FINAL PROJECT: LANGUAGE USE IN SLASH FANFICTION

ALYSSA SMITH

Abstract

Many fan scholars have remarked on the prevalence of slash fanfiction (fanfiction about same-sex relationships – relationships that are in the vast majority of cases not present in the original source text, often referred to as canon). Some theorize that this prevalence is because writing slash is a way for fans to critique traditional concepts of masculinity or subvert and resist normative narratives of gender and sexuality. There are papers that discuss the role of narrative and character/relationship development in slash fanfiction, but it is still ambiguous whether slash fanfiction does something fundamentally different from "regular" fanfiction. In this project, I aim to understand whether use of language in slash fanfiction, especially as compared to fanfiction and canon, can be construed as subversive and transformative in the same way that narrative and character/relationship development have been established to be. To do so, I trained [1] Google's open-source Word2Vec neural network on training data consisting of a billion words of news article text (to build a stable model of the English language); the text of all seven Harry Potter books (to build a baseline for vocabulary and concepts specific to or localized to the world of Harry Potter) and the text of a few hundred works of Harry Potter fanfiction [2]. I labeled vocabulary in these works of fanfiction with the genre of fanfiction (slash or non-slash) such that Word2Vec would read "Harry" in slash fanfiction as distinct from "Harry" in non-slash fanfiction or "Harry" in the source text. This meant that I had separate trained embeddings for vocabulary in each genre, and I could analyze these embeddings to understand what differences in vocabulary usage emerged [3].

Background

There have been many papers written about neural network language models; Google?s Word2Vec tool is one such method for making neural language models. Because Word2Vec uses a skip-gram model (and the gensim Python implementation I used implements skip-grams by default) dense matrix multiplications are not a necessary part of model training, and thus the model can be trained on large amounts of data fairly quickly [4]. Specifically, the skip-gram implementation is trained in the following way: suppose we have a sentence: "this class makes me sad" We want to predict a word's context given that word. So we

[1] dropbox link to .zip of neural net models (large file):https://www.dropbox.com/s/dt8irm6jw1lf6uw/fanfics.zip?dl=0—

[2] dropbox link to .zip of fanfiction directory (also pretty large file): https://www.dropbox.com/s/fxowd5h4mv89539/models.zip?dl=0

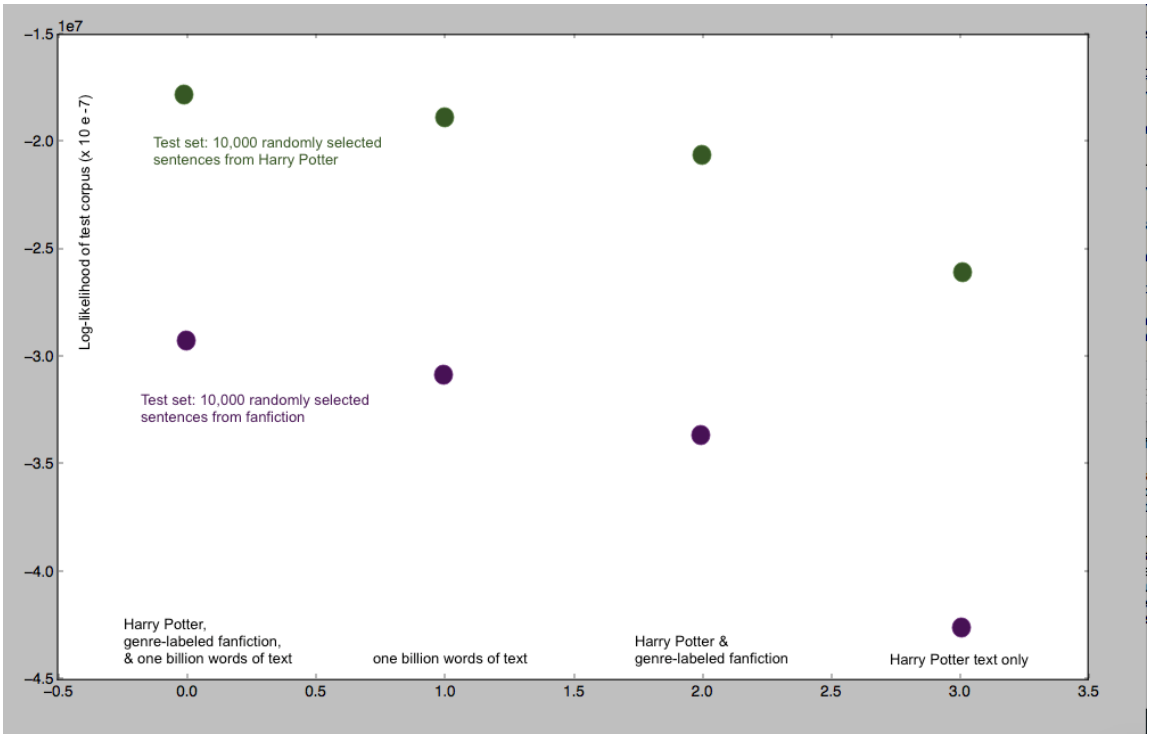
[3] Link to github repo: https://github.mit.edu/asmithh/fanfiction-embeddings-2

