A. Covariate-specific analogies

We wish to learn covariate-specific analogies of the form a is to b as c is to d. To this end, we considered experiments of the form: for fixed words a, b, c, determine words d such that for some covariate k, the quantity

$$\frac{(c_k \odot v_a - c_k \odot v_b) \cdot (c_k \odot v_c - c_k \odot v_d)}{||c_k \odot v_a - c_k \odot v_b||||c_k \odot v_c - c_k \odot v_d||} \tag{6}$$

is small, yet for other k, the quantity is large. The intuition is that under the covariate transform, $v_c - v_d$ points roughly in the same direction as $v_a - v_b$, and d is close to c in semantic meaning.

In particular, we set a = "hillary", c = "trump", and found words b for which there existed a d consistently at the top across subreddits (implying existence of strong analogies). For example, when b = "woman", d = "man" was the best analogy for every weighting. Then, for these b, we considered words d whose relative rankings in the subreddits had high variance. The differential analogies captured were quite striking: the experiment is able to reveal words whose relative meaning in relation to anchor words such as "hillary" and "trump" drifts significantly.

Table 6. Analogies task. Each best analogy d was one of the top-ranked words in every embedding. We present words whose "relative analogy" rank was enriched in some embedding. Subreddits are color-coded: green for news-related WN and N (*worldnews*, *news*), blue for left-leaning P and S (*politics*, *SandersForPresident*), red for right-leaning D (*The_Donald*), black for A (*AskReddit*).

Word b	Best analogy d	Word	High rank	Low rank
woman	man	abysmal	1351 (S), 2218 (P)	14329 (base), 14077 (D)
		amateur	1543 (P), 3966 (S)	13840 (base), 13734 (D)
		zionist	1968 (P), 2327 (S)	14173 (base), 14248 (A)
		politician	2796 (WN), 3155 (D)	11959 (base), 10386 (S)
		president	2452 (D), 3564 (WN)	12257 (base)
		nationalists	208 (S), 606 (D)	8916 (base), 7526 (A)
democrat	republican	south	3511 (P)	11091 (base)
		bigot	400 (S), 530 (A)	12888 (D), 12994 (WN)
liberal	conservative	christian	33 (P)	12756 (D), 12722 (WN)
		white	619 (P)	12824 (D), 13273 (base)
		racist	252 (P)	12930 (S), 12779 (D)
		sociopathic	1756 (D)	13744 (P), 13389 (A)
		disenfranchised	3693 (D)	11267 (base)
politician	businessman	confidence	2768 (D)	10528 (base)
		questionable	598 (WN)	13002 (base)
		irrational	2153 (N), 3430 (P)	13305 (base)

B. Sparsity results: book dataset

Number of sparse coordinates (out of 100) were as follows, by series and then book order: Harry (0, 5, 1, 3, 7, 4, 0), Chronicles (0, 0, 0, 1, 0, 0, 0), Song (8, 8, 9, 4, 11), Twilight (6, 6, 8, 7), Screwtape (5), Strike (3, 2, 2), Host (6), Vacancy (4). While not as dramatic as in the politics dataset, the presence of zero (rather than small) coordinates across multiple runs shows that there still is specificity of topics being learned. We plot the histogram of coordinate sizes in the following figure.

C. Algorithm setting notes

We also experimented with using (Duchi et al.) as the optimization method, but the resulting weight vectors in the politics dataset had highly-overlapping sparse dimensions. This implies that the optimization method tried to fit the model to a smaller-dimensional subspace, which is not a desirable source of sparsity.

