goes to the training centre of the Golds. Here, he has to prove himself by joining a house and fighting for it in an arena. One of his first allies and friends he makes is Cassius. Cassius's brother is Julian, who gets killed by the main character before him and Cassius become friends. The second time Julian appears in the topic river is when Cassius finds out about the murder. Afterwards, Cassius seeks revenge on the main character by wounding him fatally and Cassius is understandably not a topic anymore. Here, is where the main character gets found by Mustang who heals him and cares for him.

Clearly, one can track the progression of the story in the first book very well by this visualization of the results of the NMF.

At the start of the second book, the main character has successfully found a golden liege with a high social status. During this book he is travelling more often and the characters introduced in the first book appear more often in different parts of the book. This is also visualized by the topic river in figure 3 as the topics change more often than in the first book.

In the second book, the main character is challenged by Tactus, a member of house Bellona, a different highly influential house. Additionally, one can view when the main character is in contact with the Sons of Ares, vsiualized by the topic Mickey, who is a member. Another challenge is the political relationship between the main character and the closest circle of the Sovereign. This is represented by the topic Fitchner with whom the main characters interacts the most with in this circle. A different political challenge is the closest advisor of his liege. Here, he constantly is doubted by the advisor Pliny. The last

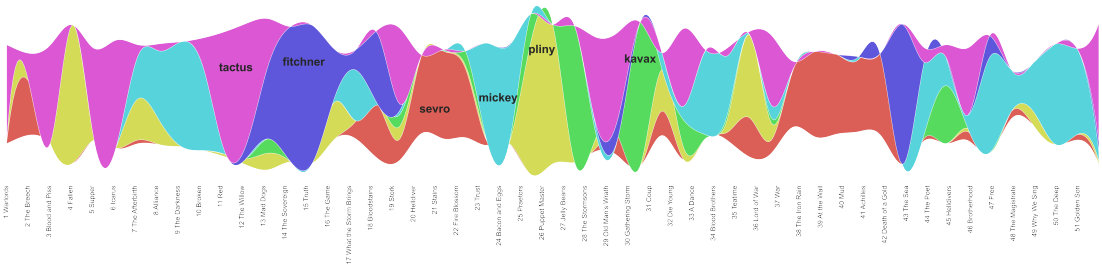

Figure 3. Topic river visualizing the events of the second book

big topics are when the main character and his allies start an assault on Mars represented by the red part and after that he realises that Fitchner was the leader of the Sons of Ares before Fitchner dies represented by the dark blue part. Naturally, the Sons of Ares have to reorganize after their leader dies which is shown in the last section of the topic Mickey.

In short, the second book was harder to visualize, because there are more frequently recurring topics. Still, the resulting topics and their progression can be comprehended by analysing the topic river and the highest ranked words of each topic.

NAMED ENTITY LINKING

Data preprocessing

Getting good results on recognizing and linking named entities from the books did not require any preprocessing of the data. Therefore, the books were only split into their chapters and then the recognizing of the entities started.

Named entity recognition

At first, the documents were forwarded into the NER pipeline of the python package stanza. It performed word tokenization and afterwards NER with a BERT model trained on the OntoNotes dataset. Here, 2,239 different entities were recognized. The results were then filtered by the tags: *ORG, PERSON, LOC, NORP, GPE, PRODUCT, FAC, WORK_OF_ART, EVENT*. These tags were chosen since the other tags did not provide important information. This reduced the amount of entities to 1,499.

It is worth to mention that some organisations, characters and Locations in the books were categorized to a different tag, e.g.: The entity *the Sons of Ares* was recognized as a work of art. Although this is not surprising as the model has not been trained on the books, it is worth to mention as this is impacting the results. After the filtering the tags of the found entities were disregarded.

