Visualizing Topic Models with PCA

Aja Klevs (ak7288) and Trevor Mitchell (tim225)

May, 2020

This document contains 2,151 words

Introduction

Blogs are a modern way for people to express their interests, activities, and inner-thoughts. Within the intersection of data science and social science, a natural question emerges. How can we leverage blog texts to gain insights into the various topics that people are interested in? In addition, do the topics that people blog about vary based upon age or gender? Answers to these questions could potentially result in commercial advantages such as better content curation and targeted marketing. In this paper, we will explore and try to answer these questions using Latent Dirichlet Allocation (LDA) and Structural Topic Model (STM) Topic Models. We will then try to visualize our findings using various applications of Principle Component Analysis (PCA).

Literature Review

There have been several similar works regarding blog data analysis throughout the years. Social science researchers have examined online identity and language use among teenagers who create blogs (Huffaker and Calvert 2005). While the research focuses on teen expression through blogs, they also dissect their findings based on age and gender. For example, they discovered that male and female authors were equally likely to author blogs about romantic relationships. In addition, there have been works that have leveraged supervised machine learning techniques to try and predict a bloggers age and gender (K Santosh, Romil Bansal, Mihir Shekhar, and Vasudeva Varma 2013). In their work, they also use LDA to build topic features. Their assumption was that different categories of people tend to write about completely different topics and thus it would help them discriminate based on gender. However, upon using these topic features, they were only able to obtain a 54% accuracy for gender classification.

Theory and Hypotheses

In our analysis, we will be exploring which topics occur in the blog authorship corpus and whether people tend to blog about different topics based upon age or gender. We are expecting to

discover a latent structure of topics that will allow us to differentiate between the most important topics based on age or gender. Furthermore, we are expecting to be able to visualize our discovered topic structures by showing some degree of separability by topic, gender and age using principle component analysis.

Data and Methods

I. About the Data

For this analysis we will use a random subset of 9,884 documents from the Blog Authorship Corpus. The corpus is composed of 19,320 bloggers, 681,288 posts containing over 140 million words. The corpus contains posts from up until 2004 and represents authors spanning a variety of ages between 13 and 47. There is an equal number of blog posts from male and female authors. We also categorized the data in accordance with the age breakdown categories presented in the description: "13-17", "23-27", and "33-48" which we call *tens, twenties* and *thirties*. For these groups, they accounted for approximately 34%, 48%, and 18% of the dataset respectively.

II. Preprocessing

Originally, we removed english stopwords based upon R's quanteda package. However, when we ran our topic models, many of our topics consisted of nonsensical words that could provide no real semantic meaning about a topic. Therefore, we decided to remove a list of search engine optimization stop words. As many blogs used words such as "im", we also removed the non-apostrophed versions of the stopwords. In addition, we performed standard preprocessing techniques such as lower casing all words and removing punctuation. However, stemming was not performed, as it is less commonly used with topic models and stemming increases the number of features in our dfm after trimming, and therefore increases computation time.

After our dfm was created, we decided to trim it based upon the document count. We used a minimum thresholds for term frequency of 168 and document frequency of 112. The minimum term frequency would prevent rare words from unproportionally skewing the topic of a document, while the minimum document frequency would allow us to avoid including words occurring often in only a small number of documents, since these words could be misleading topic indicators.

III. Methods

We used two algorithms to create our topic models; LDA and STM. We are also using STM, because we recognize that LDA makes some assumptions that may be inappropriate in this context such as the topic distribution having no dependence on document metadata. For LDA,

the user must specify the number of desired topics *k* and the algorithm will give a topic probability distribution for each document. We choose *k=8* for reasons we describe in the results section. The STM algorithm also gives a topic probability distribution for each document but we used the spectral learning method to determine the ideal number of topics *k*, which ended up being 47.

