Document-level Language Processing Using Unsupervised Learning

Timothy Chen
Seaver College
Pepperdine University
Malibu, CA 90265
timothy.chen5@pepperdine.edu

***Abstract*—A document sorting feature for the Pepperdine research social network could benefit ease of use and functionality. There is currently no system in place that groups publications by content. We propose a representation learning approach to grouping publications, allowing similar articles to be grouped based on its content and semantics, rather than just title or field of research. This study analyzes SPECTER, a document classification metric based on citations, to develop a model that groups publications based on content.**

I. Introduction

Research at Pepperdine University has thrived over the past years. Many award winning publications are released each year. However, there is no tool for keeping track of publications that happen within Pepperdine University. Our goal is to design a social network for Pepperdine research in order to document all Pepperdine publications. This tool will be useful in keeping updated the projects of Pepperdine faculty, and further Pepperdine's quality as a research institution.

We attempt to add a new feature in this social network: sorting papers by semantics. Current methods for sorting texts include sorting by author, field of study, department, or other factors unrelated to the actual content of the papers. This grouping can be done manually using database mechanics. However, such groupings are not as efficient or informative to the actual content and information found in the publications.

Therefore, the purpose of this study is to construct a classification that groups documents based on content as well as citation graphs. This study attempts to implement a deep learning based sorting mechanism for research publications. We attempt to map an array of data that represents the semantics and citations of a paper into a two dimensional graph in order to compare publications by content. Our models use unsupervised machine learning for classification of documents in order to predict the publication similarity. With an input of a specific published research paper, our model outputs a vector that represents its content, which is then clustered using unsupervised learning to a group of similar publications.

The data is represented as a vector called the SPECTER embedding. This array contains a sequence of 738 floating points representing a publication's content. These points are characterized by the citations and their importance to the publication. When a paper is cited, SPECTER takes into account the similarity of the paper content with the cited reference to predict the citation's importance to the current publication as a whole [1].

Such a model poses many potential challenges. Certain features measured within a document may not be representative of the content of the document as a whole. Language is highly variable, and certain unforeseen phrases can skew the predictions of the model. Furthermore, information may be lost when the SPECTER embedding, which has 738 fields, is scaled down to two dimensions.

A. Related Works

Machine learning researchers have been interested in the applications of natural language processing for many years [2]. Previous studies on language processing include various practical applications such as mining for information based on text within social media platforms or gauging user sentiment based on product reviews [3]. The baseline approach to processing such information is computationally analyzing individual words to match with a set, classifying sentences based on the group that the words belong in. However, this approach includes many drawbacks because langauge can not simply be classified by words. Two sentences may have the same words but completely different meanings.

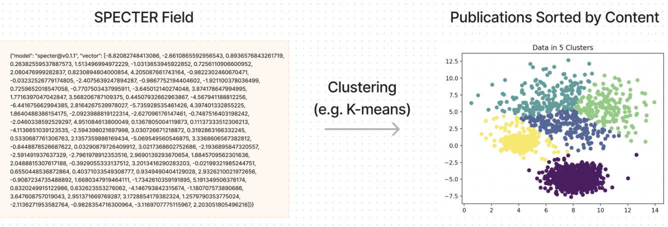
Improvements on the baseline approach includes the use of more sophisticated techniques. One model, Stanford Core NLP, utilizes grammar, entity relatedness, and sentence structure to provide a holistic view on language [4]. However, this approach still has major drawbacks due to the high variation that occurs in language. Accurate language processing requires knowledge of world events and culture that can't be computed.

Machine translation systems are also an impactful application to language processing. Modern day tools such as Google Translate are built upon statistical phrase based language processing [5]. This approach learns based on short phrases of up to three words and is a significant improvement to previous systems.

Modern day advancements in language processing include sequence to sequence learning which maps an input sequence to a vector then decodes it using long short term memory, a recurrent neural network [6]. However, most of these models are focused on token to token interpretations and have limited information on document relatedness, which would be necesary for classifying research publications.

Therefore, we analyze SPECTER, a powerful model for classifying documents based on citations graphs [1]. This model encodes the input using citation informed transformers. We take into account weighted citations based on leverage to guide the classifications. Citations provide valuable information on inter document representations. The more a cited work is referenced in a paper, the more that publication influences the classification.