[4] http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

could feed the word "me" (or, more specifically, the one-hot vector corresponding to "me") into the neural network, and we would hope to get out the context as a result: "this", "class", "makes", "sad." The model's goal is to maximize this average log-probability:
$\frac{1}{T}\sum_{t=1}^{T}\sum_{-p\leq c\leq p} logp(w_{t+j}|w_t)$
where T is the length of the sequence of training words $w_1, w_2, w_3...w_T$ and c is the size of the context of the skip-grams (a parameter in the model). It uses hierarchical softmax to converge on internal weights (the embeddings we will be talking about later in this paper are the weights for the hidden layer of the neural network) which is comparatively computationally efficient and allows us to train the model on the order of a billion words (and produce a relatively stable model of the English language).

Baseline:
I trained a few models using different permutations of training corpora. All of these models used Google's open Word2Vec neural network language model, which can be found here. Specifically, I used the Python gensim implementation for this project. To build a stable model of the English language, I hypothesized that I would need to train the model on around a billion English words; luckily there are a few such corpora available for download linked to from the Google Word2Vec page. I ended up using the dataset from the One Billion Word Language Modeling Benchmark (links to a very, very large download file: link to Google Code web page) for the English language training aspect of this model. I also knew that I would incorporate the text of the Harry Potter novels into the model (for usages and vocabulary peculiar to the Harry Potter universe) and the text of Harry Potter fanfiction (for which I would later label words with type of fanfiction – slash (characters written in same-gender relationships) and heterosexual (characters in opposite-gender relationships) such that the same word in slash fanfiction, heterosexual fanfiction, and source text would be understood as 3 distinct terms by the model). Below is a graph where I compared the log-likelihood of four models on a validation corpus of 10,000 sentences.

The green points are log-likelihood of 10,000 randomly chosen sentences from Harry Potter and the purple points are log-likelihood of 10,000 randomly chosen sentences from Harry Potter fanfiction. We can see that a model trained on Harry Potter, fanfiction, and the one-billion word benchmark corpus did best overall; training on just the one-billion word benchmark was also quite good for this use case, but it likely missed some quirks of language unique to Harry Potter and its fanfiction. Training on just Harry Potter and genre-labeled fanfiction was also fairly accurate but would probably not produce a stable model (which apparently requires a great deal more training data, even though the fanfiction corpus was quite substantial) and training on just the Harry Potter text was noticeably worse than models produced by all other training corpuses. Since the leftmost model (Harry Potter, fanfiction, and one-billion-word benchmark) did best in the use case we are concerned with, I went with that model for the rest of this project.



Methods

To understand the difference in vocabulary usage between different genera of fanfiction, I needed to distinguish when training the neural net between words that came from fanfiction and words that came from the English corpus or Harry Potter (the source text) such that the model would recognize them as distinct words. To do so, I tagged words coming from slash (here described as male same-gender relationships because there were too few works written about female same-gender relationships, even in a sample of a few hundred texts) as "word*m_m"'and words coming from heterosexual fanfiction as "word*f_m." Words coming from the English corpus and from Harry Potter were left unlabeled; I generally refer to their corresponding embeddings in the model as "baseline" vectors. Since the neural net was trained on all these texts at once, and the fanfiction dataset is fairly large, making comparisons between these baselines and the embeddings for words in fanfiction can tell us what differences there are between usage in fanfiction and baseline.

Results:

Once the model was trained, I looked at a few metrics to examine the differences between embeddings for words used in fanfiction and their baseline counterparts. I ended up plotting euclidean distance versus cosine similarity; euclidean distance versus 5-dimensional polynomial kernel; and spatial distance (reduced to 2 dimensions using T-SNE).

