Mining Big Data - Intermediate Report

Investigating Paragraph Vectors

 $\begin{array}{ll} a1632538 & Zachary\ Forman\\ a1646930 & James\ Caddy \end{array}$

September 11, 2016

Introduction

This document describes our progress towards achieving our project goals. We have three primary project goals, the first and second of which our current work has focused on:

- 1. Replicate the results claimed by Le and Mikolov in their 2014 paper [1].
- 2. Compare doc2vec backed with k-nearest-neighbours with standard approaches to determining document similarity [2].
- 3. Apply doc2vec to source code, and explore how it distributes code segments.

Current Progress

Work on the IMDB dataset

Some exploratory analysis and testing, using gensim [3] as an implementation of doc2vec, and the tests have been run as per the paper [1]. So far the results are disappointing. Despite several attempts to replicate the paper's experiment, and obtain a working sentiment classifier, so far, we have not come close to achieving the accuracy described in the paper, or achieved by other approaches. So far, only an accuracy of 68% has been obtained, in contrast with the 92% claimed by Le and Mikolov.

Work on the Wikipedia dataset

We again used gensim [3] to run doc2vec on the wikipedia corpus [4]. Even with only 2 training iterations on the corpus (compared with a desired 10-20), this took over 100 CPU hours. From the word and paragraph vectors trained on this data, we qualitatively evaluated the algorithm's performance on three axes:

- 1. Did the learned word vectors capture semantic meaning well?
- 2. Did the learned document vectors capture semantic meaning well?
- 3. Could the document vectors be used for similarity analysis?

Current Experimental Progress

All source code necessary to reproduce our results is available at https://github.com/mbd-doc2vec-team/mbd-doc2vec/, the IMDB dataset is available as referenced in Maas et al. 's work [5], and the wikipedia dataset is available at [4].

Experiments on the IMDB Dataset

The approach taken followed the method described in the paper as best as possible. Gensim [3] was used for the doc2vec implementation, with parameters as in Table 1 and Keras [6] backed with Theano was chosen for the neural network library, since neither were specified in the paper. It is unlikely that either of these choices were the cause for low performance, and more likely that our implementation was flawed. Both of these libraries were relatively easy to use which was good.

While the paper gives clear descriptions of most components and parameters in the classifier architecture, critical details are overlooked and explained poorly. Replicating the results without clear descriptions or values falls to trial and error, where any number of confounding variables could be in effect. One such ambiguity is the dimension of the document vectors, which can have large effects on accuracy.

Experiments on the Wikipedia Dataset

Similarly to the IMDB dataset, we trained doc2vec on the Wikipedia dataset. Due to the sheer size of the dataset, this was a lengthy process, with a single iteration taking in excess of 50 hours of CPU time. As a result (and due to numerous issues with disk space and memory consumption leading to crashes during model serialization) we only trained for 2 iterations of the Wikipedia dataset. We used the parameters in Table 1 for training.

Once we had trained word and document vectors, we performed some qualitative analysis. First, we compared the distribution of vectors in the trained Wikipedia model and the trained IMDB model. We did this by finding the vectors closest to a given seed word. Generally the wikipedia dataset, despite the smaller number of iterations, had better performance, as can be seen in Tables 2 and 3. This can likely be attributed to the much broader text corpus offered by Wikipedia, with concepts such as Apple as a company as opposed to apple as a fruit being unavailable in the movie-focused IMDB source.

Secondly, we evaluated the distribution of vectors internally by finding the nearest neighbors to some specific documents. An example of this can be seen in Table 4, where we find that the closest documents to the wikipedia page for xkcd include other web-comics 'MS Paint Adventures', 'Penny Arcade' and 'ctrl+alt+delete'.

Finally, we plotted the word vectors for some subsets of articles (since there are far too many to plot them all) using t-SNE [7]. A selection of these plots can be seen in the appendices (Figures 1-4). The salient feature of these plots was that the encapsulated conceptual relationships that have meaning to humans - for example, in Figure 2, it's clear that related concepts are tightly clustered. This reflects the algorithm's success in learning document level concepts.

Future Work

IMDB Dataset

Further clarification of the architecture would improve the performance far better than continued tuning. Unless there is any insight disclosed in discussion around the paper, contacting the author of the paper seems like the only way this might be achieved. Which is unfortunate, given the author has been so far unwilling to help others in discussion groups. In order to obtain higher accuracy, there is the possibility that running the training for longer could help. Both the embedding and the classifier could benefit from this, however over-fitting the data is a concern. Currently, achieving an accuracy within 10% of that claimed by Le and Mikolov [1] seems unlikely, and as such, serves as a goal for our next experiments.

