By Their Text Shall They Be Known

Individual & Community Identification From Written Corpora

LIM YONG SAN, GILBERT
National University of Singapore
April 22, 2019

His identity might be unknown, but we are at least more certain that he is not "they" – recent detailed quantitative inspection of *Beowulf* has supported the assumption that the Old English poem had been authored by a single poet rather than a partnership, despite longstanding suspicions that it had been amalgamated from two separate tales: his Danish travels, and his Draconic travails (Neidorf et al., 2019). This is merely the latest development in a long and storied tradition of bitter authorship disputes, most prominently perhaps involving the Book (Porter, 1995) and the Bard (Craig and Kinney, 2009) – and both together (Dickson, 2004). Author identification has also been sought in somewhat pettier cases, as with forged legal documents (Kelly and Lindblom, 2006).

Similar questions about attribution might naturally also apply to *communities* rather than individuals, given that those belonging to the same school of thought might well adopt compatible styles of communication. This has been extensively attested to with speech, for example as Eckert (1989) and Bucholtz (1999) studied in their works on jocks & burnouts, and nerd girls respectively. An individual's identification of belonging to each of these communities could manifest in many ways: behavioural, pragmatical, syntactical, lexical, phonological. That last is perhaps the most hackneyed way by which one's geographic community can be discerned – though seldom with the skill of a Higgins, who could place not merely one's birth, but one's career trajectory in the case of Colonel Pickering in *Pygmalion*.

Moving from spoken to written language, handwriting analysis has been a time-honoured means of establishing a writer's identity, evidenced by the continued reliance on signatures as the ultimate expression of one's deliberate will. Much research has been done in this regard, with handwriting identification techniques broadly dividable into *text-dependent* ones that directly compare known pairs of characters or words, and *text-independent* ones that employ global statistical features (Bulacu and Schomaker, 2007). It cannot be denied, however, that contemporary written corpora are generally already digitized, in their native form. In such circumstances, author identification cannot be induced from penmanship, but rather textual analysis – as with *Beowulf*.

Textual Authorship Identification Methods

Neidorf et al. (2019) took the approach of considering the entire known Old English corpus, in analyzing fine-grained features involving sound, metre and diction; through this expedient, they concluded that various partitions of *Beowulf* demonstrated consistent style, and further attributed another text *Andreas* to Cynewulf. Both conclusions rely on prior research suggesting that works by the same author exhibit similar phonetic profiles, following the approach of quantitative criticism (Dexter et al., 2017).

The above focus on phonology might not be the most appropriate with present-day corpora, as De Vel et al. (2001) point out in their work on e-mail forensics. In particular, it is stated that stylometric features such as vocabulary richness are possibly more amenable to conscious control, as opposed to syntatic features, including the placement of punctuation. Nonetheless, the popularity of stylometric features has seen approximately a thousand such style markers isolated (Rudman, 1997). Holmes (1994) offers an overview of about a dozen of the most common categories, including word and sentence length, syllable and part-of-speech distribution, and the usage of function words. Notably, Rudman acknowledged that it was the *combination* of features, and not any single feature *per se*, that is important, and Holmes foretold the extensive use of connectionist neural networks in the future.

Holmes proved quite prophetic, as neural methods would arguably come to dominate natural language processing in the 2010s. However, running counter to the trend towards deeper and deeper networks in computer vision, the shallow `word2vec` architecture (Mikolov et al., 2013) has proven itself to be exceedingly successful. In particular, `word2vec` demonstrated that it was possible to efficiently and accurately represent words as continuous vectors in some embedded representation space, given a huge training dataset. While this allowed both the prediction of a missing word given its surrounding word tokens (CBOW), and the prediction of surrounding words given a word token (Skip-gram), it was also discovered that *semantic relationships* between words could be extracted using simple algebra on word vector representations.