In this plot, the red dots correspond to word embeddings trained from heterosexual fanfiction. When I first plotted these, I noticed that while the word embeddings for heterosexual

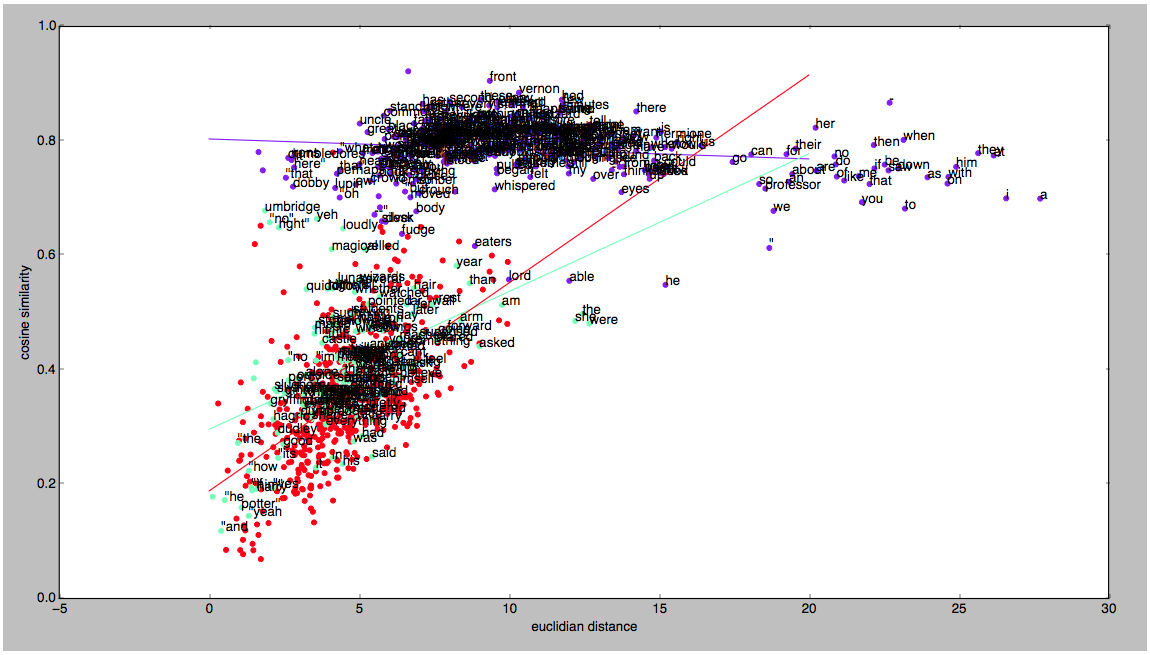


Fig. 1: Euclidean distance versus cosine distance from baseline word vector embeddings for slash fanfiction and heterosexual fanfiction: top 500 most often used words in Harry Potter.

fanfiction formed one cluster, the word embeddings for slash fanfiction formed two distinct clusters: one that overlapped with the red points quite a bit (denoted here in green), and one that had less cosine similarity to baseline and at times greater euclidean distance as well (the purple points). I defined these clusters by k-means clustering the points on a tuple consisting of (euclidean distance, cosine similarity, and cosine similarity/euclidean distance). As is visible in the image above, the clusters are well-defined and well-separated. The best-fit lines (from linear regression) indicate differences between clusters as well.

The clustering and color coding of points is the same scheme as explained for fig. 1. The regression lines are less meaningful in this graph, however, as a method of emphasizing cluster distinctiveness. The red (heterosexual) and green (some slash fanfiction) clusters are not as well separated from the purple cluster (slash fanfiction) as in fig. 1.

This figure is the result of finding vector differences between the slash fanfiction word embedding and the baseline as well as the heterosexual fanfiction word embedding and baseline for the top 500 most used words in Harry Potter. I k-means clustered these difference vectors while they were still 300-dimensional; according to this elbow plot of sum of squared error versus number of clusters, 5 clusters was the optimal number of clusters for this particular plot. Once I had labeled clusters, I used T-SNE to reduce the dimensionality of the 300-dimensional embedding vectors to 2 dimensions in order to plot. If we analyze the clusters' composition, we can see that generally a cluster is composed of one category of fanfiction or the other; there's not much mixing in clusters. The orange cluster