II. Methods



{"model": "specter@v0.1.1", "vector": [-8.82082748413086, -2.6610865592956543, 0.8936576843261719, 0.26382559537887573, 1.513496994972229, -1.0313653945922852, 0.7256110906600952, 2.080476999282837, 0.8230894804000854, 4.205087661743164, -0.9822302460670471, -0.03232526779174805, -2.4075639247894287, -0.9867752194404602, -1.921100378036499, 0.7259652018547058, -0.7707503437995911, -3.645012140274048, 3.874178647994995, 1.7716397047042847, 3.568206787109375, 0.44507932662963867, -4.567941188812256, -6.441675662994385, 2.8164267539978027, -5.735928535461426, 4.397401332855225, 1.8640488386154175, -2.0923988819122314, -2.627096176147461, -0.7487516403198242, -2.0460338592529297, 4.951084613800049, 0.1367805004119873, 0.11137333512306213, -4.1136651039123535, -2.594398021697998, 3.030726671218872, 0.3192863166332245, 0.5330687761306763, 2.1357359886169434, -5.069549560546875, 3.3368606567382812, -0.8448878526687622, 0.03290879726409912, 3.0217368602752686, -2.1936895847320557, -2.591491937637329, -2.7961978912353516, 2.9690139293670654, 1.8845709562301636, 2.0488815307617188, -0.3929055333137512, 3.2013416290283203, -0.02199321985244751, 0.6550448536872864, 0.40371033549308777, 0.9349490404129028, 2.9326210021972656, -0.9087234735488892, 1.6680347919464111, -1.7342610359191895, 5.191349506378174, 0.8320249915122986, 0.632623553276062, -4.146793842315674, -1.180707573890686, 3.647608757019043, 2.951371669769287, 3.1728854179382324, 1.2579790353775024, -2.1136271953582764, -0.9828354716300964, -3.116970775115967, 2.203051805496216]}

Fig. 1. Illustration of the output that we hope our model can produce after learning from the training set.

This study explores possible techniques to automatically classify and sort documents based on SPECTER fields. We analyze the performance our data against several different clustering algorithms including K-means, expectation-maximization, hierarchical clustering, and spectral clustering. We also attempt to visualize the results of each clustering algorithm using a 2 dimensional graph.

One drawback this model faces is challenges in projecting SPECTER, a vector of 738 dimensions, onto a 2 dimensional graph. Therefore, we employ algorithms such as PCA or spectral clustering in order to group and visualize data points of high dimensions.

A. Dataset Acquisition

The data was collected from the Semantic Scholar database. Semantic Scholar is an online research tool that provides information about publications including their SPECTER index. The publications of each of Pepperdine's 1020 faculty are queried using the semantic scholar API. These paper IDs are then used to access the SPECTER embeddings of each paper, which we will use to build our groupings.

We utilize a variety of clustering algorithms and compare them to achieve the most accurate representation of the publication categories.

B. K-means

One widely employed approach to clustering that we analyze is the K-means algorithm. In K-means, K cluster centers are initialized randomly. The data is grouped based on which cluster center they are nearest to. This is computed using the euclidean distance.

The cluster centers are then updated using the average of all the points in its group. The distances are then computed and the data is grouped again. This is repeated until convergence, or when the cluster centers stop updating.

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}.$$

Fig. 2. Euclidean distance.

The number of cluster centers is chosen using a validation method called silhouette coefficient. We test the algorithm with different number of clusters and choose the one with the highest silhouette score. The silhouette score is computed based on the average distance between each point within a cluster and the average distance between all the clusters [7]. A higher silhouette score indicates a more defined cluster and a better fit.

The advantages of the K-means algorithm is its simplicity and guarantee of convergence. However, this model suffers in its inability to adapt to outliers as well as its dependence on random initial values.

A more pertinent issue is its failure to model data with higher dimensions effectively. Clustering based on euclidean distance converges when faced with higher dimensional data points, such as SPECTER. All the points will be grouped to one cluster and our model will fail. Our model pre processes the data points using Principal Component analysis, where distinct features of the data points are analyzed to create a two dimensional representation of a vector in relation to the rest of the vectors in the data set. This allows us to generate a visual representation of the groupings produced by K means. We also propose a pre-clustering algorithm known as spectral clustering as another approach to address this inadequacy.