We then performed several types of analysis to better understand our results. For both kinds of topic models, we found the most contributing words in order to guess titles for the topics. For LDA only, we calculated the average topic scores across gender and age. Finally, we sought to visualize how our topics were distributed in space and how gender and age interacted with the topics using PCA.

In our first visualization, we computed the document frequency matrix (DFM) for the corpus, and then computed its two largest principle components. We used these to graph each document in 2d space. We then found the most probable topic for each document and labeled our graph accordingly to see if the documents clustered by topic.

Another way in which we visualized the layout of our topics was to use the *beta* matrix associated with our topic models. The *beta* matrix gives probabilities that each word in our vocabulary is associated with each specified topic. Using the words as features, we projected our *beta* matrix onto the first two principle components to visualize the spatial layout of the topics in two dimensions.

Thirdly, we performed PCA again but this time with the topic distributions as variables. We then plotted each document in terms of the top two largest principle components, and labeled each in terms of age and gender to see if there was any clustering by age or gender.

LDA Results

I. Finding the Topics

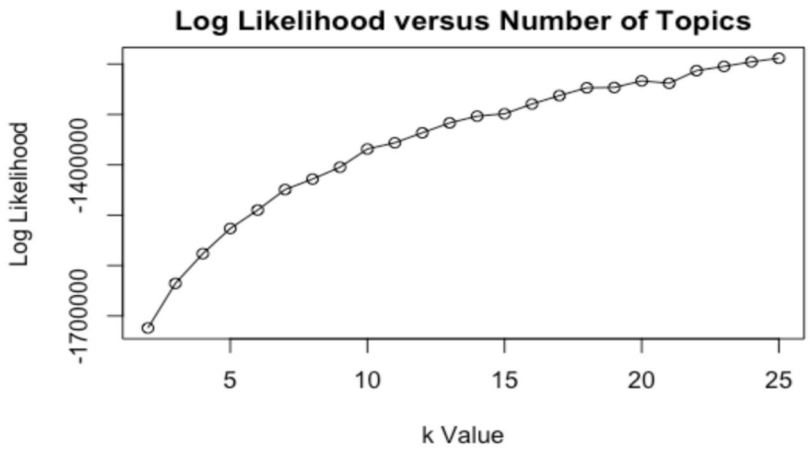

Fig. 1

As shown in Figure 1, we continued to increase the value of *k* the log-likelihood of our LDA model continues to increase. However, high values of *k* are hard to interpret. Therefore, we decided that a value of *k=8* was a good tradeoff between complexity and log likelihood.

Most Likely Words per Topic

Topic #	1	2	3	4	5	6	7	8
1st	day	work	people	bad	life	told	good	school
2nd	night	time	world	good	love	theft	time	blog
3rd	home	year	=	guy	feel	car	great	read
4th	today	job	years	yeah	time	man	show	movie
5th	house	money	point	guess	god	room	big	post
6th	morning	call	fact	stuff	friends	asked	play	good
7th	week	day	group	today	about	started	game	story
8th	hours	leave	part	pm	person	head	place	book
9th	sleep	honey	bush	guys	live	thought	music	class
10th	yesterday	phone	>	fun	people	eyes	song	write
Guess	**daily life**	**work**	**stats**	**dating**	**happy**	**crime**	**concerts**	**school**

Table 1

Table 1 shows the top ten words for each of the eight topics. The last row is our guess at a title for this topic.

