A Data Driven Analysis Framework and Erotica Writing Assistant

Since the advent of the internet, there have been an increasing numbers user-curated pornographic fanfiction and erotic narratives. With the rising popularity of erotic fiction on digital book readers and on the silver and movie screens, the corpus of erotica is an increasingly valuable, and fascinating, dataset. We present (1) `xxxwriter`, an erotica writing assistant (2) `tagmage`, a tag suggestion engine and (3) `triplex`, an erotic fiction dataset. We additionally perform a detailed analysis of trends in both the metadata and prose present in the `triplex` dataset.

1. Background

Video pornography accounts for a significant portion of web traffic, more than Netflix, Amazon and Twitter combined [5]. An average of 450 million viewers a month flock to porn sites. 14% of searches and 4% of websites are porn sites[2].

However, there is a gender disparity in porn-viewers – 70% of men partake, whereas only 30% of women do. Analysts cite the lack of female-centric content on these websites as one of the primary reasons this disparity exists [3]

The curated erotic literature industry, which is primarily generated, curated and enjoyed by women, has recently enjoyed significant growth. This industry generated 1.438 billion in sales in 2012 [4]. Perhaps most striking are the ebook sales – erotic romance ebooks comprised for 22% of ebook sales in 2011 and 44% in 2012 [4].

Uncurated erotica, typically consisting of erotic fanfiction and web erotica, is a largely untapped monetizable medium. In its short lifetime, it has already had a profound impact on recent popular culture. Consider 'Fifty Shades of Grey', which originated as a piece of uncurated erotic fanfiction, and was later published in book form (after removing all references of copyright infringement). The book later on went on to become a bestseller and the subsequent movie become highest grossing R-rated international release, at 500 million worldwide.

2. Introduction

In this paper, we (1) present `triplex` an extensive web erotica dataset, (2) `xxxwriter`, an autocomplete feature for web erotica writers (3) `tagmage`, a tag suggestion tool that suggests sets of tags that it predicts will be successful.

2.1 The `triplex` Dataset

We present a new corpus for analysts – the `triplex` dataset. A uncurated erotic literature dataset confers the following benefits:

- **As an Analyzable Dataset**: Web erotica is comprised of textual data, it is more easily analyzable than pornography. Furthermore, typically erotica is liberally tagged and structured, partitioned via html tags, rendering it a rich medium for analysis.

- **As a Colloquial Corpus**: Web erotica provides a rich corpus for natural language processing and machine learning. Corpuses for text analysis typically range in the 8K-800K word count. There is an estimated 60,000 words of erotica on a single site.

- **For Monetary Gain**: Understanding and predicting trends in web erotica may allow for (1) engineering of popular erotica (2) predicting the success particular erotic stories for publication (3) isolating popular story elements in erotica that may be transferable to print or film.

2.2 The `xxxwriter` Writing Assistant

Web erotica is famously verbose. Stories consist of a median of 60,000 words, which is substantial considering length literary classics, such as war and peace clock in at 400,000 words. Erotica writers often accrue subscribers and adhere by a strict schedule when writing multi-part narratives.

We present `xxxwriter`, a writing assistant for erotica writers – a user demographic, under-served by our community. `xxxwriter` automates the generation of text by autocompleting text on a phrase or sentence level.

2.3 The `tagmage` Tag Suggestion Tool

We present `tagmage`, a tag generation tool that aims to predict popular combinations of tags from past trends. This tool allows writers to engineer erotica by leveraging prior trends and maximize the popularity of their work.

2.4 Contributions

In this paper, we make the following contributions:

- **`xxxwriter` writing assistant**: This assistant expedites the erotica writing process by providing sentence-level autocomplete for writers.

- **`tagmage` tag completion tool**: This tool suggests sets of tags commonly associated with the input tag.

- **data analysis and discussion**: We analyze this dataset and discuss our findings. In this segment, we include a erotica length and genre analysis.

- **the `triplex` dataset**: A dataset of 27K erotic fanfictions, comprised of 1.62 billion words, scraped from `www.archiveofourown.org` – complete with tags and other metadata.

3. Example

Consider the the following two examples (1) a scenario where an erotica writer utilizes `xxxwriter` and `tagmage` to write an annotate a story in progress (2) a scenario where an erotica writer utilizes `tagmage` to find ideas for what to write about.

