

# Automatic Classification For Fanfiction Websites



A Solution For Mistagged Stories

By Sam Kamenetz

# Tags and Fanfiction

# Fanfiction and fanfiction platforms

- fan written stories based on a particular show or series (in any medium)
- websites like fanfiction.net give platform for sharing fanfiction.
- almost always include tags to describe genre and facilitate searching for content

# Typical tagging systems (Fimfiction.net)

**T** **The Man With Two Names** by donald trump card



Human

Sad

Slice of Life

Tragedy

When Fluttershy finds a human in her backyard trying to eat one of her chickens, she and a majority of the residents of Ponyville only want this dangerous creature as far away from their town as possible.

After a less than stellar first impression, Derpy Hooves and the human struggle to quell the rumors surrounding his diet, and convince the townsfolk that he's harmless.

-Author chooses any number of predefined genre tags (18)

-choose up to six characters

# Folksonomy tags (Archiveofourown.org)



An Anthem of Angels by DestinyWolfe

12 May 2016

Captain America (Movies), Captain America - All Media Types, Game of Thrones (TV), A Song of Ice and Fire - George R. R. Martin, A Song of Ice and Fire & Related Fandoms, Marvel Cinematic Universe, Marvel, Captain America/Game of Thrones crossover

**Creator Chose Not To Use Archive Warnings, Graphic Depictions Of Violence, Major Character Death,** James "Bucky" Barnes/Steve Rogers, Clint Barton & Natasha Romanov, Clint Barton/Natasha Romanov, James "Bucky" Barnes & Steve Rogers, James "Bucky" Barnes, Steve Rogers, Cersei Lannister, Loki (Marvel), Clint Barton, Joffrey Baratheon, george barnes (mentioned), Rebecca Barnes, Natasha Romanov, Brock Rumlow, Alexander Pierce, Winifred Barnes, Sarah Rogers, Joseph Rogers, Jim Morita, James Montgomery Falsworth, Stucky - Freeform, stevebucky - Freeform, and, Clintasha - Freeform, as per frickin' usual, Marvel/GoT crossover, this is going to be a trainwreck, but I'm gonna have fun wrecking the train, clintnat, Crossover, ASoIaF, GoT, some GoT characters are still themselves, and some have been replaced by appropriate Marvel equivalents, or inappropriate ones, either way, Steve is kind of a blend of several other characters, Rebecca Barnes is Sansa, Natasha is Arya, Clint is Gendry, Joffrey and Cersei are Joffrey and Cersei, unfortunately, the first chapter is really short but I promise that future chapters will be longer, if I don't update regularly I'M SORRY, my muse is just the worst, most finicky muse in the universe, Alternate Universe - Medieval, Kinda, Dark Ages, also kinda - Freeform, more explanation will come in chapter notes and such, because this is gonna be a massive trainwreck otherwise, even with notes it's gonna be a trainwreck, LMAO, Bucky is Robb, possible spoilers for various marvel movies, no spoilers for civil war, spoilers for season 1-5 of game of thrones, eventually, asoiaf spoilers, GoT spoilers, will add more tags later, Slow Burn, Romance, Love, Angst, Hurt/Comfort, but mostly there will be hurt, and no comfort, You Have Been Warned, AU, Fluff, Forbidden Love, Pining, Mutual Pining, Dark, So many tags

After the false king Joffrey Baratheon executes his father for treason, Lord James Bucky Barnes of Winterfell declares war on the South and gathers his northern bannermen to his side. But although Bucky's desire for revenge is strong, a budding romance with his childhood friend and constant companion, Steve Rogers, threatens to weaken his alliances and diminish the loyalty of the men fighting at his back.