II. Visualizing Topics in Space

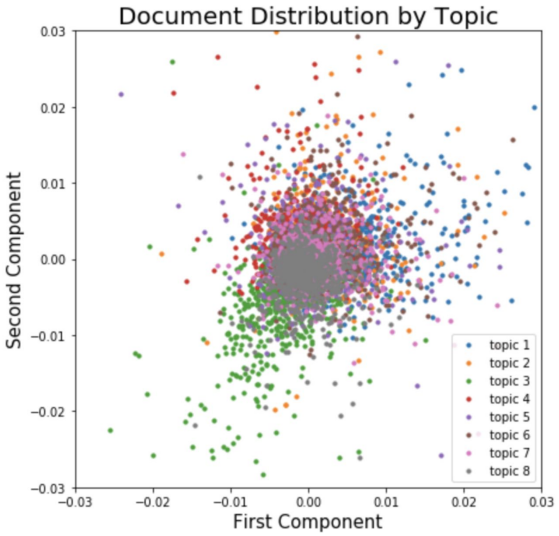

Fig. 2

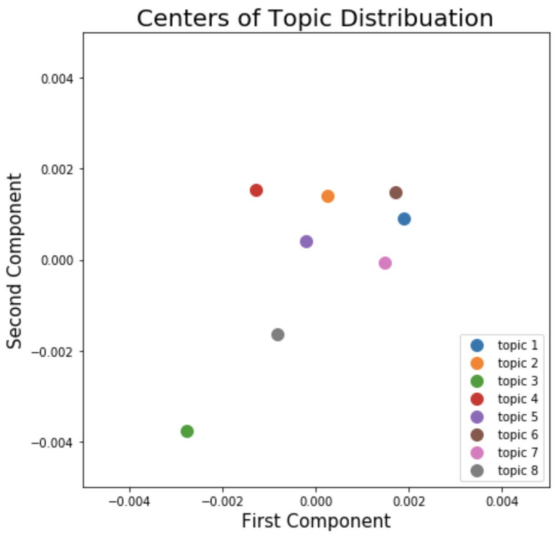

Fig. 3

Each point in Figure 2 represents a document. We represented each by its dfm vector and projected it onto the first two principle components. The documents were then labeled by the most likely topic attributed to them. We plot the centers of each topic's cluster to better visualize the distribution of the topics in Figure 3.

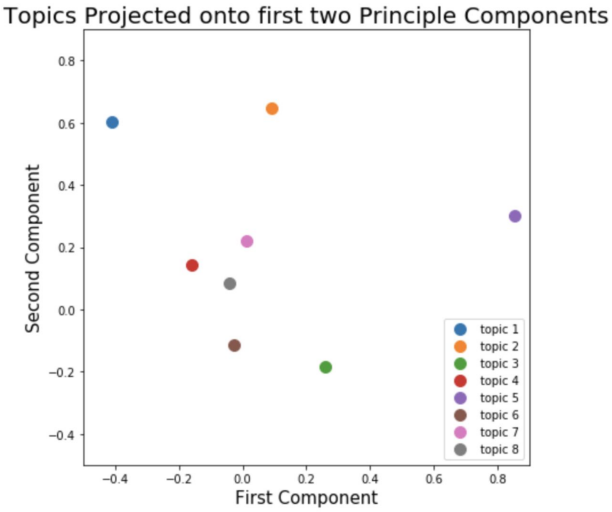

Fig. 4

As a different visualization approach, we represented each topic by its *beta* distribution on the words in the vocabulary. We then use PCA to visualize the topics in two dimensions in Figure 4.

III. Topics by Gender and Age

Topic Likelihood by Gender

Topic #	1	2	3	4	5	6	7	8
Avg. Male Score	.124	.124	.130	.122	.123	.123	.126	.127
Avg. Female Score	.130	.123	.119	.127	.129	.127	.122	.123
Difference	.006	.001	.011	.005	.005	.004	.004	.004

Table 2

Topic Likelihood by Age

Topic #	1	2	3	4	5	6	7	8
10s avg. Score	.128	.117	.116	.140	.129	.123	.122	.126
20s Avg. Score	.127	.127	.127	.118	.126	.125	.125	.124
30s Avg. Score	.123	.128	.137	.112	.122	.128	.125	.125
Age with Highest Score	10s	30s	30s	10s	10s	30s	20s	10s

Table 3

In Table 2 and Table 3 above, we find the probability that each topic is associated with each gender/age group.

For Figure 5 and Figure 6, we represented each document as a vector of its topic probability distribution. We then labeled the documents by gender and age respectively, and used PCA to visualize.