After the NER, the resulting entities were normalized using the Levenshtein distance grouping entities which had a similarity score greater than 0.8. Consequently, the amount of entities were reduced further to 999. This was a very big reduction in entities, but since the threshold was set high this should improve the result further rather than finding mostly wrong pairings. Nevertheless, some entities were grouped together even though they did not resemble the same entity in the book, for example.: *Andromeda* and *Andromedus* were grouped together whilst one entity representing the Andromeda Galaxy and the other representing a character. In addition, some entities were not grouped together even though they should have been, for instance.: the entities *Adrius* and *Adrius au Augustus* describe the same character in the book but did not have a big enough similarity score. This happened mostly when comparing the forename of a character to his full name.

Co-reference resolution

To improve the linking of the entities co-reference resolution was applied to both books by using the SpanBERT model. The model word tokenized the input documents and thereafter performed co-reference resolution. The result returned the recognized spans (or entities) and their references in the text in each chapter.

These results were then combined with the normalized Entities from the other model. In hindsight normalizing first and then applying co-reference resolution was probably not a good practice than first resolving co-references and then normalizing the found entities. However, the results were still satisfactory.

In summary, the results included the recognized entity and their appearances in each chapter of each book.

Visualization of the entity links

Visualizing the links between the found entities was done by counting the common appearances of pairs of entities in each chapter. Thereafter, these links were filtered to a count greater than 15 to maintain visibility and visualized in a network graph (figure 4) by using Gephi.

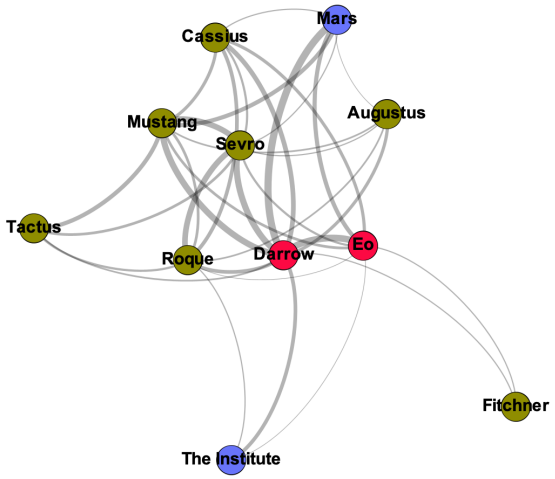

Figure 4. Relationships of the found entities with common appearances greater than 15

In this network graph character are colored by their original color in the social system and Places were colored blue.

This network graph shows the main character Darrow in the center with very strong connections to his wife and his friends Mustang, Sevro, Roque. Additionally, he is strongly connected to the Mars. On the other hand, the location *The Institute* is connected only to Darrow, Roque and Eo even though Mustang, Cassius, Sevro and Fitchner also trained at the Institute. Nevertheless, the connection to Eo is comprehensible, because Darrow reminded himself frequently to why he was enduring the brutalities during his training at the Institute.

Clustering named entities

Clustering the named entities based on the common appearances in the books resulted in the network graph shown in figure 5. For the clusters the chapters were split by a sliding window approach into parts of 10 words and a stride of 8 words.

The clustering was done with the Leiden algorithm measured by the Constant Potts Model. The resulting score was 0.776 for the resulting clusters. In the clusters the organization Sons of Ares (top right, blue) and the Bellona family (top left, pink) are clearly distinguishable from the other nodes. The big cluster in the middle is representing the characters with which the main character mostly interacts.

FURTHER IMPROVEMENTS

The project yielded satisfying results, but there are still improvements to be implemented into the process. The data preprocessing of the text was too strict to only allow words and spaces. Thus, disallowing some language constructs like contractions to be recognized by the tokenizers. Allowing these constructs could result in better tokenization and thus better topic extraction.