-free form tags determined by authors, semantically similar tags linked together

-can be esoteric, overly specific

-no understanding of relative importance of tags

# Limited tags (fanfiction.net)



[Welcome to Ponyville](#) > by [SuperCycloneProductions](#) [reviews](#)

Somewhere deep down in Tartarus is a dark factory where ponies of madness work hard for their master. He sees a weak point in time and sends one of his ponies to go and visit Ponyville to find a book. But Twilight manages to find the book and discovers that it could lead to massive destruction. Join this giant cast of ponies as they discover that not everything is as it seems.

Rated: T - English - Family/Mystery - Chapters: 11 - Words: 21,655 - Reviews: 12 - Favs: 1 - Updated: Apr 26 - Published: Feb 8 - Twilight Sparkle, Rainbow Dash

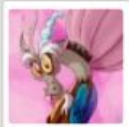
-limited to two character tags, two genre tags

-might not fully capture elements of the story

# Mistagging: why is this a problem?

- inaccurate descriptions
- authors trying to game the system
- poor understanding of tags or own story

Goneaway



#20 · 145w, 5d ago · 10 likes · 0 comments

To avoid false advertisement, ya might want to switch the Comedy tag for Tragedy, this reminds me of Planet of the Apes (the original)

**Solution: Automatic Tagging of Stories**



# Proposed experiment

- Determine if a classifier can be trained to accurately predict tags on most stories
- If so determine way to automatically know if story was mistagged by author or misclassified by classifier
  - Hypothesis: poorly tagged stories receive fewer views

# Experimental Setup

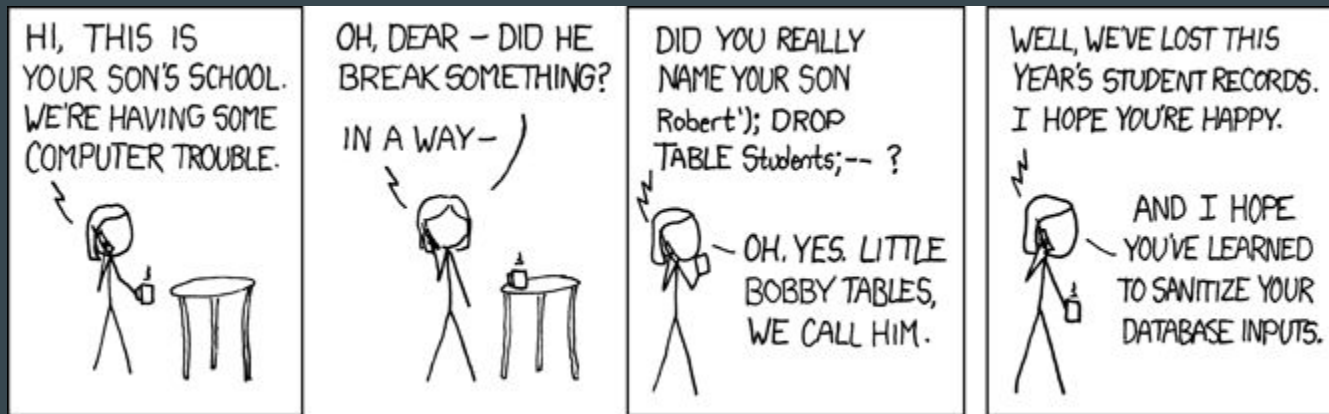
# Choosing a test set: Fimfiction.net

- users choose any number tags from predefined list
- limited to one fandom
- rich metadata (views counts, word counts, likes, etc)
- userbase created zipped archive of entire corpus available for download (5GB)



# Parsing the data, SQLite

- SQLite DB, parse through giant json file, reorganize, and store in DB
- made tables for Documents, DocumentToTags, Authors, AuthortoDocuments.
- wrote python wrapper for DB. Retrieval functions the most complicated



