community2vec: Vector representations of online communities encode semantic relationships

Trevor Martin
Department of Biology, Stanford University
Stanford, CA 94035
trevorm@stanford.edu

Abstract

Vector embeddings of words have been shown to encode meaningful semantic relationships that enable solving of complex analogies. This vector embedding concept has been extended successfully to many different domains and in this paper we both create and visualize vector representations of an unstructured collection of online communities based on user participation. Further, we quantitatively and qualitatively show that these representations allow solving of semantically meaningful community analogies and also other more general types of relationships. These results could help improve community recommendation engines and also serve as a tool for sociological studies of community relatedness.

1 Introduction

Social media usage and participation in online communities has grown steadily over the last decade (Perrin, 2015). As we increasingly live our lives online, it is important to characterize the online communities we inhabit and understand the relationships between them. Our expanding reliance on online communities also represents an exciting opportunity to understand the links between different interests and hobbies, as candid participation across online communities is more immediately and scalably measurable compared to offline communities.

Recent work has shown that vector representations and embeddings of entities are a powerful tool across a range of applications from words (Mikolov et al., 2013a) to DNA sequences (Asgari and Mofrad, 2015). In particular, the co-occurrence based embeddings of words in a corpus has been demonstrated to encode meaningful semantic relationships between them (Mikolov et al., 2013b). In this paper we extend the concept of vector embeddings to represent an unstructured collection of online communities and show that the co-occurrence of users across online communities also embeds the semantic relations between them. Further downstream applications of these results could include improved community recommendation engines and advertisement targeting.

We focus our analysis on the social sharing site Reddit, the 4th most popular website in the US (Alexa, 2017), which has user created and managed communities called subreddits.[1] Subreddits are communities centered around particular topics and interests where users can post articles and comments while also voting content up or down to make it more or less visible. To our knowledge this paper represents the first use of vector based representations of such communities to solve analogies and perform semantically meaningful calculations of relationships.

2 Related Work

Reddit is relatively understudied compared to other social networks such as Facebook, but an increasing body of work has used its data to look at topics ranging from online user behavior (Hamilton et al., 2017) to user migration across social media platforms (Newell et al., 2016). A map of Reddit using commenter co-occurrences has also been previously created using a much smaller sample of comment data (Olsen and Neal, 2015) by treating the co-occurrence matrix as a weighted graph and extracting the network backbone. Relatedly, there has been interest in developing vector representations of graph structures as shown by techniques

[1] Subreddits are typically denoted with a leading `/r/`, for example `/r/dataisbeautiful` is the "dataisbeautiful" subreddit.

Proceedings of the Second Workshop on Natural Language Processing and Computational Social Science, pages 27–31,
Vancouver, Canada, August 3, 2017.

like DeepWalk (Perozzi et al., 2014) and node2vec (Grover and Leskovec, 2016), which we could potentially use to create additional vector representations to test below. Reddit communities do not have a built-in explicit graph structure though, as there are not defined links between communities in the same manner as users can be linked by friendship requests on sites like Facebook. In this paper we show that semantically meaningful maps of communities can be created using the NLP toolbox originally created for mapping the semantic similarity of words, without a need for defining an explicit graph.

3 Method

Our method for uncovering semantic relationships between online communities begins by creating vector representations of each community based on how often users comment across communities using one of the three methods outlined below. Broadly, we follow the general framework of Levy et al. (2015), where in our modified framework communities take on the role of words and user co-occurrence the role of word co-occurrence. We then simply add and subtract these community vectors to evaluate semantic correctness. Here, we use a publicly available corpus of all Reddit comments from January 1st, 2015 through April 30th, 2017 as the input to each technique. This data set consists of roughly 1.8 billion comments across 60,978 subreddit communities.[2]

3.1 Subreddit Vectors

We first create a symmetric matrix of community-community user co-occurrences $\mathbf{X}$, whose entries $\mathbf{X_{ij}}$ indicate the number of unique users who commented 10 times or more in each subreddit.

Explicit: Our explicit subreddit representation first simply subsets the co-occurrence matrix $\mathbf{X}$ to include only the subreddits with unique author ranks between 200 and 2,201 as context subreddits (columns of $\mathbf{X}$). The choice of rank cutoff here is arbitrary but based on the idea that performance can be increased by adjusting the number of context tokens (Bullinaria and Levy, 2007). We choose the subreddits with the most unique authors because these are likely to encode the most useful information and drop the top 200 subreddits because many of these are "default" subreddits that all Reddit users are subscribed to and thus are unlikely to have as rich co-occurrence information. Then we transform this new matrix $\mathbf{X}_{:,\mathbf{201:2200}}$ using the positive pointwise mutual information metric to weigh each count by its informativeness, where $p(i,j)$ is the joint probability of seeing authors in both subreddits i and j and $p(i)$ and $p(j)$ are the probabilities of seeing an author in each subreddit respectively:

$$PMI(i,j) \equiv \log \frac{p(i,j)}{p(i)p(j)}$$

$$PPMI(i,j) = \begin{cases} 0, & \text{if } PMI(i,j) < 0 \\ PMI(i,j), & \text{otherwise} \end{cases}$$

The subreddit vectors (rows) of the resulting $PPMI$ matrix are then scaled to unit length.