The natural extension to `word2vec`, then, was `doc2vec` (Le and Mikolov, 2014), which introduced paragraph vectors that are able represent texts of arbitrary length. Clearly, the implication is that such textual embeddings possibly encode information related to authorship. Interestingly, a `community2vec` model has also been proposed with regards to subreddits (Martin, 2017). However, for `community2vec`, subreddit vectors were defined based on user participation co-occurrences, and not their actual comments. Still, enlightening relationships between various subreddit populations could be deduced[*].

[*] https://fivethirtyeight.com/features/dissecting-trumps-most-rabid-online-following/

Empirical Evaluation

For our purposes, we investigate here the utility of `doc2vec` for authorship attribution, on Reddit data from December 2005 to December 2008 inclusive. For both individual Redditors and subreddit communities, we first collect all available comments in chronological order, and consider 1000 comments randomly drawn from the earlier 90% to be available for training a model, and 100 comments randomly drawn from the remaining 10% for evaluating the performance of the trained model in predicting individual or community authorship. We consider both *lexical* models – where the lexical items within comments are considered (excluding hapax legomenon), and *part-of-speech* (POS) models, where only the parts-of-speech of the comments are considered. For example, the sentence:

"Tags are good for finding pages, but URLs are good for naming them."

has corresponding POS tags:

```
NNS VBP JJ IN VBG NNS , CC NNP VBP JJ IN VBG PRP .
```

POS tagging with the Penn Treebank Tag Set (Marcus et al., 1993) for all text was performed with the Python Natural Language Toolkit (NLTK) `pos_tag` function.

Model training was performed with `doc2vec` (Le and Mikolov, 2014), which models each individual/community using their vocabulary, as exhibited in their comment corpus. By default, the distributed bag-of-words (PV-DBOW) model is used, which considers only the frequency of items (lexical or POW), and disregards their ordering. A shallow neural network is trained via backpropagation with stochastic gradient descent, for 40 epoches in all cases. The convergence of the training procedure maps similar corpora close together, in a representative vector space, as with words in `word2vec`.

With a trained model, the authorship of each of the unseen test corpora may then be predicted, by projecting their representations into the representative vector space, and measuring their closeness to the original vector projections from the training set. The prediction for an individual/community is accurate, if its (independent) test vector is indeed closest to its training vector (Top-1 metric). The Top-5 and Top-10 metrics are also considered, where we determine whether the test vector is within the closest five or ten vectors respectively, to the training vector. We experiment with representative vector dimensionalities of 40, 100 and 200 elements, to investigate how vector dimensionality affects predictive performance.

Individuals

For individuals, the 1000 Redditors with the most comments were examined, with all Redditors having made at least 1224 comments. The predictive performance is summarized in Table 1:

		Vector Dimensionality		
		40	**100**	**200**
Lexical	**Top-1**	0.5960	0.7340	0.7980
	Top-5	0.8180	0.8820	0.9240
	Top-10	0.8650	0.9330	0.9560
POS	**Top-1**	0.2210	0.2330	0.2270
	Top-5	0.4020	0.4270	0.4210
	Top-10	0.5030	0.5290	0.5310

Table 1: Predictive Performance for Individual Models

As can be seen, predictive performance improves in general with increased vector dimensionality for lexical models, while the improvement for POS models is minimal. It should be noted that Top-1 performance is as high as approximately 80% for a vector dimensionality of 200, indicating that author prediction from lexis alone can be at least 80% accurate, from a population of a thousand individuals. Further, author prediction from POS alone – the relative distribution of about 36 POS tags – is about 23%.

The t-distributed Stochastic Neighbour Embedding (t-SNE) algorithm (Maaten and Hinton, 2008) was used to visualize the lexical representational space in two dimensions (Appendix A, Figure 1). It can be noted that there is a fair spread of individuals in the representational space, suggesting that there are clear and quantifiable distinctions in the lexical vocabulary of different individuals. However, due to the anonymous nature of Reddit individual usernames, it is difficult to directly draw correlations with demographic features.