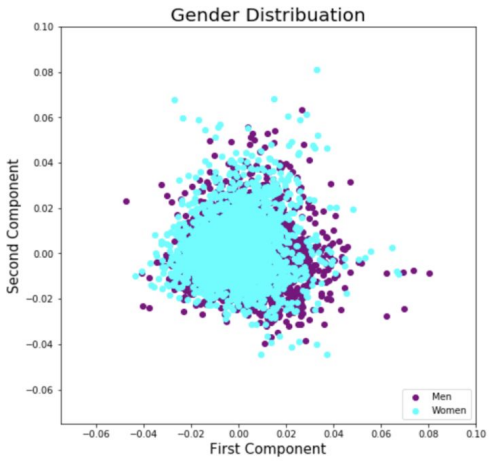

Fig. 5

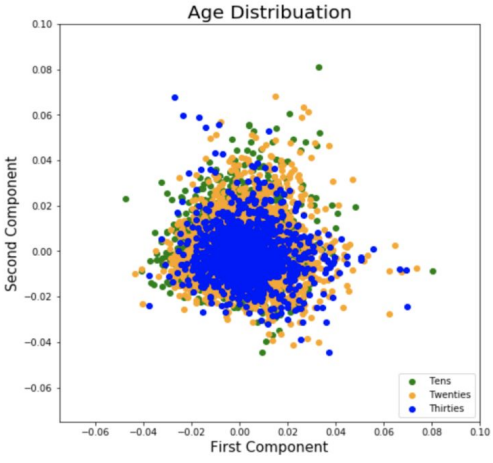

Fig. 6

STM Results

I. Finding the Topics

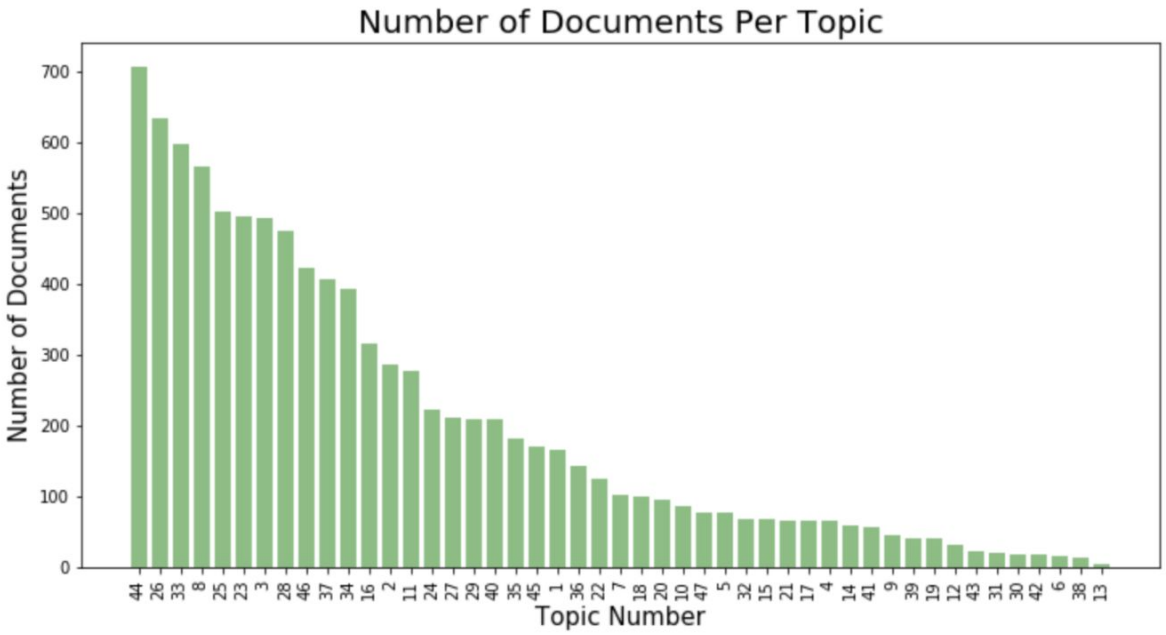

Fig. 7

Topic #	44	26	33	8	25
1st	live	feals	straight	learned	park
2nd	war	wear	lives	afraid	complete
3rd	season	anymore	hate	friend	real
4th	add	music	evil	strange	late
5th	forget	imagine	hope	watch	sun
6th	radio	answer	finished	involved	english
7th	entire	saturday	learn	woman	title
8th	songs	broke	wanted	picture	speak
9th	yesterday	written	favorite	friends	site
10th	normal	talked	font	monday	type

Table 4

As shown in Table 4, The top words for each topic here were less straightforward in terms of inferring the general themes, so unlike with LDA we did not try to guess titles for the topics.

II. Visualizing Topics in Space

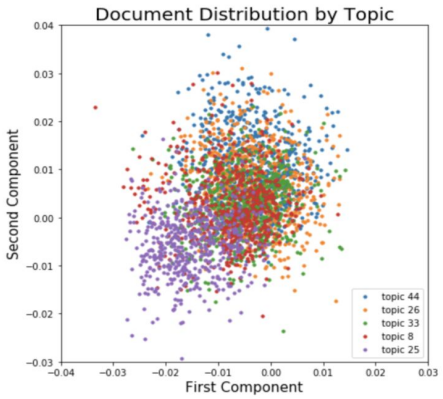

Fig. 8

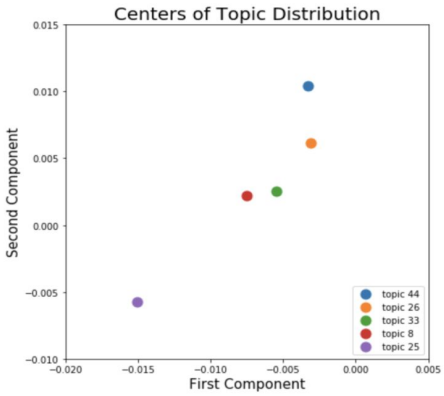

Fig. 9

In Figure 8, as with Figure 2, we represented each document by its DFM vector, and used PCA to visualize each point in 2D space. Above, we only show documents associated with the most prevalent five topics, labeled by topic. Also as with the LDA, we show the centers of the clusters below in Figure 9 for visual clarity.