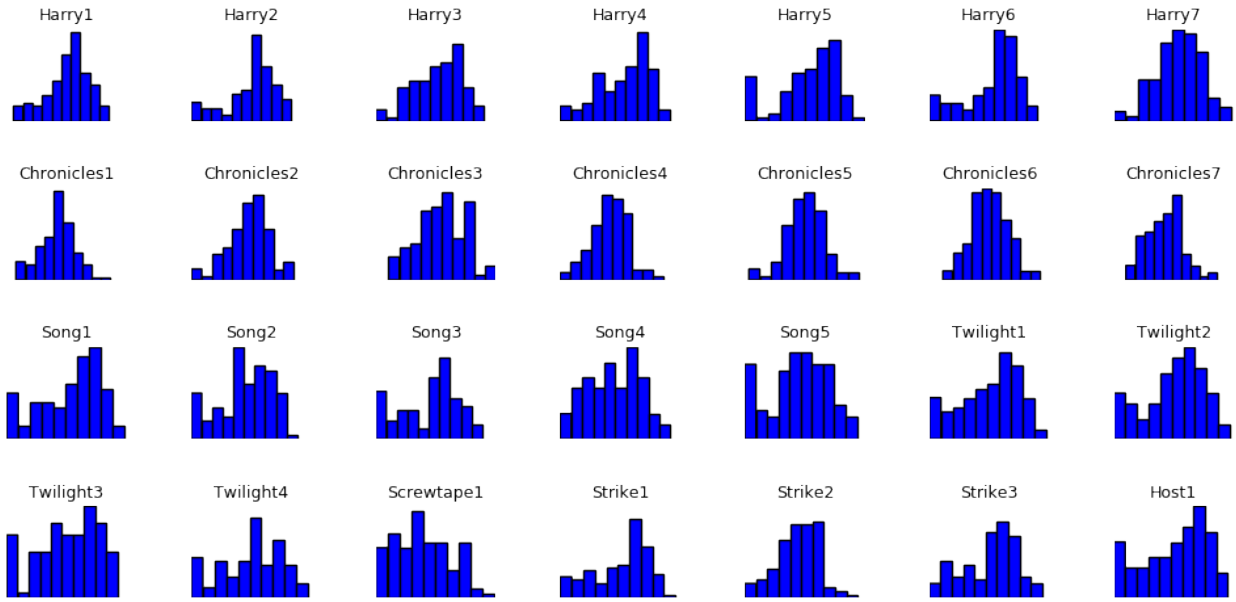
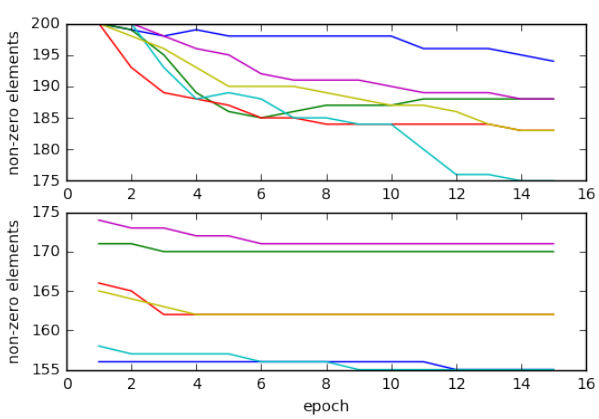
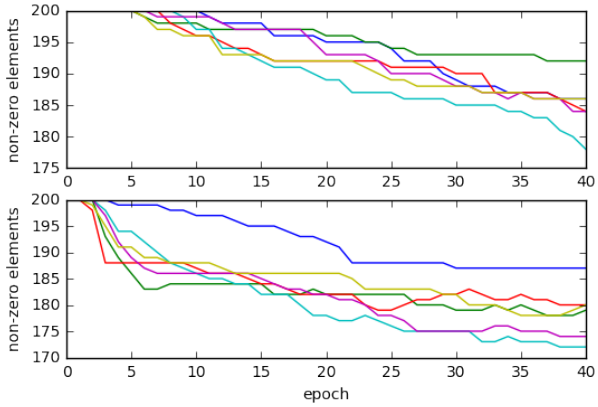


Figure 5. Histogram, sizes of weights in covariate vectors of book data.



(a) Non-zero dimensions in weight vectors by epoch and optimization method. Upper: Adam; lower: Adagrad.



(b) Non-zero dimensions in weight vectors by epoch. Upper: initialization centered around all-1 vector; lower: centered around all-0 vector.