Communities

For communities, the 100 subreddits with the most comments were examined, with all subreddits having at least 1343 comments. It should be noted that users only gained the ability to created their own subreddits from January 2008 (Tan and Lee, 2015), and as such many of these top 100 subreddits are “default” subreddits covering broad topics of interest, as defined by site administrators. The predictive performance is summarized in Table 2:

		Vector Dimensionality		
		40	**100**	**200**
Lexical	**Top-1**	0.7000	0.7600	0.7200
	Top-5	0.9100	0.9300	0.9100
	Top-10	0.9500	0.9800	0.9600
POS	**Top-1**	0.1900	0.1800	0.1700
	Top-5	0.5200	0.5500	0.5700
	Top-10	0.6600	0.6800	0.6900

Table 2: Predictive Performance for Community Models

As with individuals, the accuracy of community prediction is high, reaching 76% with lexical models, and 19% with POS models. It can be noted that a high representation vector dimensionality may be detrimental with communities, perhaps due to the nature of community versus individual corpora; since community corpora comprise multiple individuals, there may be higher internal variation as a result. As such, lower vector dimensionality capturing more general trends may be more appropriate, especially when a relatively low number of comments are being sampled.

Compared to individuals, the known topics of subreddits makes t-SNE visualization more informative (Appendix B). From Figure 2 of the 40-dimensional model, it can be observed that related topics indeed share close lexical representations – for example, the cluster of the "programming", "Python" and "ruby" subreddits at the bottom left. The similarity between certain subreddits changes with dimensionality, as can be seen by way of comparison with the 100-dimensional model shown in Figure 3; the "programming", "Python" and "ruby" subreddits have moved to the top left of the representational space, and are joined by the "compsci" subreddit. A number of other stable pairings, such as "religion" & "atheism", and "Economics" & "business", may be observed in both visualizations. Of course, it should be remembered that projection from a higher-dimensional space to a two-dimensional one inevitably loses some representational fidelity. Therefore, any relational algebra à la `community2vec` should be performed in the original representation vector space.

Moving on to t-SNE visualization with POS only (Appendix C), Figure 4 exhibits a clearly distinct cluster at the bottom right, that was not manifested with lexical models, containing the "de", "es", "it", "ja", "ru" and "tr" subreddits. It can be quickly noted that these are all non-English subreddits (German, Spanish, Italian, Japanese, Russian and Turkish respectively). Interestingly, despite being limited to POS tags, certain subreddits remain objectively similar in linguistic variation, while others move apart. For example, "Israel" was consistently close to subreddits such as "worldnews", "worldpolitics" and "history" in the lexical model representations, but is isolated by itself at the top of the POS model visualization. Simply put, this may reflect a similarity in content, but a difference in tone/style.

Finally, it can be noted that “reddit.com” is near the centre of all community visualizations. This is to be expected, since “reddit.com” appears to be a miscellaneous catch-all legacy subreddit, that moreover has by far the most comments of all subreddits in the examined time period: 4.1 million comments, to 1.4 million comments for “politics” in second place.

Conclusion

In this squib, we have briefly discussed the role of linguistic variation in authorship attribution, and explored the use of the `doc2vec` embedding method for individual and community attribution with Reddit comments. A fairly high accuracy of 80% was achieved with individual attribution using only aggregated lexical data, and without particular optimization. This has practical relevance on the expectation of anonymity in online discourse, in an environment of increasing Internet censorship by the requirement of real-name registration for commentary (Fu et al., 2013), among other means (Warf, 2011). These findings suggest that practices such as using “throwaway accounts” on Reddit to preserve anonymity may not be as foolproof as individuals hope, and that motivated investigators may well be able to link comments from various online identities as likely belonging to the same individual through computational variationist analysis, even across platforms. As such, it may be advisable for individuals concerned with privacy to consider masking their style, perhaps through generative modelling of text (Shen et al., 2017; Fu et al., 2018).