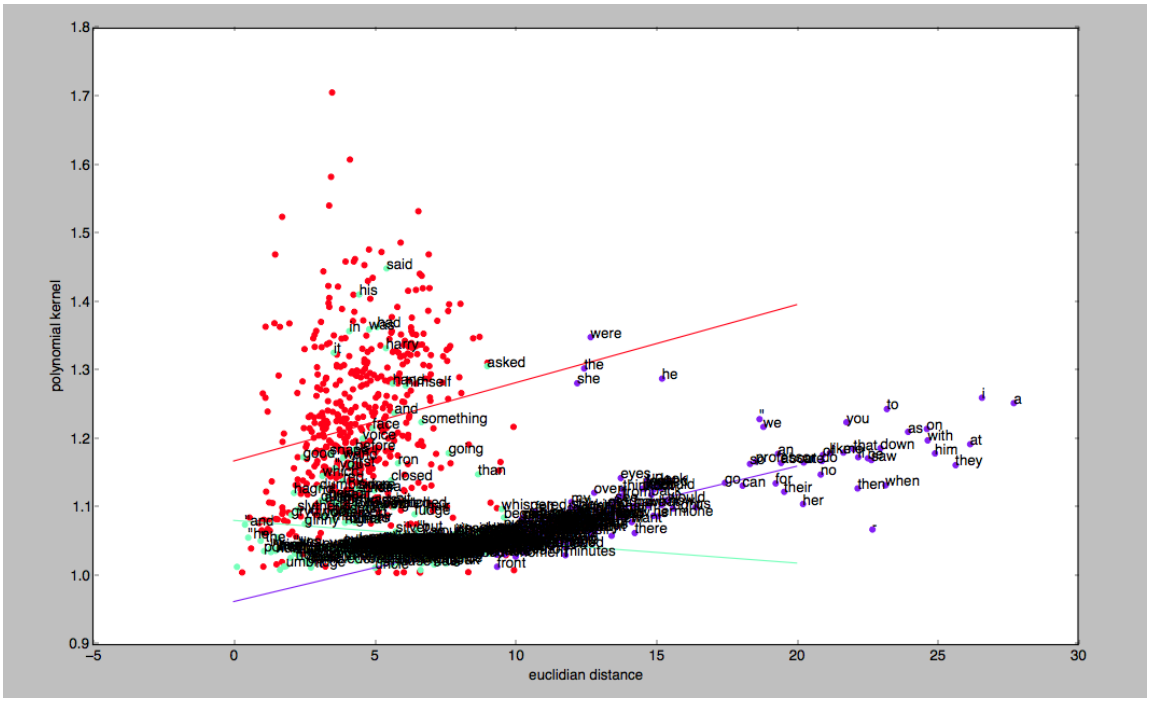


Fig. 2: Euclidean distance versus polynomial kernel from baseline word vector embeddings for slash fanfiction and heterosexual fanfiction: top 500 most often used words in Harry Potter.

(mostly triangles) is 92% slash embeddings and 8% heterosexual, while the purple cluster (mostly ?x?s) is 89% heterosexual fanfiction and 11% slash. The rest are completely pure (blue: slash, red and green: heterosexual), suggesting that differences between fanfiction embedding and slash embeddings are perhaps meaningful and/or machine-detectable.

Analysis of Results

The differences between baseline word embeddings and the fanfiction embeddings for the corresponding word are significantly dissimilar enough between slash fanfiction and heterosexual fanfiction that it would be reasonable to conclude that language usage differs in some way between the two categories. If we look at overlap between the clusters I have defined in the graphs in the previous section, we can see that there is a good deal of overlap:
Euc/Cos #1 and Euc/Poly #1 (both purple): 32.0%
Euc/Cos #1 and Euc/Poly #0 (green): 45.4%
Spatial #1 and Euc/Poly #1 (orange triangles; purple): 31.3%
Spatial #0; Euc/Poly #0 (blue triangles; green): 18.4%
Spatial #0; Euc/Cos #0 (blue triangles; green): 17.4%
Spatial #1, Euc/Cos #1: 27.0%
Note that I'm defining percent overlap as the number of points shared between both clusters, divided by the combined size of the two clusters. With that measure, overlap between