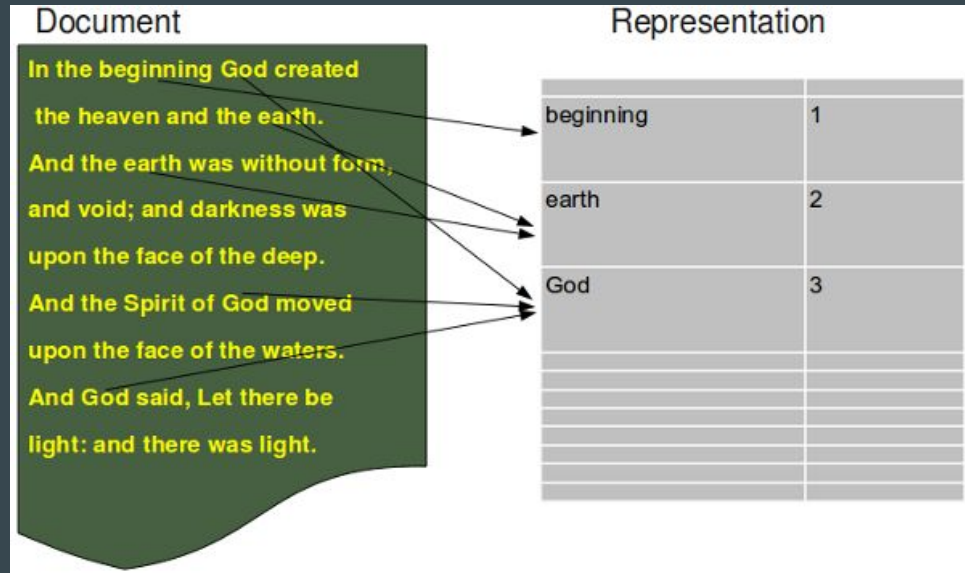
# Parsing the data: EPUBs to TXTs

- packages didn't work, had to write my own
- actually zip files, with html for chapters, plus metadata and stylesheets
- had to figure out best encoding (utf-16)
- used BeautifulSoup for parsing



# Bag-Of-Words Model and You

- concept in text classification
- Represent documents as n-dimensional vectors
- each entry in vector represents word count for particular word in given document
- does not preserve order of words



Source: [http://www.python-course.eu/text\\_classification\\_python.php](http://www.python-course.eu/text_classification_python.php)

# TF-IDF transformation

- Already explained by Nick, Beth, and Kelly. Thanks guys!
- Quick auxiliary explanation just in case:
- Stands for term frequency, inverse document frequency.
- Method of weighting the importance of words for classification
- Term Frequency: the more a word appears document, the more evidence it contributes
- Inverse Document Frequency: the more documents a word appears in, the less unique it is, and contributes less evidence

# Linear SVC

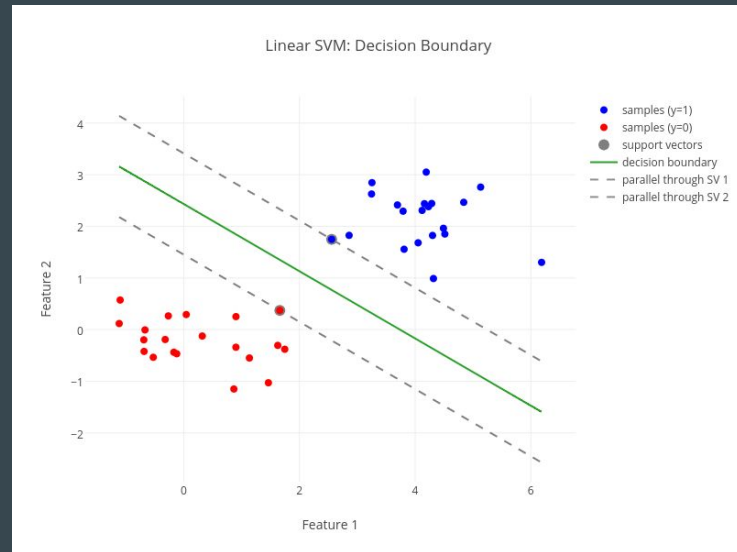
- Support Vector Classifier

- Looks at document vectors as points in n-dimensional space

- draws a “decision plane” in this space with maximum distance between the two sets of points