Works Cited

Bucholtz, Mary. ““Why be normal?”: Language and identity practices in a community of nerd girls”. *Language in Society*, vol. 28, no. 2, 1999, pp. 203–223.

Bulacu, Marius and Lambert Schomaker. “Text-independent writer identification and verification using textural and allographic features”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, 2007, pp. 701–717.

Craig, Hugh and Arthur F Kinney. *Shakespeare, computers, and the mystery of authorship*. Cambridge University Press, 2009.

De Vel, Olivier, Alison Anderson, Malcolm Corney, and George Mohay. “Mining e-mail content for author identification forensics”. *ACM Sigmod Record*, vol. 30, no. 4, 2001, pp. 55–64.

Dexter, Joseph P, Theodore Katz, Nilesh Tripuraneni, Tathagata Dasgupta, Ajay Kannan, James A Brofos, Jorge A Bonilla Lopez, Lea A Schroeder, Adriana Casarez, Maxim Rabinovich, et al. “Quantitative criticism of literary relationships”. *Proceedings of the National Academy of Sciences*, vol. 114, no. 16, 2017, E3195–E3204.

Dickson, Peter W. "Connecting the Dots: The Catholic Question and the Shakespeare Authorship Dispute". *Tennessee Law Review*, vol. 72, 2004, p. 25.

Eckert, Penelope. *Jocks and burnouts: Social categories and identity in the high school*. Teachers college press, 1989.

Fu, King-wa, Chung-hong Chan, and Michael Chau. "Assessing censorship on microblogs in China: Discriminatory keyword analysis and the real-name registration policy". *IEEE Internet Computing*, vol. 17, no. 3, 2013, pp. 42–50.

Fu, Zhenxin, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. "Style transfer in text: Exploration and evaluation". *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 663–670.

Holmes, David I. "Authorship attribution". *Computers and the Humanities*, vol. 28, no. 2, 1994, pp. 87–106.

Kelly, Jan Seaman and Brian S Lindblom. *Scientific examination of questioned documents*. CRC press, 2006.

Le, Quoc and Tomas Mikolov. "Distributed representations of sentences and documents". *International Conference on Machine Learning*, 2014, pp. 1188–1196.

Maaten, Laurens van der and Geoffrey Hinton. "Visualizing data using t-SNE". *Journal of Machine Learning Research*, vol. 9, no. Nov, 2008, pp. 2579–2605.

Marcus, Mitchell, Beatrice Santorini, and Mary Ann Marcinkiewicz. "Building a large annotated corpus of English: The Penn Treebank". 1993.

Martin, Trevor. "community2vec: Vector representations of online communities encode semantic relationships". *Proceedings of the Second Workshop on NLP and Computational Social Science*, 2017, pp. 27–31.

Mikolov, Tomas, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space". *Proceedings of Workshop at ICLR*, vol. abs/1301.3781, 2013.

Neidorf, Leonard, Madison S. Krieger, Michelle Yakubek, Pramit Chaudhuri, and Joseph P. Dexter. "Large-scale quantitative profiling of the Old English verse tradition". *Nature Human Behaviour*, 2019.

Porter, Stanley E. "Pauline authorship and the Pastoral Epistles: implications for canon". *Bulletin for Biblical Research*, vol. 5, 1995, pp. 105–123.

Rudman, Joseph. "The state of authorship attribution studies: Some problems and solutions". *Computers and the Humanities*, vol. 31, no. 4, 1997, pp. 351–365.

Shen, Tianxiao, Tao Lei, Regina Barzilay, and Tommi Jaakkola. "Style transfer from non-parallel text by cross-alignment". *Advances in Neural Information Processing Systems*, 2017, pp. 6830–6841.

Tan, Chenhao and Lillian Lee. "All who wander: On the prevalence and characteristics of multi-community engagement". *Proceedings of the 24th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, 2015, pp. 1056–1066.