3.1 Use Case 1: Continuing an Existing Story

Consider the following scenario: Emily is expected by her followers to write the latest chapter of "The Steamy Adventures of Vampire Joe Biden" by Sunday, where each chapter she has written thus far has exceeded two thousand words. She would like to expedite the writing process by leveraging past erotica without explicitly copying said erotica. Assume she enters the following sequence of words:

```
Joe Biden sucked on his lip, smiling into
```

She highlights `smiling into` and queries `xxxwriter` for suggestions on how to complete the sentence. `xxxwriter` returns with the set of suggestions listed below:

```
1. smiling into it.This kiss was everything □
   wanted and more.

2. smiling into the kiss as he sways from side to
    side.

3. smiling into □3's neck, holding him against
   the bar and ordered a cab to take him back
   now, he knows that I live in a □.

4. smiling into your kisses.

5. smiling into kisses, moaning □2's name,
   whispering I love you against my cock when I
   moved my chair closer to □'s.

6. smiling into □1's mouth and his body moved
   with renewed fluidity and grace.

7. smiling into it.This kiss was everything that
   she stood above him.
```

```
8.  smiling into kisses, moaning □3's name,
    whispering I love you's beg as you're
    stretched to the end of the speech □ is
    looking a little concerned.

9.  smiling into it in one of the most intense
    orgasm that I had escaped from her dwelling
    due to a sickness that had rocked the ship
    and disappear into nothingness.

10. smiling into his broad, sturdy chest.
```

She selects option (6) and fills in `Donald Trump` into □. Now her story reads:

```
Joe Biden sucked on his lip, smiling into Donald
   Trump's mouth and his body moved with renewed
   fluidity and grace.
```

She continues to write, and eventually gets stuck and decides to ask `xxxwriter` for a sentence.

```
Joe Biden sucked on his lip, smiling into Donald
   Trump's mouth and his body moved with renewed
   fluidity and grace. □1 winced.
```

`xxxwriter` remembers that $\square_1$ was previously bound to `Donald Trump`, and substitutes the template placeholder with his name:

```
Joe Biden sucked on his lip, smiling into Donald
   Trump's mouth and his body moved with renewed
   fluidity and grace. Donald Trump winced.
```

Assume that at this point, Emily has finished updating her erotic story and would like to tag the chapter. She enters `Vampire` into the tag. `tagmage` suggests the following tags to her:

```
We noticed your story is about 'Vampires', is
   your story also about:

  Explicit Sexual Content
  Blood Drinking
  Werewolves
  Violence
  Blood
```

She selects three additional tags, `Werewolves`, `Blood Drinking` and `Explicit Sexual Content` and then asks `tagmage` to provide more hints, and selects `Vampires` as a clue:

```
We noticed your story is about 'Vampires', does
   your story have any of these characters?

  Dean Winchester
  Sam Winchester
  Harry Styles
  Laura Hollis
  Original Female Character(s)
  Carmilla Karnstein
  Louis Tomlinson
  Original Non-Human Character(s)
  Gerard Way
  Original Male Character(s)
  Athos
  Zayn Malik
  Aramis
```

```
Pete Wentz
John Winchester
```

She notices that none of the suggestions apply to her story and submits her story to `fanfiction.net`.

3.2 Use Case 2: Starting a Story

Now suppose Emily would like to start writing a new piece of erotic fanfiction about Jazz and Prowl, Autobots from the Transformers series. She enters the `Jazz/Prowl` pairing and asks `tagmage` to provide suggestions for what she should write about. She is prompted by the following set of suggestions by `tagmage`, based on what authors have previously written about.

```
We noticed you would like to write a Jazz/Prowl
    fanfiction. Here's what other authors wrote
    about:

  Threesome - M/M/M
  Angst
  Twincest
  Bondage
  Mpreg
  Transformer Sparklings
  Alternate Universe - Canon Divergence
  Sticky Sexual Interfacing
  Explicit Sexual Content
  Sticky Sex
```

At this point, she realizes, that she really would like to write about `Sticky Sex`, and queries `tagmage` for other tags associated with sticky sex:

```
We noticed you would like to write about 'Sticky
    Sex'. Here are some related tags:

  Explicit Sexual Content
  fingering
  Bondage
  Toys
  false spikes
  Alternate Universe
  Rough Sex
  Angst
```

With this second set of tags, she can now flesh out the secondary characteristics of her erotic fiction. In short, with `tagmage`, authors are able to quickly prototype the main beats of their story.

