Abstract

Fanfiction presents an opportunity as a data source for research in NLP, education, and social science. However, answering specific research questions with this data is difficult, since fanfiction contains more diverse writing styles than formal fiction. We present a text processing pipeline for fanfiction, with a focus on identifying text associated with characters. The pipeline includes modules for character identification and coreference, as well as the attribution of quotes and narration to those characters. Additionally, the pipeline contains a novel approach to character coreference that uses knowledge from quote attribution to resolve pronouns within quotes. For each module, we evaluate the effectiveness of various approaches on 10 annotated fanfiction stories. This pipeline outperforms tools developed for formal fiction on the tasks of character coreference and quote attribution.

1 Introduction

A growing number of natural language processing tools and approaches have been developed for fiction (Agarwal et al., 2013; Bamman et al., 2014; Iyyer et al., 2016; Sims et al., 2019). These tools generally focus on published literary works, such as collections of novels. We present an NLP pipeline for processing fanfiction, amateur writing from fans of TV shows, movies, books, games, and comics. Fanfiction writers creatively change and expand on plots, settings, and characters from original media, an example of “participatory culture” (Jenkins, 1992; Tosenberger, 2008). The community of fanfiction readers and writers, now largely online, has been studied for its mentorship and support for writers (Evans et al., 2017) and for the broad representation of LGBTQ+ characters and relationships in fan-written stories (Lothian et al., 2007; Dym et al., 2019). Fanfiction presents an opportunity as a data source for research in a variety of fields, from those studying learning in online communities to social science analysis of how community norms develop in an LGBTQ-friendly environment. For NLP researchers, fanfiction provides a large source of literary text with metadata, and has already been used in applications such as authorship attribution (Kestemont et al., 2018) and character relationship classification (Kim and Klinger, 2019).

There is an vast amount of fanfiction in online archives. As of March 2021, over 7 million stories were hosted on just one fanfiction website, Archive of Our Own, and there exist other online archives of similar or even larger sizes (Yin et al., 2017). We present a pipeline that enables structured insight into this vast amount of text by identifying sets of characters in fanfiction stories and attributing narration and quotes to these characters.

Knowing who the characters are and what they do and say is essential for understanding story structure (Bruce, 1981; Wall, 1984). Such processing is also useful for researchers in the humanities and social sciences investigating identification with characters and the representation of characters of diverse genders, sexualities, and ethnicities (Green et al., 2004; Kasunic and Kaufman, 2018; Felski, 2020). The presented pipeline, which extracts text related to characters in fanfiction, can assist researchers building NLP tools for literary domains, as well those analyzing characterization in fields such as digital humanities. For example, the pipeline could be used to explore how characters are voiced and described differently when cast in queer versus straight relationships.

The presented pipeline contains three main modules: character coreference resolution, quote attribution, and extraction of “assertions”, narration that relates to particular characters. We incorporate new and existing methods into the pipeline that perform well on an annotated set of 10 fanfiction stories. This includes a novel method using
quote attribution information to resolve first- and second-person pronouns within quotes.

Fanfiction is written by amateur writers of all ages and education levels worldwide, so it contains much more variety in style and genre than formal fiction. It is not immediately clear that techniques for coreference resolution or quote attribution that perform well on news data or formal fiction will be effective in the informal domain of fanfiction. We demonstrate that this pipeline outperforms existing tools designed for formal fiction on the tasks of character coreference resolution and quote attribution (Bamman et al., 2014).

Contributions. We contribute a fanfiction processing pipeline that outperforms prior work designed for formal fiction. The pipeline includes novel interleaving of coreference and quote attribution to improve the resolution of first- and second-person pronouns within quotes in narrative text. We also introduce an evaluation dataset of 10 fanfiction stories with annotations for character coreference, as well as for quote detection and attribution.

2 Fanfiction and NLP

Data from fanfiction has been used in NLP research for a variety of tasks, including authorship attribution (Kestemont et al., 2018), action prediction (Vilares and Gómez-Rodríguez, 2019), fine-grained entity typing (Chu et al., 2020), and tracing the sources of derivative texts (Shen et al., 2018). Computational work focusing on characterization in fanfiction includes the work of Milli and Bamman (2016), who found that fanfiction writers are more likely to emphasize female and secondary characters. Using data from WattPad, a platform that includes fanfiction along with original fiction, Fast et al. (2016) find that portrayals of gendered characters generally align with mainstream stereotypes.