C. Spectral Clustering

Spectral clustering reduces the dimensionality of the data by projecting data to a lower dimension. It achieves this by first representing data as a similarity graph. Each data point is connected to its k nearest neighbor. The Laplacian matrix for this graph is formed, and its eigenvectors and projection are used to reduce dimensionality. Then, a classic clustering algorithm is applied (e.g. K means).

One potential drawback of this algorithm is its efficiency. Finding eigenvectors is computationally expensive. This may pose an issue in very large data sets. However, spectral clustering help with the issue of the curse of dimensionality and allows for robust analysis of data with large dimensions.

We utilize spectral clustering to preprocess our data, reducing the dimension of the SPECTER fields of each publication without losing information before applying PCA to a data set of lower dimension, then grouping them into a cluster. Theoretically, less information is lost using this approach.

D. Expectation-Maximization

We also utilize the EM algorithm, which uses maximum likelihood estimation to generate clusters. Each point is assigned to a randomly initialized cluster based on probability.

The mean and variance of each cluster is updated according to the probabilities computed in the first step. This is done repeatedly until convergence.

E. Hierarchical Clustering

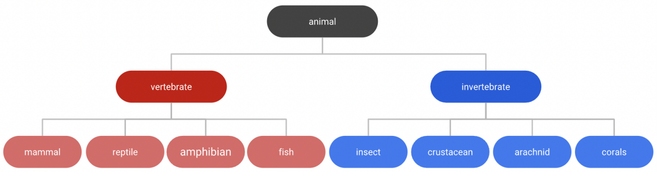


Fig. 3. An example of a cluster hierarchy.

Ideally, we want to show that publications can be related to each other despite not being of the same cluster. Some publications are less related than others. Thus, we propose a hierarchical clustering algorithm, in which our model is represented in a tree of clusters such that clusters are related through a parent node. This can be achieved by identifying similar clusters and merging them systematically from the bottom up in an algorithm known as agglomerative clustering.

With this approach, we can visually identify the potential clusters, and use the hierarchical visualization to choose the optimal number of clusters.

F. Data Visualization

The final step is to visualize the results using a two dimensional graph as a well as hierarchy diagrams. We utilize PCA, a dimension reduction algorithm to scale the data into two dimensions. PCA analyzes the eigenvectors of the covariance matrix to identify the principal components of a vector [8]. A draw back of utilizing this algorithm is the loss of some accuracy for simplicity. As the data is reduced to two dimensions, there may be some loss in information, but it enables us to visualize the clusters in a two dimensional space.

III. Results

Our research generated similar results across each clustering algorithm. Overall, the silhouette scores and cluster analysis recommended three distinct clusters across all algorithms except for spectral clustering, which had the highest silhouette score for two clusters. There were slight variations in the borders of each cluster, but the overall shape of the clusters were similar.

A. K Means

For the k means algorithm, the data was pre processed using PCA to generate two dimensional data. The k means algorithm was run with five different values for the n clusters parameter, ranging from two to six. The models seemed to recommend three distinct clusters, since it had a distinctly higher silhouette score of 0.56529. Therefore, we plot the clusters as well as the cluster centers using three clusters.