Another improvement could be to normalize the results of the co-reference resolution, because the model detected spans which can be viewed as Entities. This way NER is not applied by different models on the data. Additionally, one could further improve the results of the BERT models by validating them

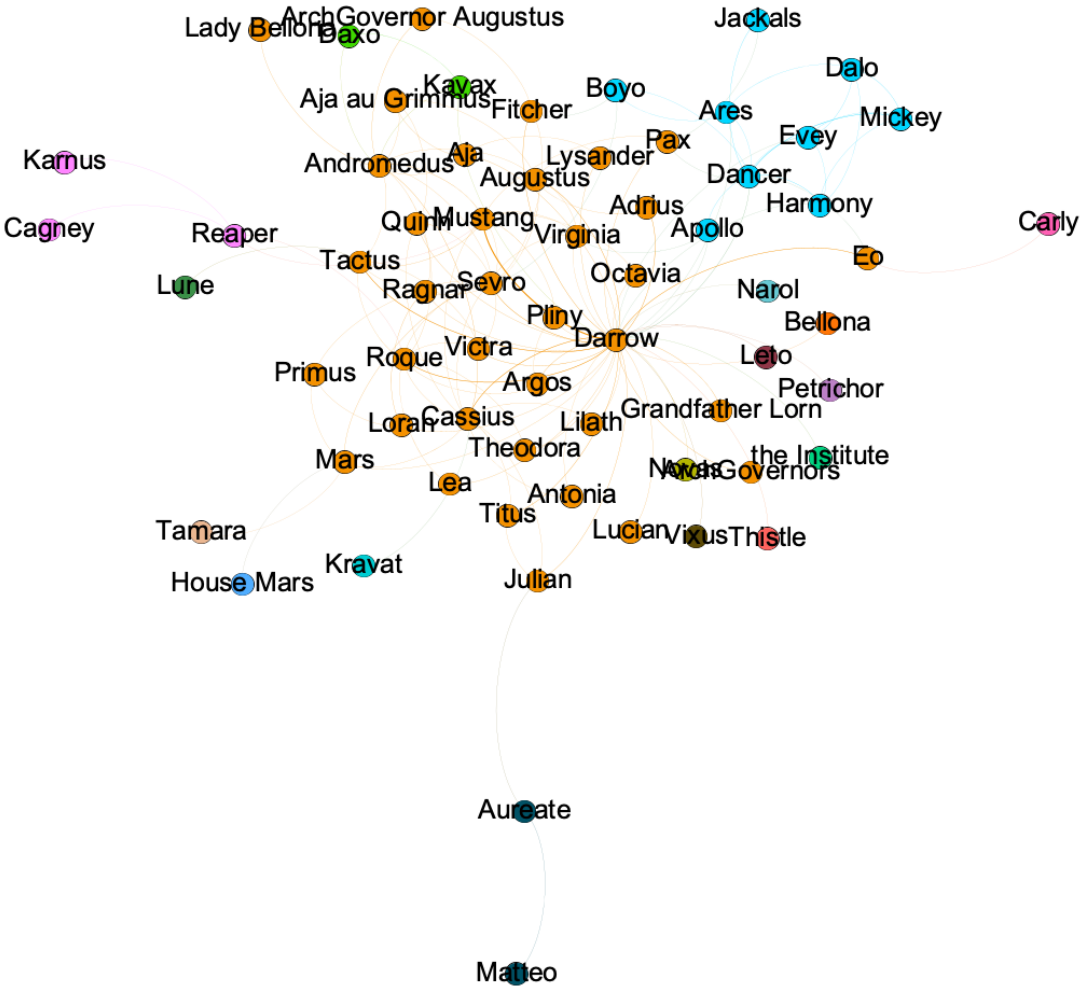

Figure 5. Relationship clusters

by a hand labeled dataset rather than only visually validating. This way one could improve the parameters of the algorithms better and could get better results.

The clustering could also be improved by adjusting the parameters resulting into more easily distinguishable clusters.

CONCLUSIONS

For this project the first two books of the book series ”Red Rising” by Pierce Brown were analyzed. For this analysis the progression of the main topics of the story were extracted and visualized by a topic river using TF-IDF and NMF scored by TC-NMPI. Additionally, named entities were extracted and co-references were resolved by BERT models. Thereafter, the relationships of these entities in the books were visualized with network graphs.