Warf, Barney. "Geographies of global Internet censorship". *GeoJournal*, vol. 76, no. 1, 2011, pp. 1–23.

Appendix

A: t-SNE visualization of Individuals (Lexical)

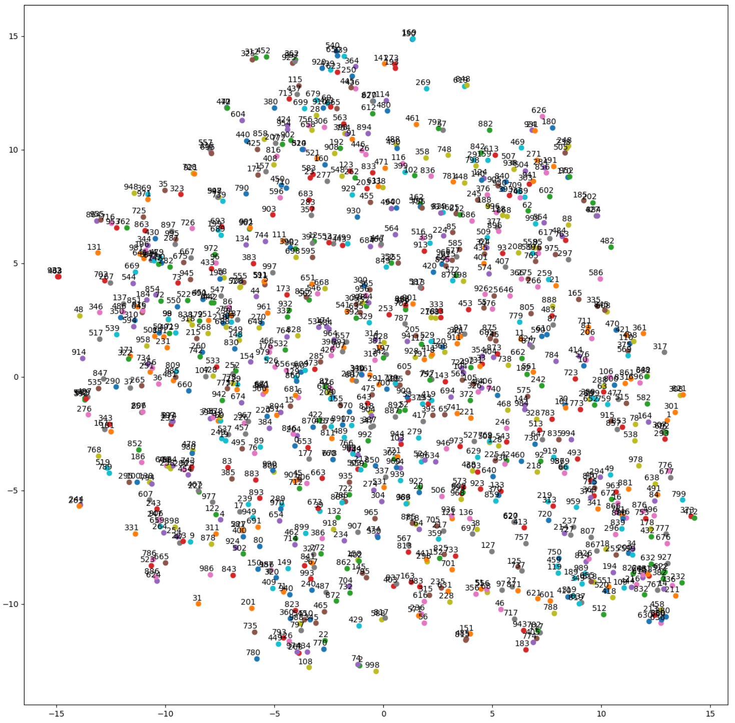

Figure 1: `doc_author_vs40_ep40` model

B: t-SNE visualization of Communities (Lexical)

