Topic Modeling by Solving Non-Convex Problem

Dongkyu Kim
kimdongk@oregonstate.edu

CS539-002 CVX OPT, Fall Term 2019

Abstract

In this term's final project, I implemented and tested the Anchor-Free algorithm and compare its performance with two traditional topic modeling algorithms, LDA and NMF, and one unsupervised learning method, Kmeans Clustering. Anchor-word topic modeling has serious scalability and identification issues. To solve this problem, Separability / Anchor-Word assumption and Sufficiently Scattered assumption were studied as explained in [1]. The problem formulation for the algorithm found that it is a non-convex problem which cannot achieve the optimal solution directly. The feasible way to solve the non-convex problem is finding an approximate solution through iteration. Based on these ideas, the Anchor-Free algorithm for topic modeling is established. For the performance test, 6090 tweets regarding the climate-change were used, and the simulation results are provided.

0. Introduction

Knowing what opinion or impression the public has has become a valuable task in marketing as well as politics. Traditionally, social science has developed qualitative research through a survey methodology. However, the advent of social network services, e.g., facebook, twitter, and Instagram, people express their thought with numerous data no matter where and when they are.

More than ten years ago, when twitter and Instagram were just introduced, the social network analysis research was arduous work due to the lack of big data handling skills. Collecting data and cleaning data into the applicable dataset was overwhelming tasks. Notably, each tweet must have been segmented in word by word, and it was challenging to understand the relationship between words. Fortunately, current machine learning skills have made those works sufficiently feasible.

In modern society, the economy, social events, culture, etc. influence each other and mutate or develop together. The rate of change is not inversely proportional to the amount of data generated. The faster the information spreads, the more considerable the amount of data, and accelerate the shift. Therefore, public opinion is also likely to change its form over time. The public's reaction is immediate, especially when there are political events such as presidential elections or economic events such as Apple's iPhone launch. Traditional survey methods for political and economic responses have a limitation in catching up on the change. Besides, data collection is proportional to time. Public opinion in the public sphere has become the past when data collection is finished.

Therefore, despite the computing power for high-speed data processing, the limited data collection period, a relatively small sample size, if a subset of public opinion in the public field can be meaningfully identified, it will be possible to respond quickly to the public opinion.

In this project, I apply Anchor-Free Correlated Topic Modeling [1] proposed by K. Huang, X. Fu, and N.D. Sidoropoulos to simulate with a small dataset. Firstly, I summarize the background and analyzed algorithm of the model and compare performances of the Latent Dirichlet Allocation (LDA) model, Non-negative Matrix Factorization(NMF) model, Kmeans Clustering, and Anchor-Free Topic Modeling Algorithm.

1. Topic Modeling: NMF

A document corpus $D(v, d)$, where $D \in \mathbb{R}^{V \times F}$, is a word($v$)-document($d$) matrix which means the term frequency of v th word in d. The topic model is represented as below [1]:

$$D \approx CW, \text{ C} \in \mathbb{R}^{V \times F} \text{ is a topic matrix.} \tag{1}$$

- $C(:, f)$ represents the probability mass function(PMF) of the topic f over a vocabulary of words
- $W(f, d)$ represents the weight of the topic f in document d.

Since $C \geq 0$ and $W \geq 0$, it becomes the separable NMF. NMF admits a unique solution if and only if C and W satisfy sparse-related conditions.

2. Assumption 1: Separability / Anchor-Word

In machine learning, a nonnegative matrix factorization(NMF) model is one of topic modeling methods with the assumption that every topic has an anchor word. The assumption is that there exists a set of indices $\Lambda = \{v_1, ..., v_F\}$ such that $C(\Lambda, :) = Diag(c)$, where $c \in \mathbb{R}^F$ [1]. Anchor-Word is such that every topic f for $f = 1, ..., F$ has a 'special' word that has non-zero probability of appearing in topic f and zero probability of appearing in other topics [1].

By the geometric interpretations of NMF, we can suppose not only C is non-negative but also it satisfies a row-sum-to-one assumption or row-stochastic assumption by the concept of probability simplex. According to Successive Projection Algorithm(SPA), one of the simplest NMF algorightm, its normalization procedure is [1]:

$$X = D^T \Sigma^{-1}, \text{where } \Sigma = Diag(1^T D^T) \tag{2}$$
$$X = WH \tag{3}$$
$$\text{where } W(:, f) = W^T(f, :)/\|W(f, :)\|_1, \tag{4}$$
$$H(f, v) = C(v, f)\|W(f, :)\|_1/\|C(v, :)\|_1\|D(v, :)\|_1 \tag{5}$$

Especially, by the row-stochastic assumption, $H^T 1 = 1$, $H \geq 0$, if $W \geq 0$ can be derived. It is constrained to the probability simplex. When H is row stochastic, $x_f \in conv\{W\}$, where $conv\{W\} = \{x \in \mathbb{R}^F | x = W\theta, \theta \geq 0, \theta^T 1 = 1\}$ denotes the convex hull, and $X(:, f) \in cone\{W\}$. If the columns of W are affinely independent, $\{W_1 - W_0, ..., W_F - W_0\}$ is linearly independent, it is simplex.