We are not aware of any text processing system for fanfiction specifically, though BookNLP (Bamman et al., 2014) is commonly used as an NLP system for formal fiction. We evaluate our pipeline’s approaches to character coreference resolution and quote attribution against BookNLP, as well as against other task-specific approaches, on an evaluation dataset of fanfiction.

3 Fanfiction Processing Pipeline

We introduce a publicly available pipeline for processing fanfiction.1 This pipeline is a command-line tool developed in Python. From the text of a fanfiction story, the pipeline extracts a list of characters, each mention of a character, as well as what each character does and says (Figure 1). More specifically, the pipeline first performs character coreference resolution, extracting character mentions and attributing them to character clusters with a single standardized character name (Section 3.1). After coreference, the pipeline outputs quotes uttered by each character using a sieve-based approach from Muzny et al. (2017) (Section 3.2). These quote attribution results are optionally used to aid the resolution of first- and second-person pronouns within quotes to improve coreference output (Section 3.3). In parallel with quote attribution, the pipeline extracts “assertions”, topically coherent segments of text that mention a character (Section 3.4).

1The pipeline is available at https://github.com/michaelmilleryoder/fanfiction-nlp.
3.1 Character Coreference Module

The story text is first passed through the coreference resolution module, which extracts mentions of characters and attributes them to character clusters. These mentions include alternative forms of names, pronouns, and anaphoric references such as “the bartender”. Each cluster is then given a single standardized character name.

Coreference Resolution. We use SpanBERT-base (Joshi et al., 2020), a neural method with state-of-the-art performance on formal text, for coreference resolution. This model uses SpanBERT-base embeddings to create mention representations and employs Lee et al. (2017)’s approach to calculate the coreferent pairs. SpanBERT-base is originally trained on OntoNotes (Pradhan et al., 2012). However, we further fine-tune SpanBERT-base on LitBank (Bamman et al., 2020), a dataset with coreference annotations for works of literature in English, a domain more similar to fanfiction.

Character Standardization. We then assign representative character names for each coreference cluster. These names are simply the most frequent capitalized name variant, excluding pronouns and address terms, such as sir. If there are no capitalized terms in the cluster or if there are only pronouns and address terms, the most frequent mention is chosen as the name.

Post-processing. SpanBERT-base resolves all entity mentions. In order to focus solely on characters, we post-process the cluster outputs. We remove plural pronouns (we, they, us, our, etc.) and noun phrases, demonstrative pronouns (that, this), as well as it mentions. We also remove clusters whose standardized representative names are not named entities and have head words that are not descendants of person in WordNet (Miller, 1995). Thus clusters with standardized names such as “the father” are kept (since they are descendants of person in WordNet), yet clusters with names such as “his workshop” are removed.

For each character cluster, a standardized name and list of the mentions remaining after post-processing is produced, along with pointers to the position of each mention in the text. This coreference information is then used as input to quote attribution and assertion extraction modules.

3.2 Quote Attribution Module

To extract quotes, we simply extract any spans between quotation marks, a common approach in literary texts (O’Keefe et al., 2012). For the wide variety of fanfiction, we recognize a broader set of quotation marks than are recognized in BookNLP’s approach for formal fiction.

The pipeline attributes quotes to characters with the deterministic approach of Muzny et al. (2017), which uses sieves such as looking for character mentions that are the head words of known speech verbs. We use a standalone re-implementation of this approach by Sims and Bamman (2020) that allows using the pipeline’s character coreference as input. Muzny et al. (2017)’s approach assigns quotes to character mentions and then to character clusters. We simply assign quotes to the names of these selected character clusters.

3.3 Quote Pronoun Resolution Module

Recent advances in coreference resolution, such as the SpanBERT-base system incorporated in the pipeline, leverage contextualized word embeddings to compute mention representations and to cluster these mentions from pairwise or higher-order comparisons. They also concatenate features such as the distance between the compared mentions to their representations. However, these approaches to not capture the change in point of view caused by quotes within narratives, so they suffer when resolving first- and second-person pronouns within quotes. To alleviate this issue, we introduce an optional step in the pipeline that uses the output from quote attribution to inform the resolution of first- and second-person pronouns within quotes.