4. Design

4.1 `tagmage` Tag autocomplete

The tag generator computes a suggested list of tags, given a single tag. In the training phase, the tag generator partitions tags into four categories (1) relationships, (2) characters, (3) freeforms and (4) warnings. The tag generator performs the following set of operations for the tags in each pair of categories:

- **Tag Frequency**: It computes the frequency of each tag over the corpus of stories, for each tag in primary category c_1 and secondary category c_2

$h[ha]^+$	haha	aa^+h	aah
$a[ha]^+$	ahaha	hn^+	hnn
f^+u^+ck	fuck	$ow+$	ow
$sh^+i^+t^+$	shit	$(ow)^+$	owowow
mm^+	mmm	o^+	oo
aw^+	aww	haa^+	haa
so^+	so	$bu?z^+$	buzz
ew^+	ew	$(ew)^+$	ew
$y^+e^+s^+$	yes	ho^+n^+k	honk
$boo+p$	boop	$o^+m^+g^+$	omg
$fu+$	fuu	$ss*h+$	shh

Figure 1: Onomatopoeia used during the text normalization process. In each case, the regex is replaced with stock word.

- **Expected Tag-Pair Frequency**: It computes the expected probability of a pair of tags occurring together by taking the product of each pair of tags.

- **Actual Tag-Pair Frequency**: It computes the actual probability of the pair of tags occurring together by counting the stories.

- **Tag-Pair Correlation Coefficient**: It takes the actual tag pair frequency relative to the expected frequency to calculate a multiplier indicating the factor of over-representation of a particular tag pair.

After computing these coefficients, the tag generator sorts the highest scoring secondary tags relative to each primary tag and indexes them. The user may then query a tag to determine which associated tags `tagmage` suggests.

4.2 `xxxwriter` Writing Assistant

The `xxxwriter` leverages a templatized N-gram Markov Chain model to generate sequences of words from a word or partial phrase. The markov chain is trained on a canonicalized subset of narratives from the `triplex` dataset. We describe the canonicalization process below:

Conversion to ASCII: We coerce the unicode stories to an approximate ascii version by leveraging the `unidXXX` library for Python.

Compilation of Slang Dictionary: Due to the colloquial nature of the dataset, we compile a slang dictionary by crawling `www.urbandictionary.com` and the online slang dictionary (`www.onlineslangdictionary.com`) for vocabulary. This vocabulary is added to our preprocessor dictionary, which already consists of the standard english and british dictionary, and allows us to handle various colloquial terms and texting abbreviations prevalent in many of these stories.

We utilize this augmented dictionary to fix the numerous misspellings and typos prevalent in these narratives. Frequent and inconsistent misspellings render the distribution of N-grams more sparse than with consistent spellings.

4.2.1 Onomatopoeia Recognition

Onomatopoeia is heavily used in erotic fiction, and highly variable across uses. Below we present an example of different manifestations of the `Oh` word:

```
Ooooh
Ohhhhh
Oooooohhhh
OOOOHHHHHH
```

Without onomatopoeia pre-processing, each instance of this particular sound effect would be considered a distinct word. Our pre-processor therefore employs a regex-based onomatopoeia recognizer. For example we would represent the Ohh sound effect as follows:

$$(O|o)^+(H|h)^+$$

We compile a list of onomatopoeia regexes which are utilized by the pre-processor to substitute the onomatopoeia with the base sound. We convert the sound to lower case to reduce the complexity of the regular expressions. In the above example, All matches of $(O|o)+(H|h)+$ will be replaced with `Ohh`. Figure **??** presents the set of onomatopoeia used.

4.2.2 Proper Noun Recognition and Templatization

We identify templatize all proper nouns prior to training the markov chain on the dataset. We utilize the augmented dictionary to identify proper nouns in the narrative. We tokenize the narrative with punctuation attached to the words and strip off the prefix and suffix punctuation. We determine if (1) the word is capitalized (2) the word does not exist in the dictionary constructed above. If both criteria are met, the proper noun is assigned an identifier and this occurrence of the proper noun, in addition to all subsequent appearances of this proper noun are replaced with the template identifier.

The templatization process allows for the markov chain to generate generic text. As the author binds new proper nouns to the template identifiers in the generated text, the templatizer concretizes future generated sentences to utilize the proper nouns the author has filled in.

4.2.3 Markov Chain Construction

The pre-processed dataset is utilized as the training set for the trigram markov chain. For this system, we utilize `markovify`, a python library that supplies a markov generator.

5. Results

We describe the experimental setup, general statistics on the corpus.

5.1 Experimental Setup

We scraped erotic stories from `archiveofourown.com` to build a corpus, consisting of 3.12 GB of data encapsulat-

ing 25571 stories and metadata. We collected these stories by scraping the search results of the website after applying filters to ensure (1) the stories are greater than 500 words (2) the stories have the *Mature* tag (3) the stories are in english. The results are sorted from most to least recent date of modification, and the resulting dataset encapsulates fictions modified from Feb 2015 to Jan 2016. The dataset consists of the first 1000 pages of the 13,675 pages of results for this study. This particular query yields 273,497 stories.