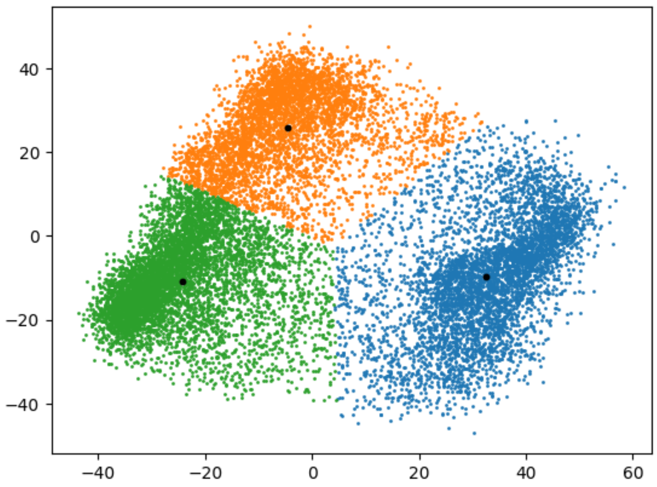


Fig. 4. Clusters generated with the k means algorithm with n clusters = 3

B. Spectral Clustering

Next, we attempt to model the data using spectral clustering and similarity graphs. This time, the silhouette scores seem to strongly favor two clusters with a score of 0.51286 for n = 2 and a significantly lower score of 0.38757 for n = 3. In fact, the scatter plot for three clusters appears to only display two clusters, indicating a failure to recognize a third cluster at all within the data.

Although in theory, spectral clustering results in dimensionality reduction, there was still too many dimensions to visualize. Furthermore, spectral clustering is the most computationally expensive and took significantly longer to run than other algorithms, especially in high dimensions. Therefore, we employ PCA again prior to the mapping of the data to a scatter plot.

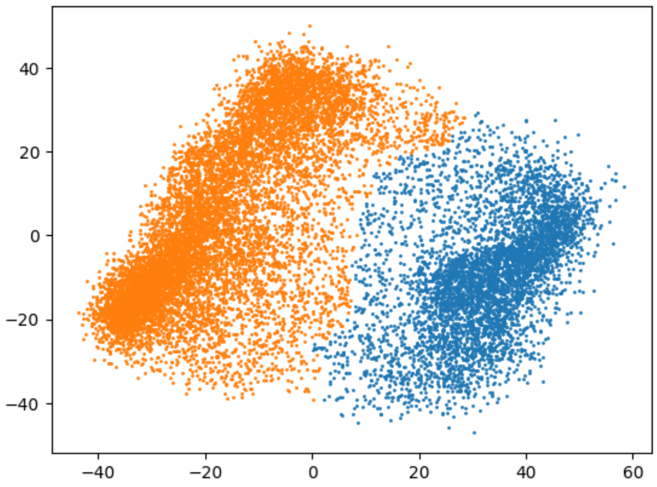


Fig. 5. Clusters generated with the k means algorithm after spectral pre clustering with n clusters = 2

Spectral clustering seems to capture the shape of each cluster quite well. The shape of the data points further away from the cluster centers seem to match with the inner shape of points near the center. The clusters produced with this methods also seem to have a smoother, more natural cut off than the clusters built using k means.

C. Gaussian Mixture Model

Next, we take a probabilistic approach to modeling the SPECTER embeddings. We propose Gaussian mixture models, which utilize the expectation maximization algorithm to fit a mixture of Gaussian models. The number of components is selected using the Bayesian information criterion. The shape of the clusters seem to match the clusters produced using k means. Although GMM seemed to be the fastest of all the algorithms, the clusters produced do not seem to match the shape of the data, producing defined, unnatural segments within the scatter plot.

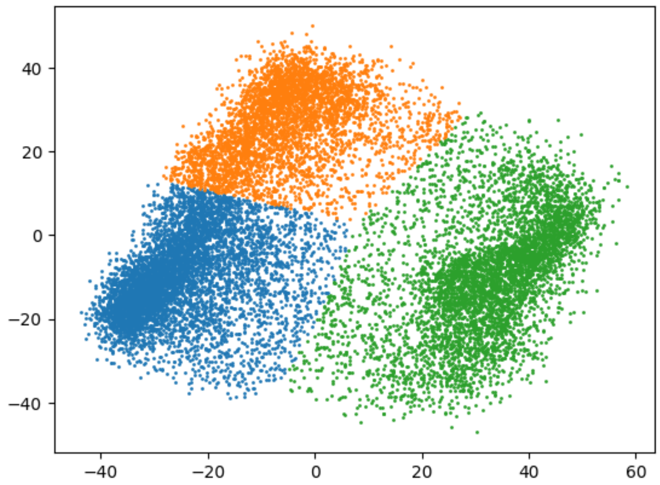


Fig. 6. Clusters generated with Gaussian Mixture Models using the expectation maximization algorithm with n components = 3.