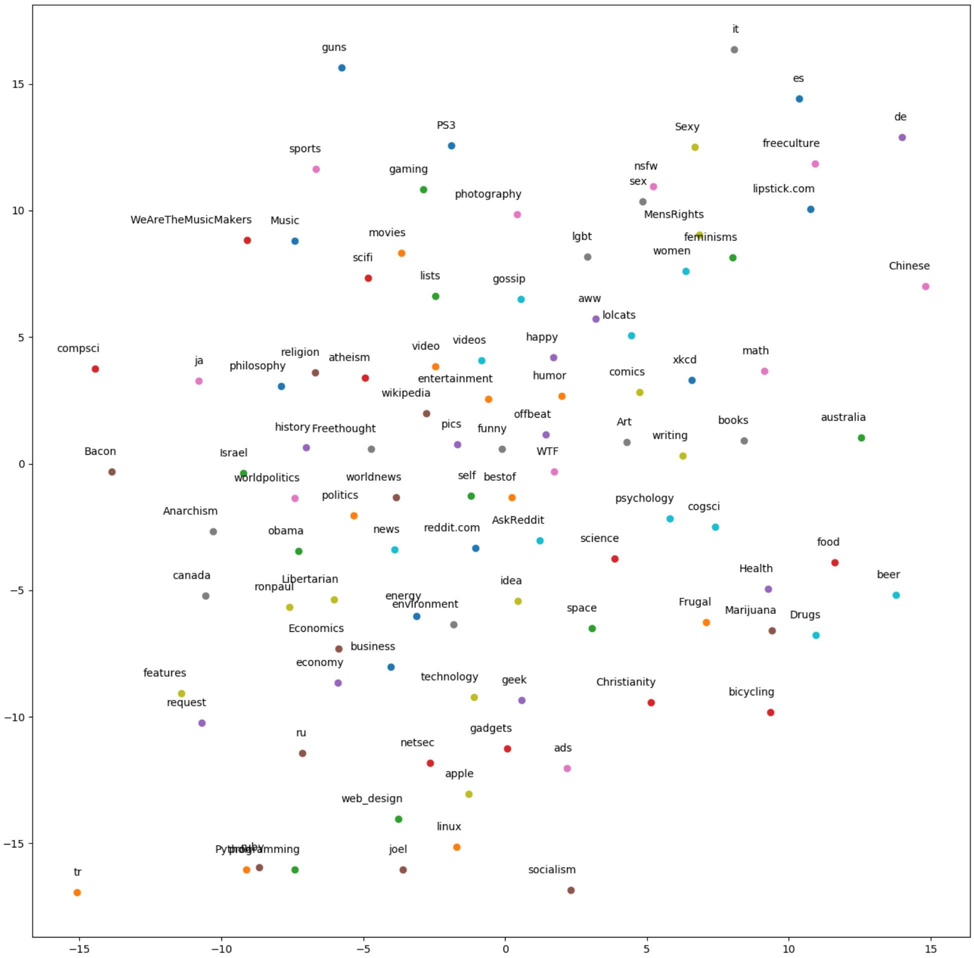

Figure 2: `doc_subreddit_vs40_ep40` model

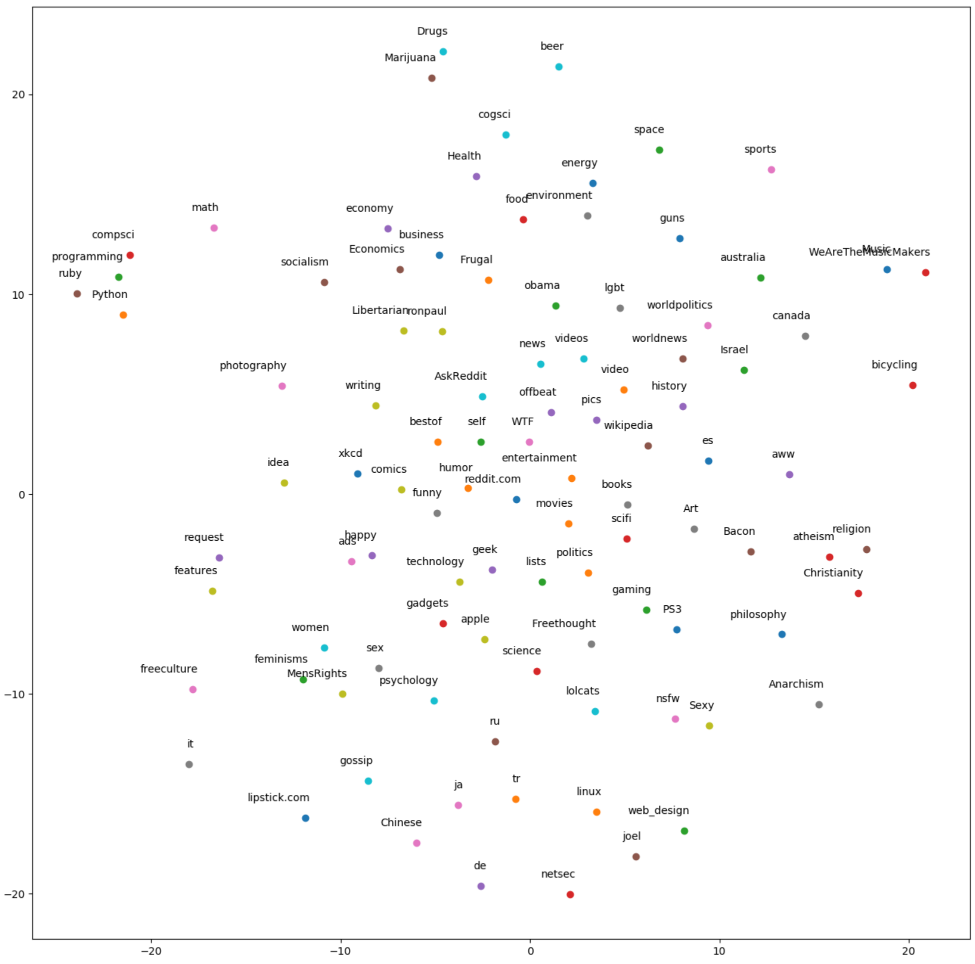

Figure 3: `doc_subreddit_vs100_ep40` model