Each erotic story consists of tagsets: including fetishes, fanfiction universes and characters. In the following experimental methodology, we would like to determine which textual clues are most associated with each tag and automatically assign missing tags to unannotated text.

We also explore automatic erotica generation and introduce the concept of the erotica turing test. We introduce the concept of computer assisted erotica generation, and put forward our erotica assistant as a technical contribution.

5.2 General Trends in Word Count

Refer to figure 3 - a logarithmic box plot of the number of words per story. Note that the median number of words per story is approximately 10,000 words. For comparison, a typical young adult novel is 55,000 to 70,000 words. Note that there is a long tail on the length of fanfictions - the longest fanfiction being around 1,000,000 words. Therefore, generally we observe that the average number of words per story is quite high, with a long tail.

Refer to figure 4 - a scatter-plot of various popularity criteria against number of words in the fanfiction. We note that the distribution is roughly normal for all of these criteria around 60,000 words. Stories shorter than 80,000 words are generally popular, whereas stories longer than 60,000 words steadily drop off in frequency and popularity.

We performed a reading level analysis over a selection of 1,000 stories and find that the reading level ranges from grade 2-9, with the bulk of the fan-fictions around the 3,4 grade reading levels. The simplicity of the prose is likely one of the reasons these stories are accessible to readers.

5.3 General Trends in Tag Usage

Erotica is generally heavily annotated – these annotations describe the relationships present in the work, and the characters and characteristics present throughout the work. We present the most popular characters, relationships and tags in Table 2. We observe several interesting trends in this table. Original characters, especially female characters top the list of characters. The subsequent top characters are from various franchises, including (1) Avengers, (2) Supernatural, (3) Harry Potter, (4) Attack on Titan, (5) Sherlock Holmes and (1) One Direction. Interestingly, with the exception of Supernatural, these five franchises do not explicitly cater to their fanbase and lack a significant romantic component in their stories.

| # Fics	Top Characters	# Fics	Top Tags	# Fics	Top Relationships
1118	Original Female Character(s)	3311	Angst	479	Harry Styles/Louis Tomlinson
1072	Sam Winchester	2608	Fluff	400	Sherlock Holmes/John Watson
990	Original Characters	2403	Romance	335	James "Bucky" Barnes/Steve Rogers
878	Steve Rogers	2193	Smut	299	Derek Hale/Stiles Stilinski
687	Original Male Character(s)	1606	Alternate Universe	256	Levi/Eren Yeager
682	Castiel	1586	Oral Sex	204	Will Graham/Hannibal Lecter
665	Tony Stark	1532	Sexual Content	204	Draco Malfoy/Harry Potter
627	Harry Styles	1492	Hurt/Comfort	191	Dean Winchester/Sam Winchester
626	James "Bucky" Barnes	1469	Anal Sex	178	Evil Queen — Regina Mills/Emma Swan
601	Louis Tomlinson	1175	Explicit Language	168	Oliver Queen/Felicity Smoak
592	Natasha Romanov	1118	Violence	168	Female Inquisitor/Cullen Rutherford
580	Harry Potter	1083	Alternate Universe - Canon Divergence	165	Kylo Ren/Rey
573	Clint Barton	973	Alternate Universe - Modern Setting	155	Clarke Griffin/Lexa
572	Sherlock Holmes	955	Fluff and Smut	153	Minor or Background Relationship(s)
528	Reader	903	Slow Burn	147	Steve Rogers/Tony Stark
514	John Watson	817	Established Relationship	146	Poe Dameron/Finn
510	Stiles Stilinski	800	Fluff and Angst	144	Gabriel/Sam Winchester
485	Liam Payne	774	Sex	143	Bellamy Blake/Clarke Griffin
473	Niall Horan	748	Rough Sex	142	Cullen Rutherford/Female Trevelyan
441	Cullen Rutherford	742	Blow Jobs	142	Female Lavellan/Solas
414	Zayn Malik	734	Eventual Smut	132	Bill Cipher/Dipper Pines
410	Hermione Granger	733	Humor	132	Clarke Griffin & Lexa
402	Thor (Marvel)	726	Emotional Hurt/Comfort	131	Marco Bott/Jean Kirstein
391	Draco Malfoy	718	First Time	129	Zayn Malik/Liam Payne
389	Derek Hale	697	Slow Build	119	Hux/Kylo Ren
386	Bruce Banner	674	Masturbation	115	Captain Hook — Killian Jones/Emma Swan
384	Varric Tethras	674	Drama	111	Clint Barton/Natasha Romanov
376	Eren Yeager	663	Anal Fingering	111	Bilbo Baggins/Thorin Oakenshield
373	Levi (Shingeki no Kyojin)	638	Love	106	Frank Iero/Gerard Way
349	Dorian Pavus	615	Canon-Typical Violence	106	Ian Gallagher/Mickey Milkovich
347	Kylo Ren	608	Sexual Tension	106	Laura Hollis/Carmilla Karnstein