It explains that columns of X all lie on the simplex spanned by the columns of W, and the vertices of the simplex correspond to the anchor words. However, it cannot avoid the disadvantage of normalization as a deflation procedure: noise amplification and serious scalability issues.

3. Problem Formulation

In order to find an identification criterion for the ground-truth E and C removing ambiguity brought by a non-trivial matrix, (F. C) propose the following identification criterion, a non-convex problem [1]:

$$\underset{E \in \mathbb{R}^{F \times F}, C \in \mathbb{R}^{V \times F}}{\text{minimize}} \quad |\det E| \tag{6}$$
$$\text{subject to} \quad P = CEC^T, \tag{7}$$
$$C^T 1 = 1, C \geq 0 \tag{8}$$

Where $E = \mathbb{E}\{WW^T\}$ is a topic-topic correlation matrix, the matrix P is a word-word correlation matrix.

From the identification criterion, the proposition 1 says,

> **Proposition 1** *Let (C_*, E_*) be an optimal solution of (6). If the separability anchor-word assumption is satisfied and rank(P) = F, then $C_* = C\Pi$ and $E_* = \Pi^T E \Pi$, where Π is a permutation matrix.* [1]

The proposition 1 is a counter-check of the identification criterion, and it is valid under the anchor-word assumption. However, it can guarantee the identifiability of C and E when the anchor-word assumption is grossly violated because the above identification criterion is non-convex, and it cannot find its exact optimality.

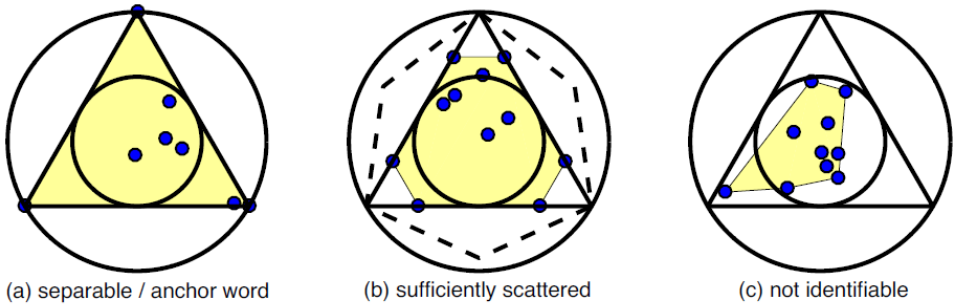

Figure 1: Illustration of the separability and sufficiently scattered conditions where M = R = 3.[1]

4. Assumption 2: Sufficiently Scattered

If the proposed criterion is under a much more relaxed condition, such as not find a exactly separable but approximately separable, it identifies topics.

> **Assumption 2. Sufficiently Scattered** *Let cone$(C^T)^*$ denote the polyhedral cone$\{x : Cx \geq 0\}$, and κ denote the second-order cone$\{x : \|x\|_2 \leq 1^T x\}$. Matrix C is called sufficient scattered if it satisfies that: (i) cone$(C^T)^* \in \kappa$, and (ii) cone$(C^T)^* \cap$ bd$\kappa = \{_f : \lambda \geq 0, f = 1, ..., F\}$, where bd$\kappa$ denotes the boundary of κ, i.e., bd$\kappa = \{x : \|x\|_2 = 1^T x\}$.* [1]

It can be seen by the concept of dual cone. $cone(C^T) = \{C^T\theta : \theta \geq 0\}$ is the dual cone of the conic hull of the row vectors of C. From it, $cone(C^T)* = \{x : Cx \geq 0\}$. Let there are two convex cones κ_1 and κ_2. If $\kappa_1 \leq \kappa_2$, then $\kappa_2^* \leq \kappa_1^*$. As such, we can say that $\kappa^* \subseteq cone(C^T)$. The dual cone of κ is another second-order cone, called ice cream cone, $\kappa^* = \{x|x^T 1 \geq \sqrt{F-1}\|x\|_2\}$, and it is tangent to every facet of the non-negative orthant. Therefore, if C satisfies the sufficiently scattered condition, which means that the rows of C are spread enough in the non-negative orthant as shown in Figure 1.(b).