- class determined by which side of classifier the test point falls on

- score of -1 to 1, magnitude indicates degree of evidence



<https://plot.ly/~florianh/157/linear-svm-decision-boundary/>

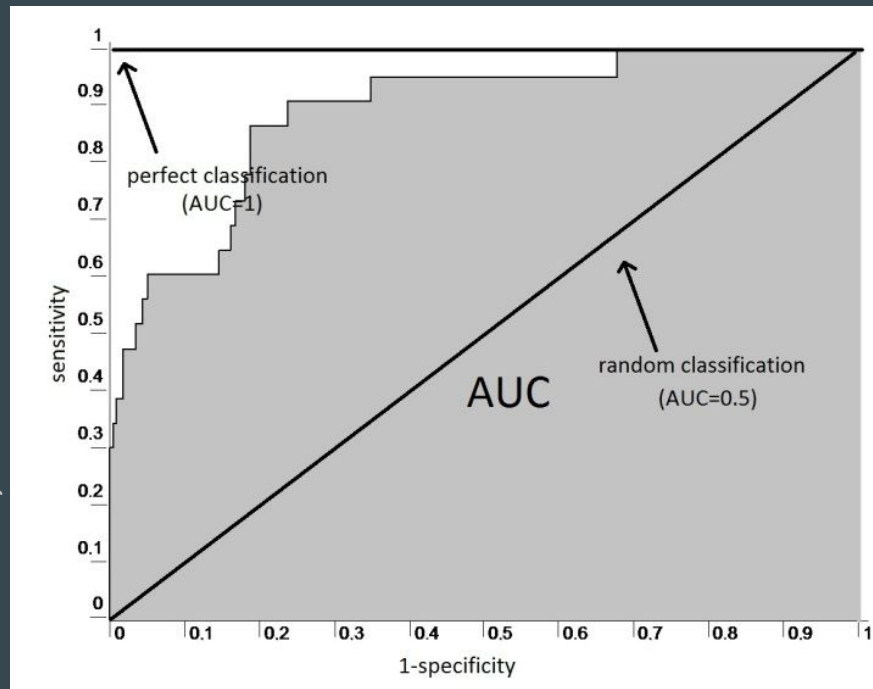


# Multilabel classification

- Linear SVC normally used for Multiclass problems, eg choosing one class from a list of all possible classes
- fanfiction problem is multilabel: must choose every class that fits
- OneVsRest Classifier in Sci-Kit Learn: series of binary classifiers (belongs to class or does not belong to class) bundled together, operating on the same test set