Figure 2: Top 20 Characters, Tags and Relationships
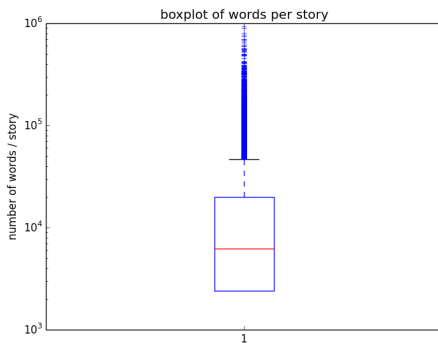
Figure 3: boxplot of log(words) / fanfiction

These characters also frequent the set of top relationships - note that we do not see original characters topping the list because original characters are likely found uniformly across fandoms. Only five of the top relationships are heterosexual relations. The majority of the top relationships are male/male relationships, and a smaller fraction are female/female relationships. The most commonly written about women are prominently featured in their respective media are well developed and have significant character arcs (`Regina Mills`, `Emma Swan`, `Female Inquisitor`, `Rey`, `Clark Griffin`). Many of the listed relationships are not, in fact, "canon", or canonical - and were rather, invented by the community.

Near the top of the tag list are the non-sexual tags `Angst`,`Hurt/Comfort`,`Emotional Hurt/Comfort`,`Fluff and Angst` - which relate to exploring and alleviating a particular character's angst / mental hangups are over-represented. This is consistent with the most popular characters, many of which have some sort of angst driven storyline or backstory. This seems to indicate redemption and recovery themes are common across fanfictions. The most popular sexual tags are `Smut`, `Oral Sex`, `Anal Sex`, `Rough Sex` and `Blowjobs`. The `Anal Sex` and `Anal Fingering` tags are likely over-represented due to the popularity of gay relationships in these stories.

5.4 Qualitative Case Study of Tag Usage

We observe the most heavily associated tags with popular relationships and characters in the Table 1. We selected four representative relationships and determined the list of the most overrepresented tags relative to the character ship. Generally, we observe that the fandom is generally, at least partially identifiable by the set of overrepresented tags. Consider the `Will Graham/Hannibal Lector` pairing, for example. Given a list of pairings, the `Cannibalism` and `Murder Husbands` tags are sufficient to pick the ro-
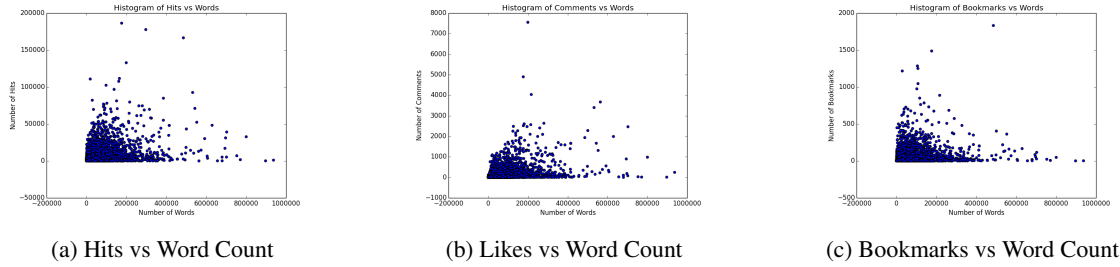
(a) Hits vs Word Count   (b) Likes vs Word Count   (c) Bookmarks vs Word Count