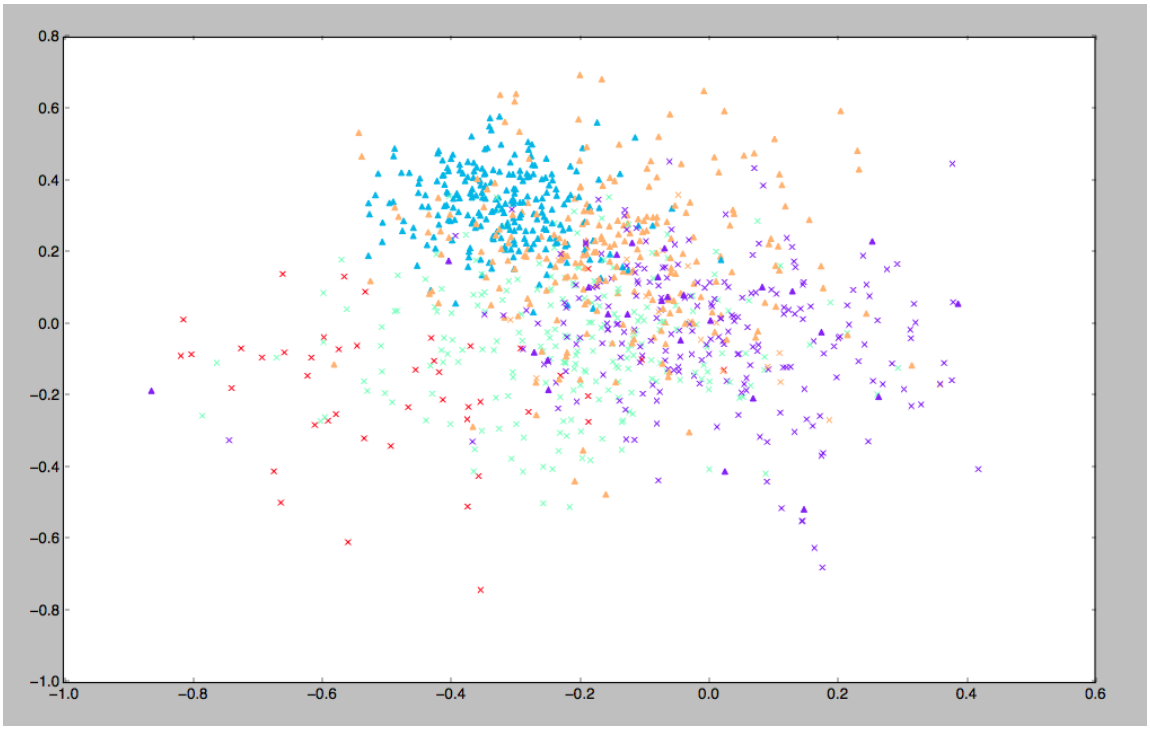


Fig. 3: Spatial difference between slash fanfiction word embeddings and baseline and heterosexual fanfiction word embeddings and baseline. Slash fanfiction embeddings denoted with a triangle; heterosexual fanfiction embeddings denoted with an ?x?. Color denotes cluster membership (5 clusters depicted in this image).

30 and 45 percent is pretty interesting. However, we need to examine the actual contents of these clusters to understand what differences might be taking place. If we sample these clusters, we can get some examples of words typical to each cluster:
sampled from intersections of cluster 1s: ”felt’, ’like’, ’out’, ’could’, ’tell’, ’right’, ’turned’, ’back’, ’by’, ’got’, ”’i’, ’go’, ’eyes’, ’not’, ’came’, ’over’, ’an’, ’him’, ’knew’, ’between’, ’can’, ’is’, ’still’, ’are’, ’look’, ’when’, ’think’, ’be’, ’me’, ’so’, ’time’, ’this’, ’see’, ’about’, ’for’, ’professor’, ’to’, ’how’,
sampled form intersections of cluster 0s: ’mcgonagall’, ’fire’, ’floor’, ’ron’, ’ginny’, ’appeared’, ’taken’, ’high’, ’corridor’, ’several’, ’rons’, ’harry”’, ”’oh’, ’dumbledore’, ’try’, ’pointing’, ’name’, ’leave’, ’yeh’, ’uncle’, ’neville’, ’luna’, ’perhaps’, ’watched’, ’whole’, ’yet’, ’stone’, ’empty’, ”’and’, ’didn’t’, ”’the’, ’cold’, ’aunt’, ’i’m’, ’don’t’, ’harry’s’, ”’it’, ’feet’, ’place’, ’passed’, ’moody’, ’upon’, ’stopped’, ’new’, ’him”’, ’madam’, ’hogwarts’, ’taking’
For each graph, when examining labeled points, I noticed that the points in the cluster #0 groupings tended to be concerned with worldbuilding and terminology peculiar to the Harry Potter universe – the vocabulary used might be used to describe the built environment, characters who form the landscape of the story, such as professors at Hogwarts, and