Wikipedia Dataset

The most pressing, and most time-consuming, future work for the Wikipedia dataset is to train the model for more generations. While the word vectors and document vectors were somewhat coherent, there was clearly noise caused by under-fitting. While training for 20 generations will take 1000 CPU hours, the benefits of training for a significant number of iterations should be worthwhile.

After this, we will compare the similarity metric that can be obtained using doc2vec with those obtained from the k-shingles method, particularly with intent to quantitatively compare the speed of finding similar documents, and qualitatively assess how well the similarity metric compares to a human notion of similarity.

Chromium Dataset

References

- [1] Quoc V Le and Tomas Mikolov. "Distributed Representations of Sentences and Documents." In: *ICML*. Vol. 14, 2014, pp. 1188–1196.
- [2] Andrei Z Broder. "On the resemblance and containment of documents". In: Compression and Complexity of Sequences 1997. Proceedings. IEEE. 1997, pp. 21–29.
- [3] Radim Řehůřek and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, May 2010, pp. 45–50.
- [4] Meta. Data dump torrents Meta, discussion about Wikimedia projects. [Online; accessed 9-August-2016]. 2016. URL: https://meta.wikimedia.org/w/index.php?title=Data_dump_torrents.
- [5] Andrew L. Maas et al. "Learning Word Vectors for Sentiment Analysis". In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA: Association for Computational Linguistics, 2011, pp. 142–150. URL: http://www.aclweb.org/anthology/P11-1015.
- [6] Keras: Deep Learning library for Theano and TensorFlow. URL: https://keras.io/.
- [7] Laurens van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE". In: *Journal of Machine Learning Research* 9.Nov (2008), pp. 2579–2605.

Appendices

Parameter	IMBD model	Wikipedia model
Vector size	200	200
Context distance	5	8
Training iterations	20	2

Table 1: Parameters used for doc2vec for each experiment

Wikipedia model	IMBD model
apricot	kahuna
cherry	mines
onion	interpreters
mango	explosive
persimmon	imperialist
indigo	catholic
hazelnut	conservatory
pomegranate	pilling
yahoo	craftsmanship
ibm	casing

Table 2: Words similar to 'apple'

Wikipedia model	IMBD model
screwball	screwball
slapstick	slapstick
erotic	romance
modernist	stand-up
realist	slap-stick
satirical	physical
absurdist	teen
bromantic	sophisticated
improvisational	erotic
surreal	frothy

Table 3: Words similar to 'romantic'

Page name	Cosine similarity
randall munroe	0.517820239067
ms paint adventures	0.443829506636
time (xkcd)	0.440876543522
penny arcade	0.424748897552
alexey pajitnov	0.424251556396
ctrl+alt+del (webcomic)	0.423852056265

Table 4: Wikipedia pages most similar to XKCD

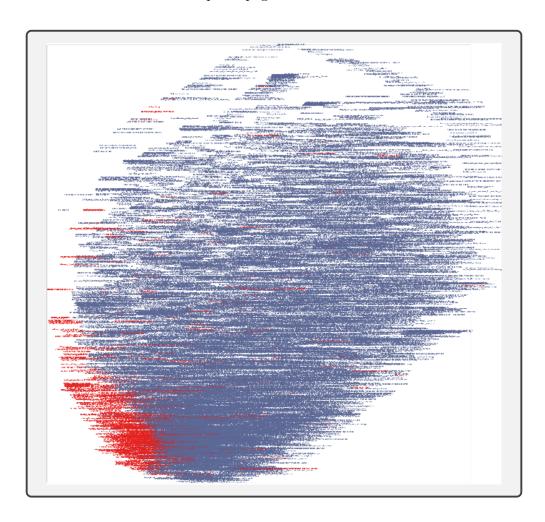
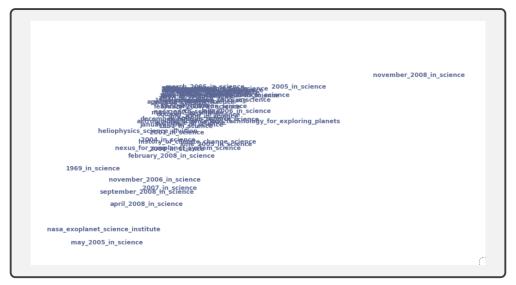