Figure 4: fanfiction popularity metrics vs number of words

Sherlock Holmes/John Watson	Corr.	Will Graham/Hannibal Lecter	Corr.	Hux/Kylo Ren	Corr.	Dean/Sam (Winchester)	Corr.
Explicit Sexual Content	102	Murder Husbands	39	Explicit Sexual Content	17	Explicit Sexual Content	66
Angst	87	Explicit Sexual Content	38	Pre-Canon	13	Wincest - Freeform	39
Fluff	51	Canon-Typical Violence	34	Angst	12	Angst	33
Johnlock - Freeform	43	Cannibalism	33	Kylux - Freeform	11	Anal Sex	31
First Time	43	Fluff	26	Post-Canon	8	Explicit Language	27
Anal Sex	39	Angst	22	Choking	8	Sibling Incest	25
Romance	37	Slow Burn	20	Spoilers	7	Top Dean	22
First Kiss	36	Hannigram - Freeform	16	Rough Sex	7	Brother/Brother Incest	22
Hurt/Comfort	30	Anal Sex	16	Hurt/Comfort	6	Bottom Sam	20
Blow Jobs	25	Murder	16	Crack	6	Anal Fingering	17
Established Relationship	23	Possessive Hannibal	15	Porn Without Plot	6	Kissing	17
Kissing	23	Hannibal is Hannibal	15	Breathplay	6	Hurt/Comfort	16
Alternate Universe	22	Hannibal is a Cannibal	14	First Time	6	Fluff	16
Smut	21	Alpha/Beta/Omega Dynamics	14	Snoke Ships It	5	Oral Sex	16
Fluff and Angst	21	Dark Will	13	evil space Nazi	5	Rimming	15
Drug Use	20	Alternate Universe	13	Hate Sex	5	Rough Sex	15
Slow Burn	20	Oral Sex	12	Anal Sex	5	Established Relationship	13
Oral Sex	20	Hurt/Comfort	11	BDSM	5	Blow Jobs	13
Teenlock	19	Top Hannibal	11	Alternate Universe	5	Smut	12
						Hand Jobs	12

Table 1: Most Highly Correlated Relationship Tags

mantic relationship from The Top Relationships of Table 2. The `Dean Winchester/Sam Winchester` pairing is similarly identifier - with the `Sibling Incest` and `Brother/Brother` incest tags significantly reducing the number of feasible relationships. This observation suggests that there is some correlation between the types of fetish tags and the associated relationship.

In many cases, the tags correlate with major characterization and plot points in the associated fandom. For example, the `Hux/Kylo Ren` relationship is closely correlated with the `Choking` and `Breathplay` fetishes, which is consistent with the force-choke technique jedi utilize.

5.5 TFIDF Analysis of Romantic Themes

For the TFIDF analysis, we utilized a word stemmer and regex tokenizer to identify the word roots, and performed tfidf analysis of the corpus associated with a particular tag versus the larger corpus. We studied a random selection of 2K fanfictions for this particular analysis and focused on popular tags that represent prevalent themes.

- **Explicit Sexual Content**: Stories with explicit sexual content have a large set of heavily correlated words, with a much higher prevalence than the other romantic themes: `eye`, `moan`, `kiss`, `pant`, `thrust`, `thigh`, `slid`, `lick`, `whisper`, `cheek`, `gentle`, `groan`, `move`, `stroke`, `gasp`, `waist`, `grin`, `clit`, `scent`, `blush`, `squeeze`, `muscle`, `whimper`, `grab`, `jerk`, `rub`, `tighten`, `arousal`, `lust`, `erection`. Words somewhat strongly inversely correlated with this tag include: `glance`, `stop`, `shook`, `nod`. We note that two of these words are overrepresented in the angst tag and postulate that explicit angsty stories are more likely to have other mature themes, such as violence, instead of sexual content. We note that the high correlation of particular words to the tag indicates it is possible to construct a wordlist that can detect if the work is pornographic from this analysis.

- **Love**: We note that the following in-prose words are most closely correlated with the `love` tag: `smile`, `eye`, `cheek`, `kiss`, `whisper`, `chuckle` and `love`. We similarly note that the following in-prose words are weakly inversely-correlated with the `love` tag and appear more often in the comparison corpus: `shrug`, `snort`, `nothing`, `mutter`, `murmur`, `frown`, `snap`, `ignore`.

- **Angst**: We note that the following in-prose words are most closely correlated with the `angst` tag: `shook`, `sigh`, `glance`, `nothing`, `cheek`, `chuckle`, `swallow`. The weakly inversely correlated words are `frown`, `smirk`, `surprise`, `notice`, `glare`, `nod`, `grin`, `annoy`.

5.6 Fanfiction Generation

We analyze the fidelity of our erotica markov chain in the following section. We consider three use-cases (1) generate random sentences (2) generate short sentences (3) generate completions starting with a particular word. We disable support for templatization to maximize the comedic effect of the sentence generation, but note the Markov training routine does support templatizing out identifiers.