An important feature of the sufficiently scattered condition is that it does not need to normalize the data columns to a simplex, and it could eliminate the disadvantage of normalization: amplifying noise as serious scalability issues.

As the result of the above condition,

> **Lemma 1.** *If $C \in \mathbb{R}^{V \times F}$ is sufficiently scattered, then rank(C) = F. In addition, given rank(P) = F, any feasible solution $\widetilde{E} \in \mathbb{R}^{F \times F}$ of Problem.* [1]

Therefore,

> **Theorem 1.** *Let cone$(C^T)^*$ denote the polyhedral cone$\{x : Cx \geq 0\}$, and κ denote the second-order cone$\{x : \|x\|_2 \leq 1^T x\}$. Matrix C is called sufficient scattered if it satisfies that: (i) cone$(C^T)^* \in \kappa$, and (ii) cone$(C^T)^* \cap$ bd$\kappa = \{\lambda_f : \lambda \geq 0, f = 1, ..., F\}$, where bd$\kappa$ denotes the boundary of κ, i.e., bd$\kappa = \{x : \|x\|_2 = 1^T x\}$.* [1]

The theorem 1 can be considered as a more natural application of the sufficiently scattered condition to co-occurrence/correlation based topic modeling, which explores the symmetry of the model and avoids normalization.

5. Anchor-Free Algorithm: Solving Non-convex problem

The identification criterion gives a non-convex optimization problem, simplex constrained least squares [1]:

$$\underset{E,C}{\text{minimize}} \qquad \|P - CEC^T\|_F^2 + \mu|\det E| \tag{9}$$

$$\text{subject to} \qquad C^T 1 = 1, C \geq 0 \tag{10}$$

Algorithm 2: AnchorFree

input : $\boldsymbol{D}, F$.
$\boldsymbol{P} \leftarrow$ `Co-Occurrence`$(\boldsymbol{D})$;
$\boldsymbol{P} = \boldsymbol{B}\boldsymbol{B}^T, \boldsymbol{M} \leftarrow \boldsymbol{I}$;
repeat
  for $f = 1, \ldots, F$ **do**
    $a_k = (-1)^{f+k} \det \bar{\boldsymbol{M}}_{k,f}, \ \forall\, k = 1, ..., F$;
    `// remove` k`-th row and` f`-th column of` $\boldsymbol{M}$ `to obtain` $\bar{\boldsymbol{M}}_{k,f}$
    $\boldsymbol{m}_{\max} = \arg\max_{\boldsymbol{x}} \boldsymbol{a}^T\boldsymbol{x}$ s.t. $\boldsymbol{B}\boldsymbol{x} \geq 0,\ \mathbf{1}^T\boldsymbol{B}\boldsymbol{x} = 1$;
    $\boldsymbol{m}_{\min} = \arg\min_{\boldsymbol{x}} \boldsymbol{a}^T\boldsymbol{x}$ s.t. $\boldsymbol{B}\boldsymbol{x} \geq 0,\ \mathbf{1}^T\boldsymbol{B}\boldsymbol{x} = 1$;
    $\boldsymbol{M}(:, f) = \arg\max_{\boldsymbol{m}_{\max}, \boldsymbol{m}_{\min}} (|\boldsymbol{a}^T\boldsymbol{m}_{\max}|, |\boldsymbol{a}^T\boldsymbol{m}_{\min}|)$;
  end
until *convergence*;
$\boldsymbol{C}_\star = \boldsymbol{B}\boldsymbol{M}$;
$\boldsymbol{E}_\star = (\boldsymbol{C}_\star^T\boldsymbol{C}_\star)^{-1}\boldsymbol{C}_\star^T\boldsymbol{P}\boldsymbol{C}_\star(\boldsymbol{C}_\star^T\boldsymbol{C}_\star)^{-1}$;
output: $\boldsymbol{C}_\star, \boldsymbol{E}_\star$

Figure 2: AnchorFree Algorithm [1]

In here, $\mu \geq 0$ works as a balancer between the data fidelity and the minimal determinant criterion.