REFERENCES

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., and Zettlemoyer, L. S. (2017). Allennlp: A deep semantic natural language processing platform.

Haapala, A., Määttä, E., CD, J., Ohtamaa, M., and Necas, D. (2021). Ztane/python-levenshtein: The levenshtein python c extension module contains functions for fast computation of levenshtein distance and string similarity.

Honnibal, M. and Montani, I. (2017). spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2019). SpanBERT: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529*.

O'Callaghan, D., Greene, D., Carthy, J., and Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645–5657.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., El-Bachouti, M., Belvin, R., and Houston, A. (2013). Ontonotes release 5.0.

APPENDIX

Topic 1

word	NMF score
mustang	0.583
pax	0.296
army	0.262
sevro	0.242
jackal	0.238
tactus	0.226
apollo	0.225
house	0.218
proctors	0.195
howlers	0.168

Topic 2

word	NMF score
eo	0.214
laurel	0.178
wife	0.173
old	0.166
dance	0.148
uncle	0.147
father	0.143
song	0.138
drill	0.136
tinpots	0.128

Topic 3

word	NMF score
cassius	0.387
titus	0.358
roque	0.206
house	0.153
tituss	0.142
sevro	0.135
vixus	0.122
tribe	0.111
castle	0.11
lea	0.103

Topic 4

word	NMF score
harmony	0.362
mickey	0.355
dancer	0.238
ares	0.095
make	0.091
body	0.09
puzzles	0.081
wings	0.078
mickeys	0.075
skin	0.073

Topic 5

word	NMF score
test	0.135
matteo	0.113
datapad	0.11
julian	0.109
green	0.103
luna	0.101
augustus	0.1
city	0.099
stylus	0.087
agea	0.086

Topic 6

word	NMF score
matteo	0.338
dance	0.203
dancer	0.163
dances	0.13
perform	0.104
darrow	0.085
baton	0.083
copse	0.083
stomp	0.083
aureate	0.083

Topic 7

word	NMF score
julian	0.306
jackal	0.146
bag	0.121
fist	0.088
kill	0.085
let	0.08
sacrifice	0.076
wouldnt	0.074
stone	0.072
strikes	0.071

Topic 8

word	NMF score
house	0.134
golden	0.13
test	0.116
lightning	0.112
students	0.1
chosen	0.096
box	0.093
elves	0.093
floatchairs	0.093
goblet	0.093

Table 2. 10 highest ranked words of each topic extracted by NMF of the first book

Topic 1	
word	NMF score
sevro	0.351
bridge	0.162
wall	0.143
ship	0.141
men	0.139
blues	0.13
ragnar	0.127
grays	0.126
mud	0.125
ships	0.125

Topic 2	
word	NMF score
pliny	0.453
augustus	0.19
leto	0.144
table	0.129
liege	0.122
power	0.113
lorn	0.106
kavax	0.105
mustang	0.105
archgovernor	0.082

Topic 3	
word	NMF score
kavax	0.52
daxo	0.414
jelly	0.238
orion	0.147
fox	0.125
percent	0.125
beans	0.111
haggle	0.093
anticipation	0.091
headim	0.091

Topic 4	
word	NMF score
mickey	0.152
dancer	0.13
evey	0.118
mustang	0.111
sevro	0.103
gold	0.099
harmony	0.097
door	0.095
look	0.086
hands	0.083

Topic 5	
word	NMF score
fitchner	0.343
aja	0.282
sovereign	0.262
oracle	0.132
virginia	0.09
boyo	0.083
box	0.081
student	0.075
olympic	0.068
praetorians	0.068

Topic 6	
word	NMF score
tactus	0.163
roque	0.158
jackal	0.127
victra	0.125
karnus	0.118
bellona	0.114
augustus	0.099
family	0.091
lorn	0.079
come	0.078

Table 3. 10 highest ranked words of each topic extracted by NMF of the second book