Figure 1: t-sne [7] used to visualize mining (red) and science (blue) wikipedia pages. Concepts such as 'School of Geological Sciences' are nearer the meeting point of the groups.



```
mining industry_of_senegal
mining/industry_of_senegal
mining/industry_of_senegal
mining/industry_of_the_congo
mining_industry_of_the_congo
mining_industry_of_mights_oper_mining_in_arizona
mining_injughts_instruction_arizona
mining_injughts_instruction_arizona
mining_industry_of_the_central_african_republic
mining_industry_of_cameroon
mining_industry_of_cesotho
office_of_surface_mining_and_reclamation_act_of_1977
molybdenum_mining_in_the_united_states
copper_mining_in_the_democratic_republic_of_the_congo
mining_industry_of_angola
```

```
53rd_world_science_fiction_convention
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           science-fiction_plus
                                       science_fiction_studieseashington_science_fiction_association
the_yeaជក្រក្នុងក្នុងក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រុមក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្បក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួនក្រកួ
                                                              the best science fiction of the year 9 the year science fiction of the year substance fiction the year science fiction the year's best science fiction the year's best science fiction the year's the year's limited for the year's the
                                                                                                                                  1937_in_science_fiction
                                                                           the_year's_best_science_fiction:_sixth_annual_collection
46th_world_science_fiction_convention
                                                                                                                                                                                                   the\_year's\_best\_science\_fiction:\_fourteenth\_annual\_collection\_20th\_world\_science\_fiction\_convention
                                                                                                                                                                      ..._avience_iction
21st world_science_fiction_convention
engenfictionsels.the.uear.4-
                                                                                                                                      the best signed actions of the best signed and the best signed action of the best signed action 
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      30th_world_sci
                                                                                                                                                                                                                                         science_fiction
new_wave_science_fiction
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    the_five_fists_of_science
                                                                                                                                                                                                                                 the_best_science_fiction_of_the_wear_3
                                                                                                                                                         the_best_science_fiction_stories:_1952 robest_adams_science_fiction_sy_thirders_of_america
                                                                                                                                                                                                                               the \_year's\_best\_science\_fiction: \_seventh\_annual\_collection
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    science_fiction_film
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              nonscience
the_year's
list_of_science_fictionoraldid_fanta_syjentia_fiction:_1971
                                                                                                                                                                                                                                                    science fiction fastasy translation awards the besty seign disting of the ign as sociation
```

Figure 2: Some conceptual clusters from Figure 1.

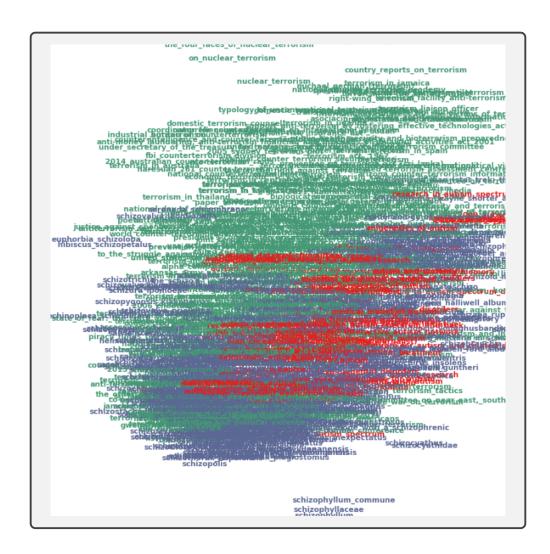


Figure 3: t-sne [7] used to visualize autism (red), terrorism (blue) and schizo.* (blue) Wikipedia pages. Note how autism and schizophrenia are clustered together, in the top right.

		schizor	ohonia
schizovalva_isochorda			
schizovalva_triplacopis orthern_ireland)_act_2006	sociological	_and_cultural_aspects_of_autism	neoschizomer isoschizomer
schizax			
schizovalva_alaopis	autism_research_ce	entre outline of autism	
		journal of autism and develop	mental disorders
schizoxylon autism_is_a_wo	rld	jeuu.je.juu.e.u.juu.e.e.e.	diploschizia_
lelphis	athletes_again	st_autism	
autism_servi	ce_dog	schizothyriaceae	diploschizia_habecki
autism_friendly			dip
n_research		maia chung autism and disabilit	ies foundation
sex_differences_in_autism	schizophasia	maia_chung_autism_and_uisabilit	ies_roundation
n_shaw_(autism_researcher)			
bout_schizophreni <mark>persecution_of_pe</mark> ople	_with_autism _{schizother}	pleistodo	ntes_schizodontus
schizothorax g	ıriseus	revista_autismo	schizophyte
utism_30	metaschizother n_treatment_evaluation		
	auti	ism cymru	
itha cycloschizon	duti	schizotrichaceae	
schizoparme		research_centre	schizophonic_(us3_al