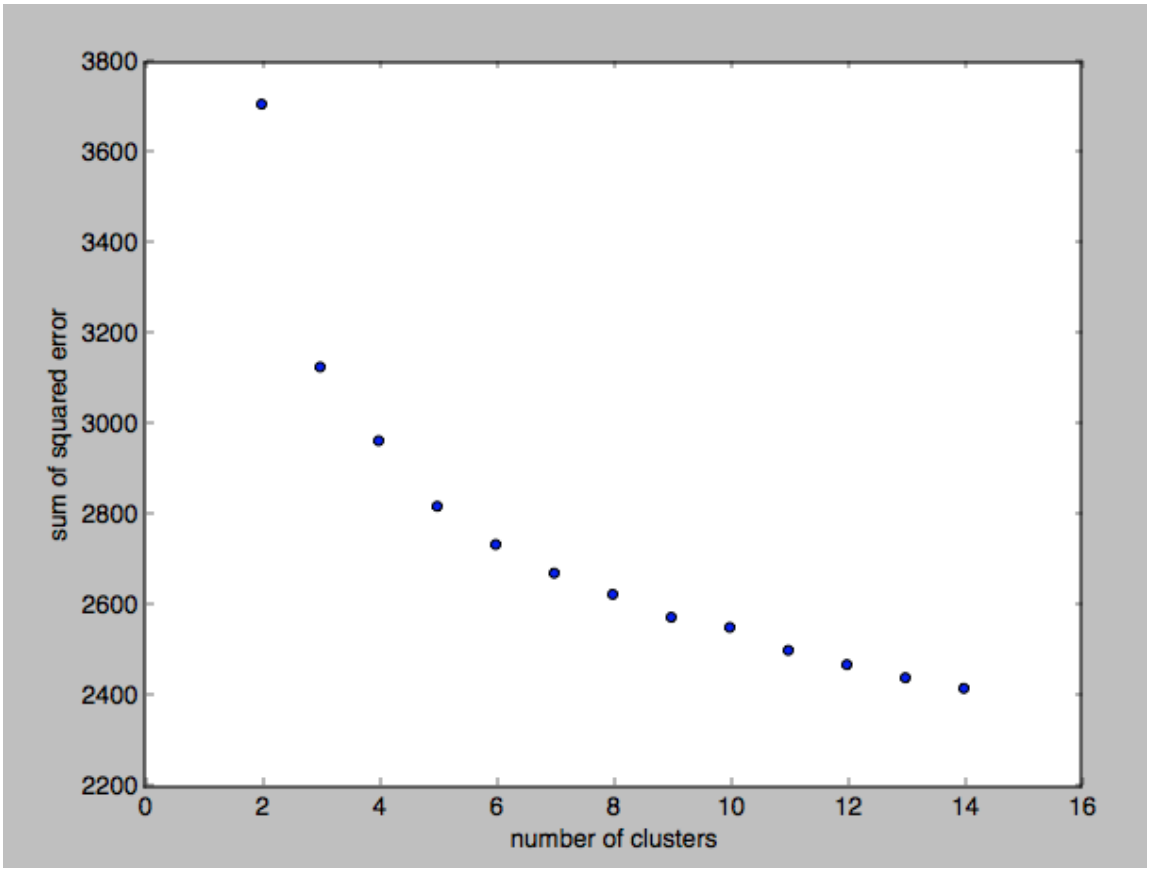


Fig. 3a: elbow plot of sum of squared error versus number of clusters for fig. 3.

words pertaining to the way magic is written in Harry Potter (?wand,? ?spell,? ?eaters? as in Death Eaters). Here are some plots zoomed in on the graph from the euclidean/cosine plots and euclidean/polynomial plots of cluster #0:
The euclidean/cosine plot (left) has more words concerned with characters grouped in cluster 0, while the euclidean/polynomial plot captures some words that might overlap slightly with the purple cluster (cluster 1) but also captures a greater range of world-building vocabulary. I would hypothesize that this is because fanfiction authors may be inclined to keep the built environment and canonical universe of the Harry Potter books constant and prefer to focus on the actions of the characters. Indeed, if we look at cluster #1, we find that it is comprised of words are either a) style markers (words that are not significant in sentence meaning or as concepts in and of themselves, but can be used statistically to match writing of unknown origin to an author with a known body of work using these words as features) or b) words that indicate character relations. Once again, here are zoomed-in images of the cluster #0 groupings for euclidean/cosine and euclidean/polynomial plots:
I hypothesize that slash fanfiction writers have perhaps a more cohesive sense of style –