# Evaluation of classifier: AUC metric

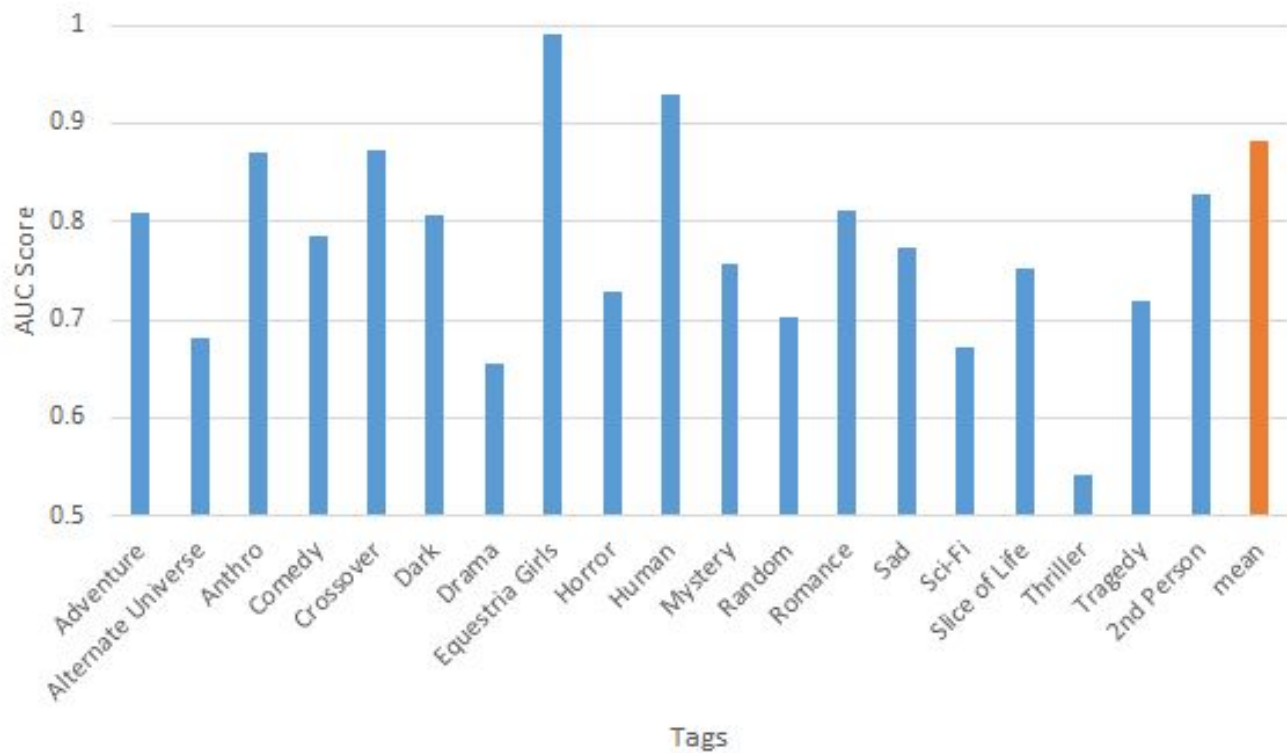
- Area Under the Receiving Operator Characteristic
- ROC plots true positive rate against false positive rate
- Results can be interpreted as how likely classifier is to give a true positive a pass over how likely classifier is
- score of 1 is perfect, .5 is no better than random



<https://www.kaggle.com/c/GiveMeSomeCredit/forums/t/14823/why-are-we-predicting-a-probability-rather-than-a-binary-response/100798>

# Results

Evaluation of Classifier using AUC metric



# Results: continued

Training Set	AUC Score
Entire Training Set	0.8815
top 1/3 most viewed of test set	0.9052
top 1/3 least viewed of test set	0.8527

# Side by side comparisons:

```
id: 21241
[('Slice_of_Life', 0.359), ('Romance', 0.167), ('Adventure', -0.094), ('Comedy',
title: Luna the Matchmaker
description:
  Princess Luna plays matchmaker for the ponies holding the Elements of Harmony.
predicted:
('Romance', 'Slice_of_Life')
actual:
['Dark', 'Romance', 'Sad', 'Slice_of_Life']
```

# Discussion

- Some tags more easily predicted than others, most likely due to nature of the genre
- overall successful classifications
- noticeable but marginal impact on performance based on testing set. inconclusive

# Conclusions

- Automatic Classification feasible method for ensuring tag accuracy. With fine tuning, could be implemented as a feature eg (recommended tags when posting a story)
- In future, experiment with hierarchical tag structures



# Work Cited

Christopher Manning. *Introduction to information retrieval*. Cambridge University Press, Cambridge, UK New York, 2008.

Andrew McCallum. Multi-label text classification with a mixture model trained by em. In *AAAI'99 workshop on text learning*, pages 1–7, 1999.

Bohdan Pavlyshenko. Classification analysis of authorship fiction texts in the space of semantic fields. *CoRR*, abs/1210.5965,

Wikipedia. Receiver operating characteristic — Wikipedia, the free encyclopedia. <http://en.wikipedia.org/w/index.php?title=Receiver%20operating%20characteristic&oldid=715266020>, 2016. [Online; accessed 09-May-2016].