Prior work (Almeida et al., 2014) proposed a joint model for entity-level quotation attribution and coreference resolution, exploiting correlations between the two tasks. However, in this work, we propose an interleaved setup that is modular and allows the user of the pipeline to use independent off-the-shelf pre-trained models of their choice for both coreference resolution and quote attribution.

More specifically, once the quote attribution module predicts the position of each quote \(q_i\) and its associated speaker \(s_i\), the first-person pronouns within the quote (e.g. I, my, mine, me) are resolved to the speaker of that quote, \(s_i\). For second-person pronouns (e.g. you, your, yours), we assume that they point to the addressee of the quote \(a_i\), which is resolved to be the speaker of the nearest
Table 1: The most popular 10 fandoms on Archive of Our Own by number of works, as of September 2018. We annotate 1 story from each fandom to form our test set.

<table>
<thead>
<tr>
<th>Fandom</th>
<th>Primary media type(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marvel</td>
<td>Comics, movies</td>
</tr>
<tr>
<td>Supernatural</td>
<td>TV show</td>
</tr>
<tr>
<td>Harry Potter</td>
<td>Books, movies</td>
</tr>
<tr>
<td>DCU</td>
<td>Comics, movies</td>
</tr>
<tr>
<td>Sherlock Holmes</td>
<td>Books, TV show</td>
</tr>
<tr>
<td>Teen Wolf</td>
<td>TV show</td>
</tr>
<tr>
<td>Star Wars</td>
<td>Movies</td>
</tr>
<tr>
<td>Doctor Who</td>
<td>TV show</td>
</tr>
<tr>
<td>The Lord of the Rings</td>
<td>Books, movies</td>
</tr>
<tr>
<td>Dragon Age</td>
<td>Video game</td>
</tr>
</tbody>
</table>

Table 2: Fanfiction evaluation dataset statistics

<table>
<thead>
<tr>
<th>Evaluation Dataset</th>
<th># stories</th>
<th># words</th>
<th># character mentions</th>
<th># quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>22,283</td>
<td>2,808</td>
<td>876</td>
</tr>
</tbody>
</table>

4 Fanfiction Evaluation Dataset

To evaluate our pipeline, we annotate a dataset of 10 publicly available fanfiction stories for all mentions of characters and quotes attributed to these characters, which is similar in size to the test set used in LitBank (Bamman et al., 2020). We select these stories from Archive of Our Own\(^2\), a large fanfiction archive that is maintained and operated by a fan-centered non-profit organization, the Organization for Transformative Works (Fiesler et al., 2016). To capture a representative range of fanfiction, we choose one story from each of the 10 most popular fandoms on Archive of Our Own when we collected data in 2018 (Table 1). Fandoms are fan communities organized around a particular original media source. For each fandom, we randomly sampled a story in English that has fewer than 5000 words and does not contain explicit sexual or violent content.

Two of the authors annotated the 10 stories for each of the tasks of character coreference and quote attribution. All annotators were graduate students working in NLP. Statistics on this evaluation dataset and the annotations can be found in Table 2.

These stories illustrate the expanded set of challenges and variety in fanfiction. In one story, all of the characters meet clones of themselves as male if they are female, or female if they are male. This is a variation on the practice of “genderswapping” characters in fanfiction (McClellan, 2014). Coreference systems can struggle to keep up with characters with the same name but different genders. Another story in our test set is a genre of fanfiction called “songfic”, which intersperses song lyrics into the narrative. These song lyrics often contain pronouns such as I and you that do not refer to any character.

For quote attribution, challenges in the test set include a variation of quotation marks, sometimes used inconsistently. There is also great variation in the number of indirect quotes without clear quota-

\(^2\)http://archiveofourown.org/
tives such as “she said”. This can be a source of ambiguity in published fiction as well, but we find a large variety of styles in fanfiction. One fanfiction story in our evaluation dataset, for example, contains many implicit quotes in conversations among three or more characters, which can be difficult for quote attribution.

Annotation details and inter-annotator agreement for this evaluation dataset are described below. An overview of inter-annotator agreement is provided in Table 3.