PCA: We also create a dense vector representation of subreddits by calculating the principal components of the $PPMI$ transformation above applied to the matrix $\mathbf{X}_{:,\mathbf{1:5000}}$, which is $\mathbf{X}$ subset to the top 5,000 context subreddits by unique author ranks. We extract the top 100 principal components and scale each subreddit vector to unit length.

GloVe: Finally, we create a second dense vector representation of subreddits by running the GloVe algorithm (Pennington et al., 2014), originally developed to create embeddings for word-word co-occurrence matrices, on the raw co-occurrence matrix $\mathbf{X}$. The resulting size 100 GloVe subreddit vectors are again scaled to unit length.

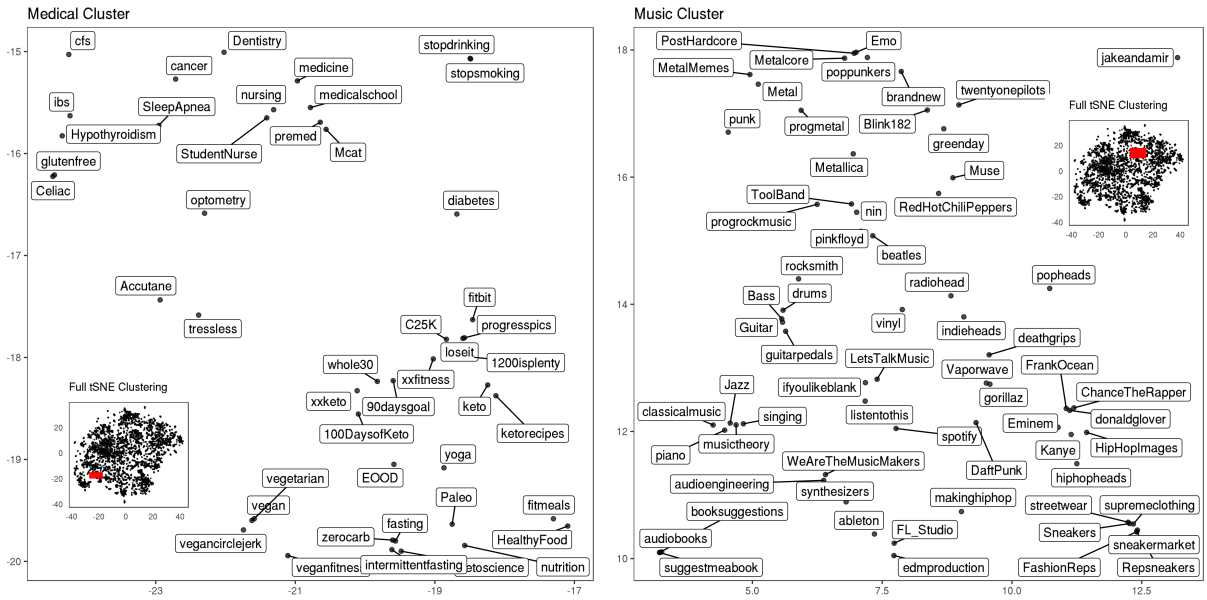
3.2 Subreddit Algebra

Combinations of subreddit representations (subreddit algebra) are performed through standard vector addition and subtraction. The similarity between two subreddits is defined here as the *cosine similarity*, given by:

$$\text{cosine similarity}(\vec{A},\vec{B}) = \frac{\vec{A}\cdot\vec{B}}{\|\vec{A}\|\|\vec{B}\|}$$

Where $\vec{A}$ and $\vec{B}$ are the vector representations of subreddit A and B respectively. Subreddits are ranked in similarity by ordering from largest cosine similarity to smallest.

[2] Reddit data available at: https://bigquery.cloud.google.com/table/fh-bigquery:reddit_comments.all_starting_201501



(a) View of subreddits representing medical interests and health conscious lifestyles.

(b) View of subreddits representing music genres and performing groups.

Figure 1: Examples of semantically meaningful clusters in t-SNE visualization of GloVe subreddit vectors. Zoomed-in region of t-SNE visualization indicated in red on figure insets.