D. Hierarchical Clustering

Finally, we implement the agglomerative approach to hierarchical clustering in order to group our data. Hierarchical clustering can provide a general visualization of the differences when a variable number of clusters are used. Figure 7, a dendrogram of our model, shows a decrease in model compatibility, represented as distance between divisions, as more clusters are used.

We utilize the dendrogram to select the optimal number of clusters to use in our data. The number of clusters with the maximum distance between the horizontal splits is used as the optimal number of clusters. In this case, three clusters indicate the best fit. This is further evidenced by the silhouette scores displayed in figure 9.

Thus, three clusters seems to be the optimal approach for hierarchical clustering. The scatter plot of the clusters is shown in figure 8. This model appears to unevenly divide the less dense data in the center of the chart. Most of the data that is in this area is assigned to clusters on the left of the chart, where as other algorithms split the data more evenly.

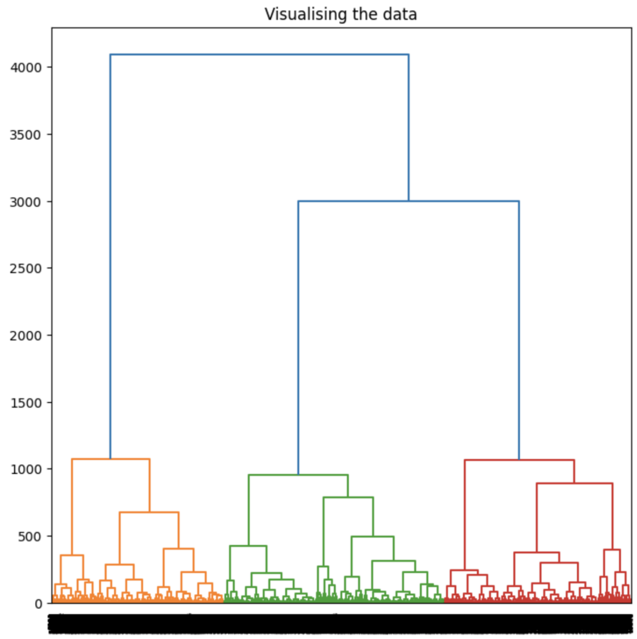


Fig. 7. A dendrogram of our model produced using hierarchical clustering.

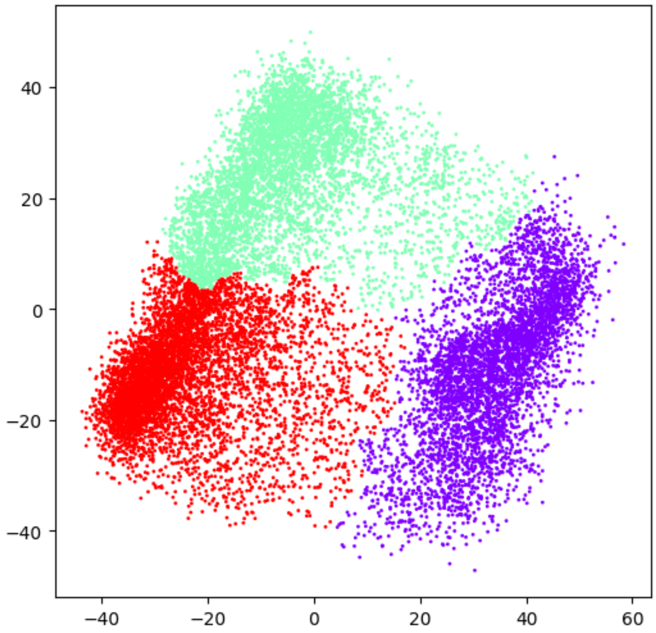


Fig. 8. Clusters generated with hierarchical clustering using three clusters.