III. Topics by Gender and Age

Since the topics were so difficult to interpret, we decided to forgo finding the breakdowns in terms of age and gender per topic.

As with the LDA model, we represented each document as a vector of its topic probability distribution. We then labeled the documents by gender and age respectively, and used PCA to visualize.

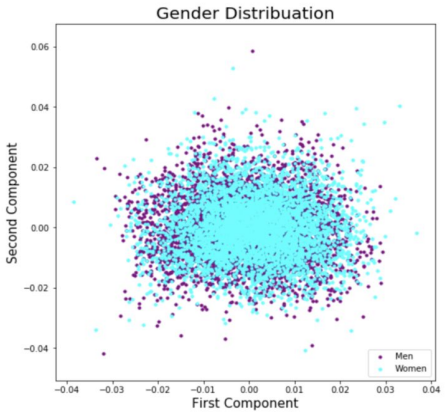

Fig. 10

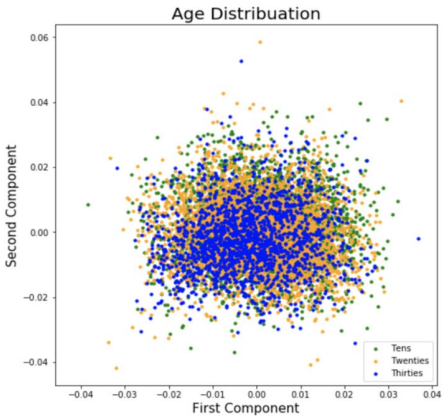

Fig. 11

Discussion of LDA Model

I. Visualizing Topics in Space

In our first visualization (fig. 2 and fig. 3) it is evident that there exists slight differences in which words apply to each topic. Especially topic 3 (which we guessed to be statistics) seems to occupy a different area of the 2d space. The various sizes of the clusters in fig 2. suggests some topics, like topic 3, are more eclectic in terms of word use while some seem to have smaller vocabulary such as topic 8, which we guess to be school.