4.1 Character Coreference Annotation

To annotate character mentions in our evaluation dataset, annotators (two of the authors) were instructed to identify and group all mentions of singular characters, including pronouns, generic phrases that refer to characters such as “the boy”, and address terms. Possessive pronouns were also annotated, with nested mentions for phrases such as $<\text{char1}> <\text{char2}> \text{his} <</\text{char2}> <\text{char1}> <\text{sister}</\text{char1}>$. Determiners and prepositional phrases attached to nouns were annotated, since they can specify characters and contribute to characterization. For an example, $<\text{char1}> \text{an} <\text{char2}> \text{old friend of} <</\text{char2}> <\text{char1}> \text{my} <</\text{char2}> <\text{char1}> \text{parents}</\text{char1}>$. Note that “parents” is not annotated in this example since it does not refer to a singular character. Appositives were annotated, while relative clauses (“the woman who sat on the left”) and phrases after copulas (“he was a terrible lawyer”) were not annotated, as we found them to act more as descriptions of characters than mentions.

After extracting character mentions, annotators grouped these mentions into character clusters that refer to the same character in the story. Note that since we focus on characters, we do not annotate other non-person entities usually included in coreference annotations. Full annotation guidelines are available online.

To create a unified set of gold annotations, we resolved disagreements between annotators in a second round of annotation. The final test set of 10 annotated stories contains 2,808 annotated character mentions.

In Table 3, we first provide inter-annotator agreement on extracting the same spans of text as character mentions by comparing BIO labeling at the token level. Tokens that begin a mention are labeled B, tokens that are inside or end a mention are labeled I, and all other tokens are labeled O. Which mentions are identified affects the agreement of attributing those mentions to characters. For this reason, we provide two attribution agreement scores. First, we calculate agreement on mentions annotated by either annotator, with a NULL character annotation if any annotator did not annotate a mention (Attribution (all) in Table 3). We also calculate agreement only for character mentions annotated by both annotators (Attribution (agreed) in Table 3). Character attribution was labeled as matching if there was significant overlap between primary character names chosen for each cluster by annotators; there were no disagreements on this.

For all these categories, inter-annotator agreement was 0.84 Cohen’s $\kappa$ or above, “near perfect”, for character coreference (Table 3).

4.2 Quote Attribution Annotation

Two of the authors annotated all quotes that were said aloud or written by a singular character, and attributed them to a list of characters determined from the character coreference annotations. Annotation was designed to focus on characters’ voices as displayed in the stories. Thus characters’ thoughts were not annotated as quotes, nor were imagined or hypothetical utterances. We also chose not to annotate indirectly reported quotes, such as “the friend said I was very strange” since this could be influenced more by the character or narrator reporting the quote than the original character who spoke it. However, we did annotate direct quotes that are reported by other characters.

Inter-annotator agreement on quote attribution

<table>
<thead>
<tr>
<th></th>
<th>Character Coreference</th>
<th>Quote Attribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraction (BIO)</td>
<td>0.95</td>
<td>0.97</td>
</tr>
<tr>
<td>Attribution (all)</td>
<td>0.84</td>
<td>0.89</td>
</tr>
<tr>
<td>Attribution (agreed)</td>
<td>0.95</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Table 3: Inter-annotator agreement (Cohen’s $\kappa$) between two annotators for each task, averaged across 10 fics. Extraction (BIO) is agreement on extracting the same spans of text (not attributing them to characters) with token-level BIO annotation. Attribution (all) refers to attribution of spans to characters where missed spans receive a NULL character attribution. Attribution (agreed) refers to attribution of spans that both annotators marked.

---

was 0.89 Cohen’s $\kappa$ on the set of all quotes annotated by any annotator (see Table 3). Attribution agreement on the set of quote spans identified by both annotators was very high, 0.98 $\kappa$. Token-level BIO agreement for marking spans as quotes was 0.97 $\kappa$. The final test set of 10 stories contains 876 annotated quotes.

5 Pipeline Evaluation

We evaluate the pipeline against BookNLP, as well as other state-of-the-art approaches for coreference resolution and quote attribution.

5.1 Character Coreference Evaluation

We evaluate the performance of the character coreference module on our 10 annotated fanfiction stories using the CoNLL metric (Pradhan et al., 2012; the average of MUC, $B^3$, and CEAFE) and LEA metric (Moosavi and Strube, 2016).