	schizovatva_un	iitincta	schizo	parme			wales auti	sm research co	entre			
isophanes schizovalva_ochro	tincta	. schizoscyphus							ear_east,_south_a	sia,_central_as	ia_and_counterterroris	m
in_the_grip_of_ushi20v	itinskhizavalva_	xylotincta -schizodiscus	formations_of	the united s	tates_army_durii	ng_the_war_c	n_terrorism				schizella	schizos
			hizochila		schiz	ophrenia res phanomor	earch forun	esgyWizoporella	unicornis			
na	acinzo		schiz	ochlamys	schizonhrenia	of modern e	thical theor	ies	Cannearing		mayoschizocera	
harpalus chilos	chizontuperischi	zojchizozygia ^{schizo} gyn	schizoid_avoidant_	penavior						schizothyr	ium_pomi	
musa_schizocarpa deoxyse			chizomeris schiz	ophrenia_disc	rder_10_(period	ic_catatonia)	schizochla		schizopterio	dan)
		oglosšáhi/aktiv/akthora	cnizomeris x_wangchiachii						enia (sepultura all		rme straminea	schizoceratomyia
slamic terrorism in the	balkans				i lenside rop <u>t</u> rin a i	enteur strogb	(han they)	schizophre	rnia_(sepuitura_aii	bum)	_	schizo
	mloschizon		hizothorax_skardu dermining israeli								schizochiton	schizo
vestigation_measures_		chizoth <u>oraxo</u> myzasčan	Mfopogon			aschiza		eleutherosch	izonidae			
terrorism_in_the_euro	pean union (video game)	john_ni	cholas_brown chai	roffeentget	SEMESTHORAX NO	submibus .	schizothora	x_dolichonema	res	trepia schizos	anala	
PSMd Stts_of_terrorism	against_american	75 schizothorax rotun	dimaxillaris			-			schizothorax_zai		- para	
	gioritying_	schizothorax molesw	izothorax_lissolabi	atus schizotho	rax_heterochilus					total and total		
errorism_in_greece		schizothorax_molesw	Path political Selvi	opyge daine	liischizothorax_r	asus	schizo	pashizopodidae		schizolobi	um_amazonicumgchizot	horax_microstomus
urban_terrori												angiostoma
many	schizothorax_me	eridionalialliance_for_tl	he_restoration_or_]	peace_and_co	Schizol	thorax_pelza	ni					
rism in kenya			schizotho	rax_curvilabia	tus ithorax_heteropl ongi	schiz		udoaksaiensist	nizogregarinina			
uide to genre-terrorism			scl	hizothorax_ch	ongi schizothorax b	iddulnhi					megaschizomus	
uide_to_genre-terrorism		eco-terrorism_i	n_fiction osa regional_spec	ial forces con	inter-terrorism	onference			schizosaccharp	70XSR5haromy	es pombe	
		schizoo	cosa					izothorax_ocor	nori	,	ies_pennie	
schizoseridead	schiz	zothlisa of didtional ch	aracters on the ar	atisimo bineratora	richardsonii	schizothor	orax_nudive	ntris				rhynchoschizus
schizoserideae					iggthorax_pregla							
enthus				.mzura_umee	ma		hizothorax	epidethorax_				
list_of_terrorism_films					schizothorax			scnizotnorax_	granami			
	combating_t&	Aignan_Uhter			schizothorax	beipanensis				schizotho	rax_prenantischizotho	rax_huegelii
				schiz	ophonic	schizot	horax_arge	ntatus				
			chizopyge_niger		schizochytriu	m limacinum			schizothorax_e	longatus		
						-	schia	othorax				
chizothorax_cryptolepis	thorax_davidi	schizothorax_dulonge	ensis			othorax oligo	lands.					
			schizot	thorax vunna	nensis vunnaner		iepis					
											schizoda	ctylidae
			schizothe	ardai radhamara	y aisnamėnisią epas		opyge		schize	ochytrium		
				-								
				rax_yunnane	nsis schizoid man							
		schizothorax macr	ophthalmus	the_	scnizoio_man			rehizothora	x plagiostomus			
	schizoth	norax_nepalensis						scinzotnora	x_pragrostomus			

uggish_schizophrenia ^C	•	pigenetics_of_autism	history_of_schizophrenia		
hiza_koenigiana		heritability_of_autism	post-schizophrenic_depre	ession	
	effects_of_estrogen_c	on_schizophrenia epigenetics_of_schi	schizophrenia zophrenia	schizachne	
	schizophrenia_and_smoking	dia	agnosis_of_schizอกกid_schizophrenia		
	prognosis_of_schizophrenia schizotypy	pediatric_schizo	schizoaffective_disorder phrenia construction_of_schizophrenia		schizo
nics) childhood_autism_rat	ting_scale				
	management of schizophrania causes of Schizophrenia	_schizophrenia			
	con	nparison_of_bipolar_diso	schizophreniform_disorder rder_and_schizophrenia		
interpre	etation_of_schizophrenia				
	sci	hizaster			
n_embassy_bioterrorism_hoa	dopamine_hypothesis of schizo mechanisms	pphrenia s_of_schizophrenia			
	animal_model_of_	visual_processing_abno schizophrenia	ormalities_in_schizophrenia schizo	chilus	
schizotypal_personal	lity_disorder	autism_and_m	nemon/		

Figure 4: Some conceptual clusters from Figure 3. $\,$