Finding an optimal solution is very difficult, but we can find its approximation as following:
Since P is sysmmetric and positive semidefinite, we can apply square root decomposition and eigen decomposition of sparse matrix.

$$P = BB^T, \text{where } B \in \mathbb{R}^{V \times F} \tag{11}$$

$$B = CE^{1/2}Q, Q^TQ = QQ^T = I, E = E^{1/2}E^{1/2} \tag{12}$$

The respresenting coefficient of $CE^{1/2}$ in the range space of B must be orthonormal since P is symmetric. Thus, the above approximation() can be re-written:

$$\underset{E,C,Q}{\text{minimize}} \qquad |\det\ E^{1/2}Q| \tag{13}$$

$$\text{subject to} \qquad B = CE^{1/2}Q, C^T1 = 1, C \geq 0, Q^TQ = I \tag{14}$$

Let $\widetilde{E} = E^{1/2}Q$,

$$\underset{\widetilde{E},C}{\text{minimize}} \qquad |\det\ \widetilde{E}| \tag{15}$$

$$\text{subject to} \qquad B = C\widetilde{E}, C^T1 = 1, C \geq 0 \tag{16}$$

The objective problem (15) is proportional to the objective problem (6). Now, we can say that both have the same solutions. Also, Q is unitary which does not affect the determinant. So, let $M = Q^TE^{-1/2}$, then, it can be reformulated as below:

$$\underset{M}{\text{maximize}} \qquad |\det\ M| \tag{17}$$

$$\text{subject to} \qquad M^TB^T1 = 1, BM \geq 0 \tag{18}$$

However, the objective problem is still non-convex. It can be solved by a dynamical system identification problem as an alternating linear program. If all the columns of M except the fth one are fixed, $detM$

becomes a linear function with respect to $M(:, f)$:

$$\det M = \sum_{k=1}^{F} (-1)^{f+k} M(k, f) \cdot \det \bar{M}_{k,f} = a^T M(:, f) \tag{19}$$

$$\text{where } a = [a_1, ..., a_F]^T, a_k = (-1)^{f+k} \cdot \det \bar{M}_{k,f}, \forall_k = 1, ..., F \tag{20}$$

$$\bar{M}_{k,f} \text{ is a matrix obtained by removing the } k\text{th row and } f\text{th column of } M \tag{21}$$

Maximizing $|a^T x|$ subject to linear constraints is still a non-convex problem, but it can be solved by maximizing both $a^T x$ and $-a^T x$ after picking the solution with larger absolute objective. Then, repeatedly updating the columns of M produces an alternating optimization(AO) algorithm which is computationally feasible.

6. Simulating Algorithms

Six thousand ninety tweets regarding climate change have been used [3]. The data was originally collected to see the public's sentiment on climate change. The size of the data is relatively small, but the data size would be enough to achieve the purpose of this study, daily monitoring of topic changes in public. The small size dataset would be a very harsh condition for each algorithm, and we can expect the distinctive results difference between all four algorithms. Also, the anchor-free algorithm must work well in the small size data as they claimed. The data does not have any fixed label or categorization, and it would be the same circumstance of the future application by twitter data crawling. One of the interesting points of data from current social network services is hashtags. Hashtags are not included in the main context, but they reflect the content generator or user's interests. Thus, hashtags have roles not only as a keyword for search but also as the topic word itself.

For this reason, the cleaned dataset includes hashtags. Interestingly, twitter limits 280 characters per each tweet, and it causes shorten expression and abbreviation. It is hard to guess their meanings and usages. They may become noise in the dataset.

Stop words are removed by using nltk, a natural language toolkit. The standard tf-idf data as the D matrix by scikit-learn is used, and for other data pre-processing procedure, I observed the direction of the paper [2]. In order to compare the performance, LDA, NMF, and Kmeans clustering are applied using scikit-learn. However, due to the characteristics of the data, it does not have its label to evaluate its accuracy. Thus, in this experiment, only the diversity of the topic clustering would be evaluated by the qualitative test analysis. The simulation results are shown in table 1 to 4.

I let each algorithm produce ten different topics with 10 top words and its weight for the experiment purpose. The number of topics should be adjusted for practical use. LDA shows the weakest performance among the four models. You can see many overlaps on every top two words. To avoid it, those overlapped words, such as 'climat,' 'chang', 'global,' and 'warm,' need to be included in the stop words and removed for the experiment. NMF provides fewer overlaps than LDA, and topics are distinguishable. Kmeans clustering also shows diverse topics, but clustered words are less meaningful than other algorithms. Interestingly, NMF and Kmeans clustering show a hashtags cluster, topic 9 of NMF, and Topic 3 of Kmeans clustering. Single-character elements are found in NMF and Kmeans.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	climat	global	global	global	climat	global	climat	chang	climat	climat
2	chang	warm	warm	warm	chang	warm	chang	climat	global	chang
3	via	snow	scientist	climat	warm	al	make	new	chang	confer
4	believ	like	climat	chang	via	day	energi	#climate	warm	peopl
5	global	volcano	scienc	come	global	gore	wors	agenc	great	talk
6	#tcot	dc	chang	know	news	one	allergi	take	senat	world
7	warm	storm	skeptic	green	good	effect	global	obama	bill	graham
8	trial	chang	un	tell	report	earth	warm	warm	stop	fight
9	clinic	climat	debat	caus	human	watch	expert	feder	immigr	bolivia
10	live	iceland	real	com	thing	chang	public	global	tackl	bill