4 Evaluation

We quantitatively evaluate the efficacy of subreddit algebra by assessing its ability to identify local sports team subreddits from combinations of league and geography subreddits. Additionally, we qualitatively evaluate our the results by identifying specific interesting subreddit relationships and visualizing the subreddit vector space as a whole.

4.1 tSNE Clustering

To check that our vector representations of subreddit communities are reasonable, we used t-SNE (Maaten and Hinton, 2008) to project the high-dimensional vector representations of each subreddit into two dimensions for visualization. Examples of typical semantically meaningful clusters that we can observe in these t-SNE projections are given in Figure 1. Figure 1a shows that medical and health related subreddits cluster together and Figure 1b shows the dense clustering of music and band related subreddits and clustering within this larger group by music genre.[3] These natural groupings suggest that our vector representations are reasonable and are encoding semantically relevant information about each subreddit.

[3]To aid in visualization, we only project the top 5,000 and 2,500 subreddits by unique author count for the medical and music GloVe based clusters respectively.

4.2 Automated Semantic Relationship Test

In order to quantitatively evaluate the ability of the subreddit vectors to encode semantic relations, we created a list of subreddit combinations where we have a strong expectation for the outcome subreddit. Conveniently, sport, location, and team subreddits have a natural analogy structure. Specifically, for the NBA, NFL, and NHL sports leagues we created a list of geographic location subreddits (e.g. `/r/sanfrancisco`) that when combined with a league subreddit (e.g. `/r/nba`) should result in that location's local league affiliate (e.g. `/r/warriors`).[4] Performance on this task for an individual league-location pair is assessed by calculating:

$$\text{median}(SR(\vec{S},\vec{T}), SR(\vec{L},\vec{T})) - SR(\vec{S}+\vec{L},\vec{T})$$

Where $\vec{S}$ is the league subreddit, $\vec{L}$ is the location subreddit, and $\vec{T}$ is the target subreddit. $SR(\vec{A},\vec{B})$ is the rank of the subreddit B when all subreddits are ordered by decreasing cosine similarity to subreddit A.

The decrease in similarity ranking for each sports league across each of the three vector representations was then evaluated for significance by

[4]In total we use 92 league-location combinations.

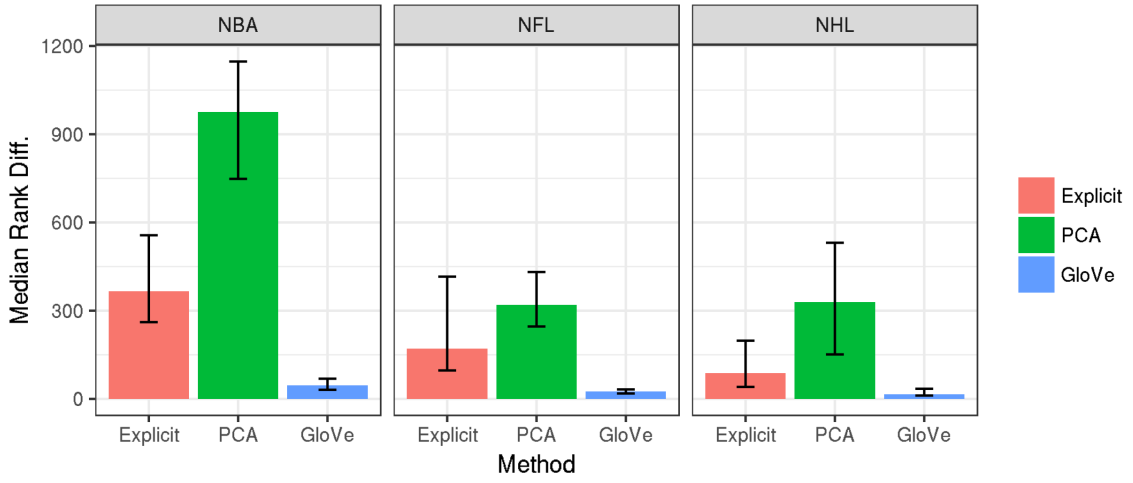


Figure 2: Comparison of different vector representation's performance for identifying local sports teams in each league.

Method	League	$\vec{S}+\vec{L}$: $\vec{T}$ Median Rank	Median Rank Diff.	p-value
Explicit	NBA	7	365.5	1.9e-9
	NFL	5	170.8	8.3e-7
	NHL	4	87.5	1.9e-9
PCA	NBA	212	976.3	4.7e-8
	NFL	13	320.1	9.3e-10
	NHL	41.5	330	3.4e-4
GloVe	NBA	7	46.5	1.8e-6
	NFL	2	25	1.5e-5
	NHL	1	16.5	1.8e-6

Table 1: Results of automated testing of subreddit vector representation semantic encodings.

a two-sided Wilcoxon signed-rank test for symmetry of the rank changes around 0. The median decrease in target subreddit rank between $SR(\vec{S}+\vec{L},\vec{T})$ and median$(SR(\vec{S},\vec{T}),SR(\vec{L},\vec{T}))$ for each sports league-vector representation pair is shown in Figure 2.[5] Interestingly, both the explicit and PCA vector representations appear to perform best, but all three methods show significant performance on the task as indicated in Table 1.