We compare our approach against different state-of-the-art approaches used for coreference resolution in the past. Along with BookNLP’s approach, we consider the Stanford CoreNLP deterministic coreference model (CoreNLP (dcoref); Raghunathan et al., 2010; Recasens et al., 2013; Lee et al., 2011) and the CoreNLP statistical model (CoreNLP (coref); Clark and Manning, 2015) as traditional baselines. As a neural baseline, we evaluate the more recently proposed BERT-base model (Joshi et al., 2019), which replaces the original GloVe embeddings (Pennington et al., 2014) with BERT (Devlin et al., 2019) in Lee et al. (2017)’s coreference resolution approach.

Micro-averaged results across the 10 annotated stories are shown in Table 4. The FanfictionNLP approach is SpanBERT-base fine-tuned on LitBank, with the post-hoc removal of non-person and plural mentions and clusters (as described in Section 3.1). Note that these results are without the quote pronoun resolution module described in Section 3.3. Traditional approaches like BookNLP and CoreNLP (dcoref, coref) perform significantly worse than the neural models, especially on recall. Neural models that are further fine-tuned on LitBank (OL) outperform the ones that are only trained on OntoNotes (O). This suggests that further training the model on literary text data does indeed improve its performance on fanfiction narrative. Furthermore, the SpanBERT-base approaches outperform their BERT-base counterparts with an absolute improvement of 4-5 CoNLL F1 percentage points and 6 LEA F1 percentage points. Post-hoc removal of non-person and plural entities improves CoNLL precision on characters by more than 12 percentage points over SpanBERT-base OL.

5.2 Quote Attribution Evaluation

Using our expanded set of quotation marks, we reach 96% recall and 95% precision of extracted quote spans, micro-averaged over the 10 test stories, compared with 25% recall and 55% precision for BookNLP.

For attributing these extracted quotes to characters, we report average F1, precision, and recall under different coreference inputs (Table 5). To determine correct quote attributions, the canonical name for the character cluster attributed by systems to each quote is compared with the gold attribution name for that quote. A match is assigned if a) an assigned name has only one word, which matches any word in the gold cluster name (such as Tony and Tony Stark), or b) if more than half of the words in the name match between the two character names, excluding titles such as Ms. and Dr. Name-matching is manually checked to ensure no system is penalized for selecting the wrong name within a correct character cluster. Any quote that a system fails to extract is considered a mis-attribution (an attribution to a NULL character).

As baselines, we consider BookNLP and the approach of He et al. (2013), who train a RankSVM model supervised on annotations from the novel
Table 5: Quote attribution evaluation scores. Scores are reported using the respective system’s coreference (system coreference), with gold character coreference supplied (gold coreference) and with gold character and gold quote spans supplied (gold quote extraction). Attribution is calculated by a character name match to the gold cluster name. If a quote span is not extracted by a system, it is counted as a mis-attribute. Micro-averages across the 10-story test set are reported. We include Muzny et al. (2017)’s approach in the FanfictionNLP pipeline.

### 5.3 Quote Pronoun Resolution Module Evaluation

We test our approach for resolving pronouns within quotes (Section 3.3) on character coreference on the fanfiction evaluation set. We show results using gold quote attribution as an upper bound of the prospective improvement, and using quote attributions predicted by Muzny et al. (2017)’s approach adopted in the fanfiction pipeline. As shown in Table 6, post-hoc resolution of first-person (I) and second-person (you) pronouns with perfect quote annotation information (Gold QuA) substantially improves the overall performance of coreference resolution across both CoNLL and LEA F1 scores (by 1.6 and 3.5 percentage points respectively).

Similarly, coreference resolution using information from a state-of-the-art quote attribution system (Muzny et al., 2017) also results in statistically significant, although smaller, improvements across both metrics (by 0.3 percentage points and 0.8 percentage points respectively) on the 10 fanfiction stories. These results suggest that our approach is able to leverage the quote attribution outputs (speaker information) to resolve the first and second-person pronouns within quotations. It does so by assuming that the text within a quote is from the point of view of the speaker of the quote, as attributed by the quote attribution system.