C: t-SNE visualization of Communities (POS)

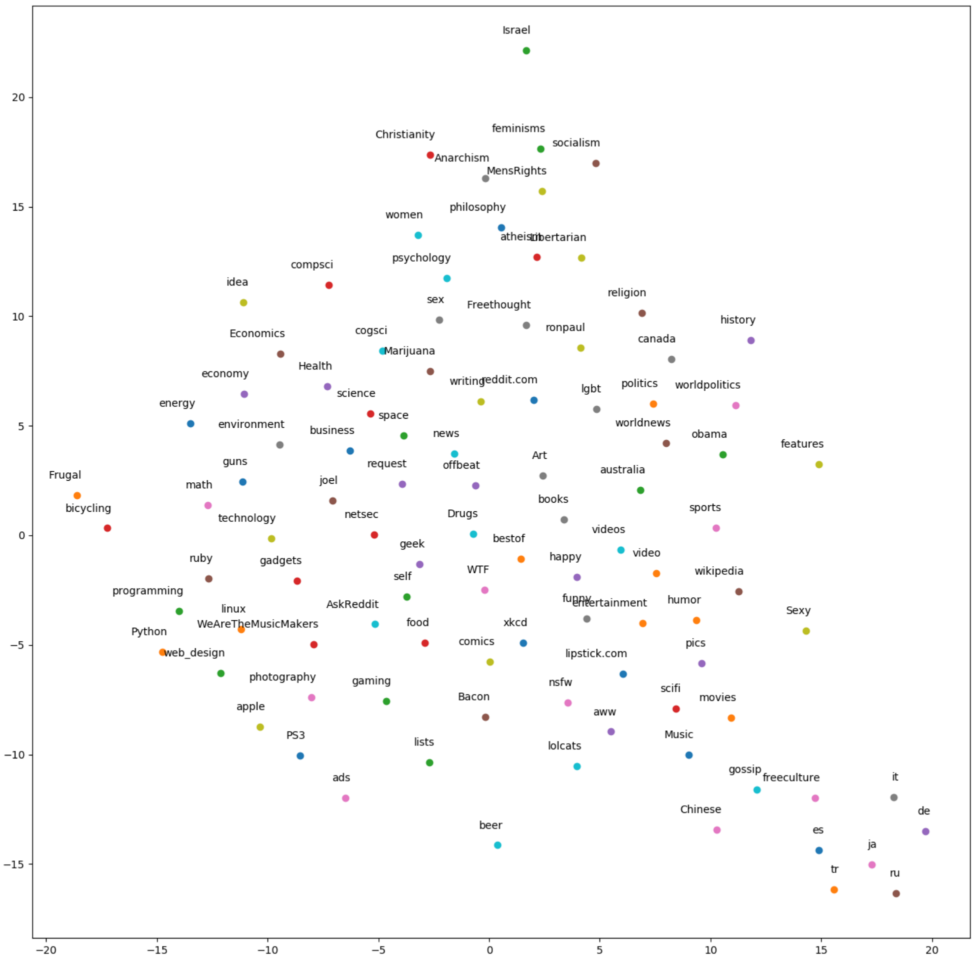

Figure 4: `doc_subreddit_pos_vs100_ep40` model