Closer inspection of the results reveals though that while the PCA method has the largest improvement in target subreddit rank (Median Rank Diff. in Table 1), it also has the highest median subreddit ranks for the target subreddits after performing subreddit algebra of the three methods ($\vec{S}+\vec{L}$: $\vec{T}$ Median Rank in Table 1). This observation suggests that while the PCA representations benefit the most from algebra they also have the least accuracy for identifying the target subreddit overall.[6] In contrast, for algebra using either the explicit or GloVe vector representations, the target subreddit is often the most similar result.

4.3 Selected Semantic Examples

In addition to the automated test, we also identified several interesting analogy tasks to run using subreddit algebra.[7] Because we do not necessarily have subreddits for representing concepts such as "man" or "woman" we cannot reproduce exactly classic cases like $king-man+woman=queen$, but for the cases where we could form robust analogies the results are encouraging, as shown in Figure 3.

Of note is that we can reproduce country:capital relationships similar to those found in word embeddings using community participation across subreddits and also can reproduce analogies that

[5] More specifically the Hodges-Lehmann pseudomedian, with 95% CI

[6] Also, PCA based representations do not necessarily have the linear substructure seen in GloVe embeddings.

[7] We use the explicit representations here.

```
/r/berlin - /r/germany + /r/unitedkingdom = /r/london
/r/chicagobulls - /r/chicago + /r/minnesota = /r/timberwolves
/r/running + /r/weightlifting = /r/fitness
/r/personalfinance - /r/frugal = /r/wallstreetbets
/r/books + /r/fiction = /r/HFnovels
```

Figure 3: Selected semantic algebra examples.

subtract a component (Chicago) of a whole (Chicago Bulls NBA team) and add a different location (Minnesota) to get that locality's NBA team (Minnesota Timberwolves). We can also find communities specific to medium-genre combinations such as the historical fiction book community `/r/HFnovels`. Finally, we see some surprising examples, such as subtracting the community for frugality from the community for managing personal finances results in the community for taking extreme risks on the stock market, `/r/wallstreetbets`.

5 Conclusions

Our work here shows that vector representations of communities can encode meaningful analogies and semantic relationships in the same way as has been previously seen for words. Notably, the explicit vector representations perform competitively with the GloVe embeddings on the semantic task we tested, suggesting that the semantic meanings are present in the raw vectors and are simply preserved through the embedding process. Future directions we are pursuing involve supplementing the vector representations with data on comment voting scores, using posts or views in lieu of or supplementally to comments and looking at diachronic subreddit embeddings to analyze the patterns of subreddit relationships over time.

Acknowledgments

We would like to thank Will Hamilton for his valuable comments and suggestions on the manuscript.

References

Alexa. 2017. http://www.alexa.com/siteinfo/reddit.com. *Alexa Rankings* .

Ehsaneddin Asgari and Mohammad R. K. Mofrad. 2015. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLOS ONE* 10(11):1–15. https://doi.org/10.1371/journal.pone.0141287.

John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods* 39(3):510–526. https://doi.org/10.3758/BF03193020.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. *CoRR* abs/1607.00653. http://arxiv.org/abs/1607.00653.

William L. Hamilton, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Loyalty in online communities. *CoRR* abs/1703.03386. http://arxiv.org/abs/1703.03386.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* 3:211–225. https://transacl.org/ojs/index.php/tacl/article/view/570.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* 9(Nov):2579–2605.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781. http://arxiv.org/abs/1301.3781.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Hlt-naacl*. volume 13, pages 746–751.

Edward Newell, David Jurgens, Haji Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. 2016. User migration in online social networks: A case study on reddit during a period of community unrest.

Randal Olsen and Zachary Neal. 2015. Navigating the massive world of reddit: using backbone networks to map user interests in social media. *PeerJ Computer Science* .

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. pages 1532–1543. http://www.aclweb.org/anthology/D14-1162.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. *CoRR* abs/1403.6652. http://arxiv.org/abs/1403.6652.

Andrew Perrin. 2015. Social media usage: 2005-2015. *PewResearchCenter* .

A Supplemental Material

All code and league-location-team combinations are available at `https://github.com/trevormartin/papers`.