Random Sentence Generation: Consider the following set of randomly generated sentences.

```
Shooter guided himself in-between Hiro's checks.

I ball my hand into his shorts, afterwards
    continuing to work in a team photo.

One way to get inside the truck, starting it with
    sex.

I cry, trying to burst out of her filthy clothes.

Levi said and tugged the offending shirt off of
    his broom, then shifted forward, making room
    for her to hear.

Gungnir lay by his side was enough to force me
    into a meme, still hunted me.

That was all she could and screamed at the same
    time until Lukas felt something churn inside,
    a feeling of intense pleasure that he's
    never been able to find.

I sped off to get a closer look at the prologue.

This is more Felida's style, but I think I want
    to talk to Rick and nipped at his collarbone.

There is the push he needed in his life, Armin
    kissed Jean's cheek and breathes a soft sigh
    from Inho that blended with Sangjun's groan
    of absolute despair.Why, oh why, did she just
    say?! Oh hell no he did NOT want to happen.

Connington paused a moment, his face totally
    changes and he knows that we make out.

He's already dreading a conversation he wasn't a
    good sign or not, so many things he always
    does... nothing out of the building collapsed
    ahead of him.

Perhaps he should give them some privacy.

And I saw a yellow dress that matched her
    favorite lipstick, which Rowan secretly
    called classic siren red.
```

Note that the cohesiveness of the sentence degrades with the length of the sentence. The training data is verbose enough that fandoms are typically mixed, resulting in cross-multiverse interactions between characters. Also consider the baseline porniness of these sentences is relatively tame. Since sentences are built up randomly, weighted by ngram frequency - this indicates without additional constraints, a bulk of the text in the training data is not pornographic.

At least one sentence generates a typo: `in-between Hiro's checks` – this n-gram was likely meant to be `in-between Hiro's cheeks`. Typos such as this one humanize the generated sentences and give the impression it was written by an author.

Short Sentence Generation: We note that the shorter sentences are more coherent. The next sentences where generated by applying a 250 character limit to the markov chain. The chain is sampled until a sentence that meets these constraints is generated. We eliminate sentences that are too short.

```
He'd always assumed that his brother had actually
    found me too forward.

Minho groaned and picked the first movie and he
    had always been.

Thankfully, Trelawney had already disappeared
    from the screen Bulma could see the relief
    reflected back at her.

Four cups of unspiked coffee sit on the floor and
    gives him a quick kiss, making a show of
    taking her out into the dark.

She pushed herself off her sweat covered skin.

I just can't yet tell if it's because the idea of
    saving me from that strange state in which
    he was only a flame test lab.

Daryl threw up one finger in me gently, and his
    heart rate begins to calm and settle into a
    comfortable position.

It was amazing how sexy he thought she could love
    him and he greedily explored her.

The flames turned a pale green swaddling blanket.

Lying still, he invited the lord in.

Spirits of past generations lived in its walls,
    phantoms danced in the wind as he points at
    the moment.
```

Note that these sentences are more cohesive than the prior set of sentences. Interestingly, they are more explicit, on average, than the previous generation scheme.

Autocomplete Generation: We consider the secondary use-case where we would like to autocomplete a sentence.

The following set of autocompletions are generated utilizing `hot` as the starting word. In this set of completions, the markov chain interpreted `be.Stiles` as a single word –

we observe that neglecting spacing following punctuation is a common mistake in these stories and many of the generated sentences contain mistakes of this nature.

```
hot on top of his staff to support her.

hot damn, was that a great ruler does not lead to
    ruin.

hot folds already slick with moisture, his
    movements as Peter would be.Stiles would
    chase their prey down and Peter would humor
    him.

hot water rushes over his bare stomach, the tips
    of her serviceable shoes.

hot mess right now, considering you just said,
    kisu-mi feels his knees buckle under him and
    across the sand to join her.

hot dew out through the entire room, which made
    Khan inhale deeply.

hot wet, heat firing the connection between Lex
    and Kal.

hot release quickly follow yours.
```

The next set of autocompletions are generated utilizing `fingers` as the starting word. Note that autocompletions starting with fingers contain both violent and sexual autocompletions - though it is unclear if the violent autocompletion is also sexual.

```
fingers tighten round her calves as her legs
    would be excruciatingly painful.

fingers in soft circles with his tongue.

fingers tender through the strands of my hair and
    threw it on the floor with the second ball,
    and Rin came hard..

fingers grabbing onto the stall door and carried
    Lexi into their bed.

fingers pleasing her pussy.

fingers choking him, strong as vices, until the
    air was much thicker than Harry's, but that
    made aiming the weapons mounted on top
    difficult.
```