Table 7 shows the qualitative results on three consecutive quotes from one of the stories in our fanfiction dataset. For the first two quotations, FanfictionNLP incorrectly resolves your/you to the char-
<table>
<thead>
<tr>
<th>Quote</th>
<th>Speaker (Muzny QuA / Gold QuA)</th>
<th>Addressee (Muzny QuA / Gold QuA)</th>
<th>FanFictionNLP</th>
<th>FanFictionNLP + I + You (Muzny QuA / Gold QuA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Alright, give me [your] phone. These questions are lame.&quot;</td>
<td>Caitlin / Caitlin</td>
<td>Cisco / Cisco</td>
<td>you = Caitlin</td>
<td>you = [Cisco / Cisco]</td>
</tr>
<tr>
<td>&quot;Would [you] rather give up showering for a month or the Internet for a month?&quot;</td>
<td>Caitlin / Caitlin</td>
<td>Cisco / Cisco</td>
<td>you = Caitlin</td>
<td>you = [Cisco / Cisco]</td>
</tr>
<tr>
<td>&quot;[You] know what, don’t reply to that one, [I] don’t want to know.&quot;</td>
<td>Cisco / Caitlin</td>
<td>Caitlin / Cisco</td>
<td>I = Cisco You = Cisco</td>
<td>I = [Cisco / Caitlin] You = [Caitlyn / Cisco]</td>
</tr>
</tbody>
</table>

Table 7: Coreference Resolution of first- and second-person pronouns in three consecutive quotes from one of the fanfiction stories in our dataset. Results show the impact of the Quote Attribution predictions on the performance of the algorithm described in Section 3.3.

6 Ethics

Though most online fanfiction is publicly available, researchers must consider how users themselves view the reach of their content (Fiesler and Proferes, 2018). Anonymity and privacy are core values of fanfiction communities; this is especially important since many participants identify as LGBTQ+ (Fiesler et al., 2016; Dym et al., 2019). We informed Archive of Our Own, with our contact information, when scraping fanfiction and modified fanfiction examples given in this paper for privacy. We urge researchers who may use the fanfiction pipeline we present to consider how their work engages with fanfiction readers and writers, and to honor the creativity and privacy of the community and individuals behind this “data”.

7 Conclusion

We present a text processing pipeline for the domain of fanfiction, stories that are written by fans and inspired by original media. Large archives of fanfiction are available online and present opportunities for researchers interested in community writing practices, narrative structure, fan culture, and online communities. The presented text processing pipeline allows researchers to extract and cluster mentions of characters from fanfiction stories, along with what each character does (assertions) and says (quotes).

6 Ethics

Though most online fanfiction is publicly available, researchers must consider how users themselves view the reach of their content (Fiesler and Proferes, 2018). Anonymity and privacy are core values of fanfiction communities; this is especially important since many participants identify as LGBTQ+ (Fiesler et al., 2016; Dym et al., 2019). We informed Archive of Our Own, with our contact information, when scraping fanfiction and modified fanfiction examples given in this paper for privacy. We urge researchers who may use the fanfiction pipeline we present to consider how their work engages with fanfiction readers and writers, and to honor the creativity and privacy of the community and individuals behind this “data”.

7 Conclusion

We present a text processing pipeline for the domain of fanfiction, stories that are written by fans and inspired by original media. Large archives of fanfiction are available online and present opportunities for researchers interested in community writing practices, narrative structure, fan culture, and online communities. The presented text processing pipeline allows researchers to extract and cluster mentions of characters from fanfiction stories, along with what each character does (assertions) and says (quotes).

We assemble state-of-the-art NLP approaches for each module of this processing pipeline and evaluate them on an annotated test set, outperforming a pipeline developed for formal fiction on character coreference and quote attribution. We also present improvements in character coreference with a post-processing step that uses information from quote attribution to resolve first- and second-
person pronouns within quotes. Our hope is that this pipeline will be a step toward enabling structured analysis of the text of fanfiction stories, which contain more variety than published, formal fiction. The pipeline could also be applied to other formal or informal narratives outside of fanfiction, though we have not evaluated it in other domains.

Acknowledgements

This work was supported in part by NSF grant DRL 1949110. We acknowledge Shefali Garg, Ethan Xuanyue Yang, and Luke Breitfeller for work on an earlier version of this pipeline, and Matthew Sims and David Bamman for their quote attribution re-implementation. We also thank the fanfiction writers on Archive of Our Own whose creative work allowed the creation and evaluation of this pipeline.

References


Alexis Lothian, Kristina Busse, and Robin Anne Reid. 2007. *Yearning void and infinite potential: Online slash fandom as queer female space*. *English Language Notes*, 45(2).