Table 1: LDA

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	global	climat	via	bill	new	snow	peopl	scienc	#climate	make
2	warm	chang	news	senat	agenc	gore	world	scientist	chang	wors
3	caus	fight	humanitarian	graham	feder	al	confer	debat	#global	blizzard
4	effect	energi	india	immigr	form	dc	earth	say	#eco	time
5	great	take	us	compromis	obama	#tcot	right	report	great	allergi
6	stop	legisl	polit	prepar	report	mean	bolivia	climat	#p	snowstorm
7	volcano	green	com	put	studi	storm	mother	palin	us	get
8	could	carbon	climat	limbo	york	washington	day	un	#green	c
9	believ	u	need	exit	administr	cold	indigen	snake	take	blame
10	like	nation	chang	legisl	propos	jr	summit	oil	fact	season

Table 2: NMF

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	scientist	rfk	#tcot	trial	wors	york	chang	warm	ocean	#globalwarming
2	scienc	jr	#teaparty	clinic	allergi	new	climat	global	intensifi	#saveterra
3	debat	month	#gop	collagen	season	fall	via	snow	show	fact
4	art	ago	#tlot	scream	make	citi	bill	gore	rise	tell
5	climat	mean	#ocra	cliniqu	blizzard	commun	senat	volcano	faster	get
6	warm	cold	disprov	clip	snowstorm	weather	energi	caus	water	movement
7	global	dc	#p	art	ever	check	u	believ	impact	support
8	chang	snow	#climategate	chang	time	big	new	al	carbon	next
9	say	f	warm	climat	c	scam	fight	could	nation	say
10	heat	washington	global	day	blame	go	immigr	storm	year	chang

Table 3: Kmeans

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	collagen	rfk	limbo	matter	habitat	safeti	warm	climat	climat	volcan
2	cliniqu	jr	exit	money	wildlif	silli	global	chang	chang	erupt
3	scream	month	bill	resourc	protect	mine	exit	agenc	safeti	senat
4	clinic	ago	graham	better	action	regul	limbo	new	silli	volcano
5	trial	mean	put	part	help	call	put	form	mine	iceland
6	clip	cold	climat	save	#saveterra	coal	graham	feder	senat	climat
7	art	dc	unveil	live	take	global	bill	administr	regul	chang
8	chang	snow	chang	believ	#climatechange	warm	climat	propos	via	trigger
9	climat	global	senat	global	chang	effort	chang	obama	peopl	compromis
10	keep	warm	talk	warm	climat	well	gore	studi	world	global

Table 4: Anchor-Free

The anchor-free algorithm shows distinctively diverse topics. Topics include health, politics, geology, economy, and sociology. Notably, 'collagen' and 'clinic,' and 'volcano' and 'erup' are very distinguishable words, and they are ranked at the top in each topic with higher weights. It does not have any single-character element in topics

Single-character elements are found in NMF and Kmeans, but it does not have any single-character element in topics. Singe-character elements are complicated to guess its meaning or relationship in a topic. There is a high possibility that they work as noise in the dataset.

7. Conclusion

In this project, I compared the performance of Anchor-Free algorithm with other traditional topic modeling algorithms using a relatively small size of Twitter data for a quick public opinion poll. To solve the problems of the existing anchor-word-based algorithms, identifiability, and amplifying noise, we reviewed the Anchor-Free Algorithm to solve the problem through the approximation solution approach of the non-convex problem. Also, it is confirmed that the limitation of data size is low, and it could find a solution for the problem by the implementation. Unfortunately, the accuracy could not be measured due to the nature of the data in this project. This project has been verified through text data, and I hope to conduct verification through image data from Instagram in future studies.

References

[1] Kejun Huang, Xiao Fu, and Nicholas D. Sidiropoulos. *Anchor-Free Correlated Topic Modeling: Identifiability and Algorithm.* Advances in Neural Information Processing Systems (NIPS 2016), Dec. 2016, Barcelona, Spain.

[2] Wei Xu, Xin Liu, and Yihong Gong. *Document clustering Based on Non-negative Matrix Factorization.* In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pages 267–273. ACM, 2003.

[3] Crowdflower. *Sentiment Analysis of Global Warming/Climate Change.* https://data.world/crowdflower/sentiment-of-climate-change