The next set of autocompletions are generated utilizing `bit` as the starting word. Note that sentences that utilize both bit the verb and bit the descriptor are generated. The second use of the word, of course, is utilized in a more sexual context.

```
bit tense, they clearly had together.Chewing on
    his thumb with his other friends, but Tom
    became a real possibility of those after they
    had finished Hot Fuzz and were halfway
    through the second corridor on the fourth
    floor classroom they're meeting in, shift
    towards the fifth floor.
```

```
bit into skin drawing blood from one another in
    tandem, until she shivers with her orgasm,
    the tension of the previous evening before he
    was ready.

bit snobby and a bit years to make sure she could
    spot him in seconds-- it just wasn't working
    anymore?

bit cocky, and at times had to be there too, is
    Sanji doing the same thing had happened
    without Thorin's notice: Love had been placed
    upon her.

bit of dress torn away to show the colorful range
    of his hair, willing him up his body.

bit drunk, but then again, Dean wears so many
    layers of clothes and then crawls into his
    own palm.
```

The next set of autocompletions are generated utilizing the two word prefix: `hard cock` as the starting set of words. The markov chain is unlikely to generate innocuous sentences with this prefix.

```
hard cock stroke her sensitive spot again and
    again that this young alien goddess used her
    pussy to experiment on some blood to mask her
    eagerness for him to hold her in my sleep
    all the way down to where their voices were
    little more than their underwear.

hard cock touched Alexander's.

hard cock bobbed at the attention of the masked
    figures, and in one of the bowls of chocolate
    and ran over her bottom lip, not even asking
    for entrance that he would only give them
    time to find Thomas, Newt wishes he could be
    moving or joking or laughing, there was this
    tasty little morsel in his mouth.

hard cock quickly, adjusting the pillow so
    beautifully.

hard cock of the man who was becoming more and
    more into Barney's space than Barney would
    like.

hard cock pressing up against him, fist pressed
    to Naruto's stomach, the handle of the shower
    .
```

We note the `xxxwriter` text generator is able to successfully generate reasonable standalone sentences and autocomplete sentences given a word without any outside knowledge of grammatical structure. While these sentences, are not yet at the level of human-scribed fanfiction, we believe that the fidelity of the text generation may be improved with grammatical annotations and a text generation algorithm that takes into consideration that overarching grammar of the sentence. For this reason, for the time bieng, `xxxwriter` is a writing assistant rather than an automated fanfiction gen-

eration framework.

Fanfiction Turing Test: We propose a fanfiction turing test, where machine generated fanfiction is considered sufficient if a reader is unable to differentiate human-generated fanfiction from machine generated fanfiction. We postulate passing the fanfiction turing test is easier than passing the classical turing test because fanfiction is expected to have less consistency and grammatical integrity and allowed to spuriously fail coherence tests. We note, by simply introducing typos and misuses of words, the credibility of the fanfiction as a human generated work is increased.

6. Conclusion

We contribute (1) the `triplex` dataset, an expressive and heavily annotated dataset [1], (2) `xxxwriter`, an erotica writing assistant and sentence autocompleter, and (3) `tagmage` a tag suggestion engine.

7. Future Work

In the future we hope to provide the following improvements to our system.

- **example-based onomatopoeia regex generation**: We hope to leverage example-based synthesis techniques to automatically generate regular expressions for onomatopoeia and therefore reduce the developer's burden in providing sound effect specifications.

- **onomatopoeia classifier**: We will construct an onomatopoeia classifier to automatically extract training data for the regex generator by ascribing onomatopoeia classes to words in each document.

- **porniness classifier**: We are planning to construct a porniness classifier that rates the porniness of a passage utilizing information from the TFIDF analysis. From this information we can plot porniness curves and determine if there are correlations between the porniness trajectory and the popularity of the story.

References

[1] S. Achour. Artifact for ficxxx framework. URL `https://github.com/sarachour/triplex`.

[2] B. News. Web porn: Just how much is there? - bbc news, 2016. URL `http://www.bbc.com/news/technology-23030090`.

[3] P. O'CONNELL. Web erotica aims for new female customers, 2016. URL `http://www.nytimes.com/1998/08/13/technology/web-erotica-aims-for-new-female-customers.html`.

[4] P. Perspectives. Yes, yes, yesssss...! erotic romance sales still sizzle - publishing perspectives, 2014. URL `http://publishingperspectives.com/2014/01/yes-yes-yesssss-erotic-romance-sales-still-sizzle/#.VtRk4pMrIUE`.

[5] T. H. Post. Porn sites get more visitors than netflix, amazon and twitter combined, 2016. URL `http://www.huffingtonpost.com/2013/05/03/internet-porn-stats_n_3187682.html`.