IV. Discussion

In this study, we successfully grouped publications by their content, or SPECTER embeddings. The data set, despite consisting of all publications of all of Pepperdine University's faculty, only represents a small proportion of publications obtainable in the semantic scholars database. However, our experiments show that publications can be clustered and sorted automatically by their content. For any new publication that

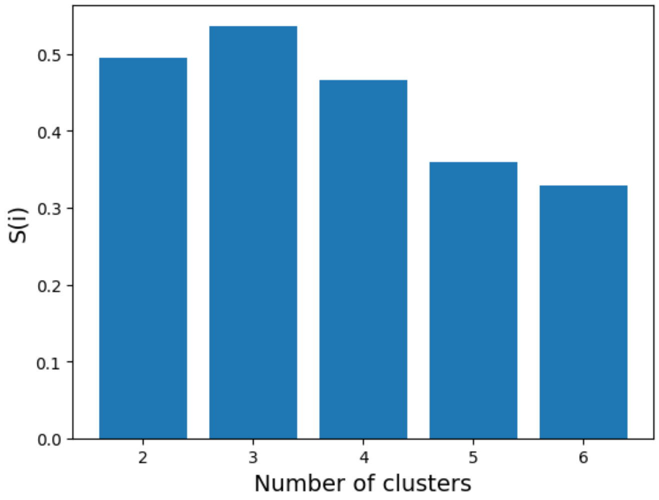


Fig. 9. A histogram of the silhouette scores for hierarchical clustering

needs to be tested on our sample, we can obtain the SPECTER embedding through semantic scholars and use it to assign the publication to a cluster using the `predict` method.

One limitation of our model is that it does not output the exact type of paper each cluster is. Since the input data consists of numeric vectors, there is no explicit definition of what type of paper each cluster represents. Our model simply predicts publications that are similar by content. For now, this feature can be implemented manually by looking at papers within the same cluster and assigning a label to describe the group. In further studies, we hope to automate this process, using language processing tools to extract the explicit category each cluster represent.

Another drawback is the issue of dimensionality. SPECTER embeddings obtained by semantic scholar consist of 738 dimensions each. Scaling it down to two dimensions most likely resulted in information loss, even if it was through well researched algorithms like PCA. Therefore, the outputs seem to be all grouped together, with only two or three distinct clusters. These clusters may represent broader categories, such as field of study, but they fail to capture more specific measures of similarity such as publications on similar topics. In reality, the models should be able to accurately group several clusters into several fields of research. Our models do achieve accurate specifications of inter document similarity, as data points that are close to each other do have similar content. However, it struggles to classify the data into many clusters since the dimension reduction made the data points too close to each other for the algorithms to accurately pinpoint more than three clusters.

V. Conclusion

Our research succeeds in showcasing the capability of unsupervised learning models to successfully classify and group semantic data at a document level. Previous studies on language processing such as sentiment analysis achieves groupings on a much smaller scale, like sentences or a short review. However, our model utilizes new metrics in language processing, specifically SPECTER, to group much larger data such as research publications. This is especially beneficial because larger documents take more time to read manually, where as a shorter review can be more easily processed by humans. Our models can be use in applications such as Pepperdine's social network for research publications in order to track and group Pepperdine University's research publications more efficiently. Users can access publications of similar content in order to streamline the document search process. New publications can be automatically grouped into the database without an administrator needing to manually enter it into a group. We hope that this new technology encourages connectedness within different departments of research institutions and ground breaking interdisciplinary research.

References

[1] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2270–2282, Online. Association for Computational Linguistics.

[2] Chowdhary, KR1442, and K. R. Chowdhary. ”Natural language processing.” Fundamentals of artificial intelligence (2020): 603-649.

[3] Hirschberg, Julia, and Christopher D. Manning. ”Advances in natural language processing.” Science 349.6245 (2015): 261-266.

[4] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, D. McClosky, “The Stanford CoreNLP Natural Language Processing Toolkit,” in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, System Demonstrations (Association for Computational Linguistics, Stroudsburg, PA, 2014), pp. 55–60.

[5] P. Koehn, F. J. Och, D. Marcu, “Statistical phrase-based translation,” in Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (Association for Computational Linguistics, Stroudsburg, PA, 2003), pp. 48–54.

[6] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. ”Sequence to sequence learning with neural networks.” Advances in neural information processing systems 27 (2014).

[7] Shahapure, Ketan Rajshekhar, and Charles Nicholas. ”Cluster quality analysis using silhouette score.” 2020 IEEE 7th international conference on data science and advanced analytics (DSAA). IEEE, 2020.

[8] Shlens, Jonathon. ”A tutorial on principal component analysis.” arXiv preprint arXiv:1404.1100 (2014).