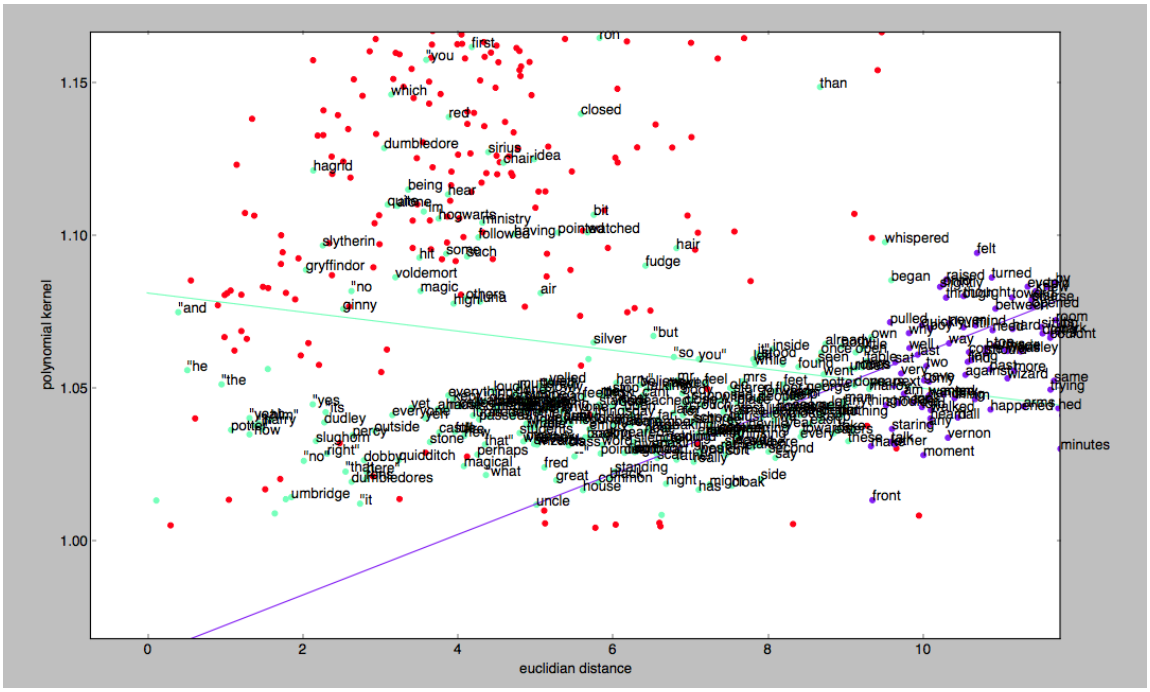


Fig. 4a: euclidean/cosine difference plot: cluster #0

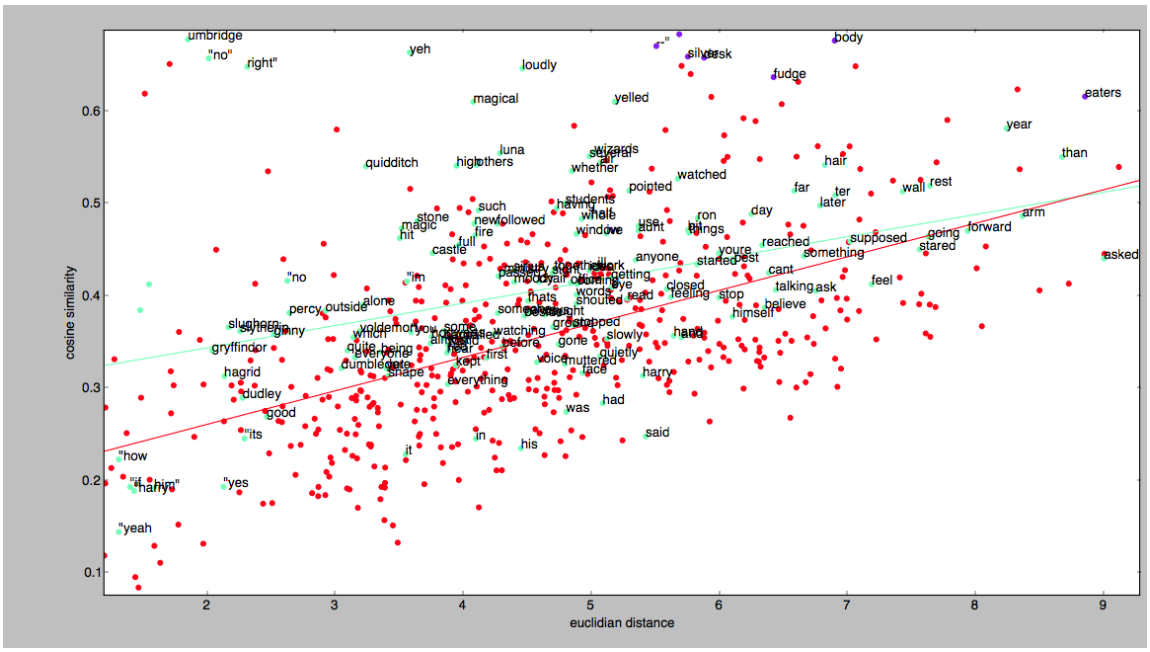


Fig. 4b euclidean/polynomial difference plot: cluster #0

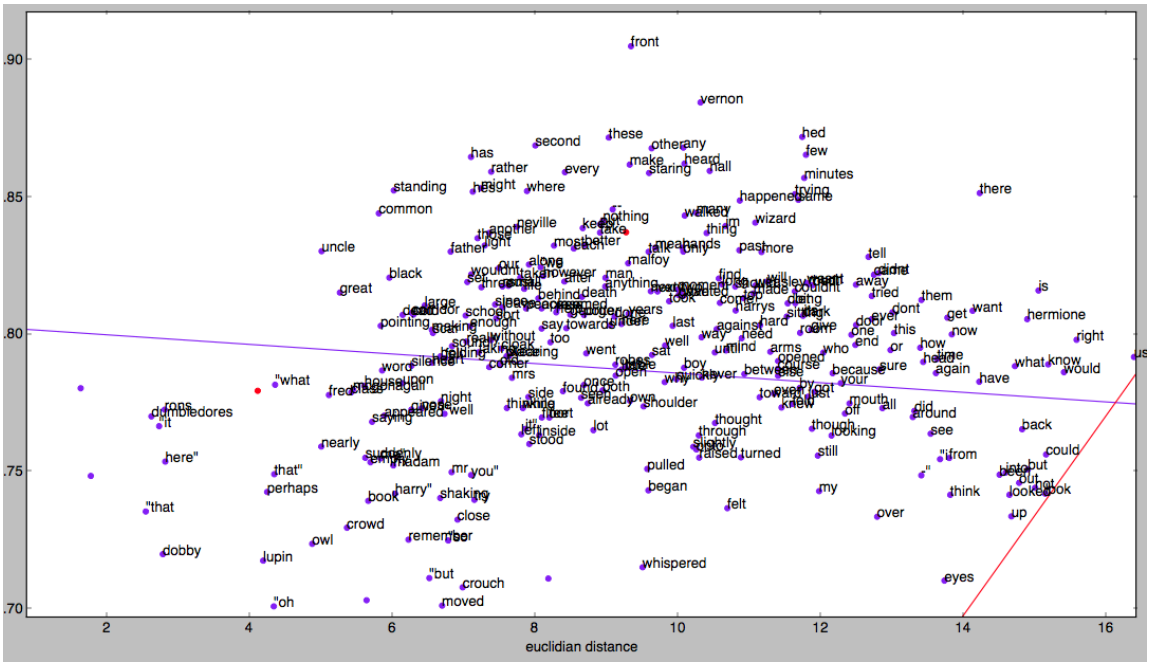


Fig. 4a: euclidean/cosine difference plot: cluster #1

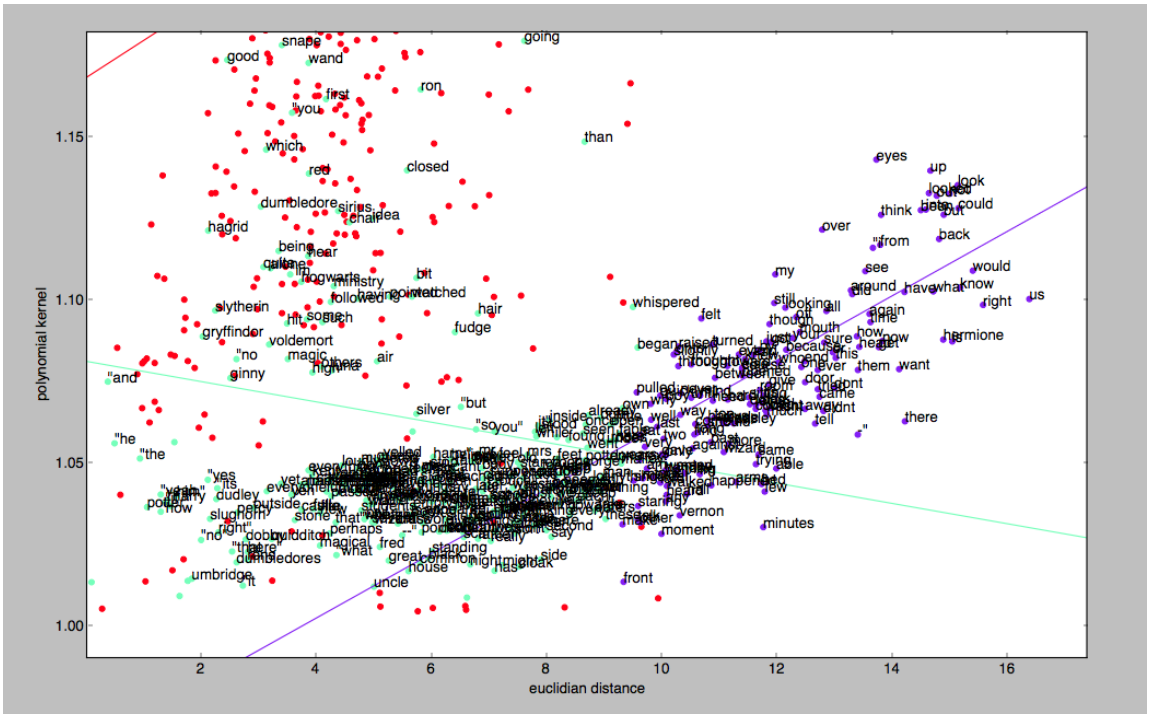


Fig. 4b euclidean/polynomial difference plot: cluster #1

perhaps the most prolific or most-read writers dominate the largest and most-read works, or influence each other's writing style in different ways such that this effect is apparent for style. The fact that words suggesting character interaction are also in this grouping (and clustered so distinctly from word embeddings for worldbuilding in slash) isn't something I'm completely sure how to explain, but I think it might be that slash fanfiction writers are in fact doing something significantly different with character relationships from heterosexual fanfiction. However, understanding this phenomenon more thoroughly would require examining context around word usage and looking at how (and where) it differs and looking at overarching aspects (plot and character development) of fanfiction – problems that are more difficult in natural language processing.