Our other method of visualizing the spatial layout of the topics (fig. 4) suggests that topics 4, 7, and 8 (dating, concerts, and school) are the closest together in terms of words associated with the topic. This seems sensible, given a presumed audience overlap which is partially verified when we look at the topics by age below.

II. Topics by Gender and Age

Overall, we did not find huge differences in topics in either gender or age. In fact, in fig 5. and fig. 6 there is no visible clustering by gender or age. The only inference we could make from those visualizations is from the various cluster sizes. For instance, from fig. 6 we may infer that the authors in their thirties may use a smaller range of vocabulary.

Overall, topic probabilities for the different ages and genders were very evenly distributed (Table 2 and Table 3). The only topic with a significant differential in terms of gender was topic 3, which we guessed to be statistics. The topics most associated with authors aged 13-17 were topics 1,4,5 and 8 (daily life, dating, happiness, and school). The topic most associated with authors in their twenties is topic 7, which we guessed to be concerts. The topics most associated with people in their thirties and forties are topics 2, 3, and 6 (work, statistics, and crime). However, the topic most unevenly distributed among age groups was topic 4, which we guess to be dating. According to the LDA model, teenage authors are significantly more likely to talk about topic 4 than any other age group.

Discussion of STM Model

Unfortunately, due to the large number of topics and the challenge of interpreting these topics, we received far less information from the STM model results. As with the LDA model, it seems as though some topics have reasonably different words associated with them, and therefore occupy different parts of space. It also seems like there are no detectable differences between topics in terms of gender or age.

Overall Conclusions

After performing LDA with eight topics and STM with spectral learning (which produced 47 topics), we found much more use from the LDA in terms of topic interpretability and meaningful associations with age and gender. This could possibly be due to the smaller number of topics utilized in LDA. Perhaps the topic distribution of this dataset is not really a function of gender or age.

For both topic models we were able to create some form of meaningful spatial representation of the topics, and for the LDA model these spatial representations fit somewhat well with our intuitions about the topics.

Both topic models indicate that there is not much difference between topics based on age or gender. The only exceptions could be the topic we guessed to be statistics, which was much more associated with male authors than female authors, and the topic we guessed to be dating which was much more associated with teenage authors than older ones.

Before drawing any conclusions from these topic models, we cannot rule out any systematic bias in data collection. Furthermore, this dataset was last updated in 2004 and we must not assume that the language of blog authors has remained constant in the interim.

With these qualifiers in mind, it is conceivable that male authors in the 1990s and 2000s used more language associated with statistics, and that teenagers of that same era also used language more associated with romance or dating. Further experimentation with varying numbers of topics and more modern datasets could provide further insights into the differences, or lack thereof, between blog topics for authors of different genders and ages.

Refferences

Huffaker, D. and Calvert, S., 2020. Gender, Identity, And Language Use In Teenage Blogs. [online] Available at: <Select Gender, Identity, and Language Use in Teenage Blogs David Huffaker-Sandra Calvert - https://onlinelibrary.wiley.com/doi/full/10.1111/j.1083-6101.2005.tb00238.x> [Accessed 8 May 2020].

Santosh, K., Bansal, R., Shekhar, M. and Varma, V., 2013. [online] Pdfs.semanticscholar.org. Available at: <https://pdfs.semanticscholar.org/af51/a290f62912f24b11aabdfa81e3c0e6e430af.pdf> [Accessed 8 May 2020]