Big Data Clustering Topics using Jensen-Shannon Divergence

INFO 590: Big Data Application and Analytics

Final Project Report

Sameer Alam(samalam@indiana.edu)
Santhosh Soundararajan(soundars@umail.iu.edu)
School of Informatics and Computing

***Abstract*— An exploratory analysis on the Darwin's research topic models with a different distance metric between the topic vectors, i.e. using *Jensen-Shannon divergence (JSD)* instead of KL divergence.**

***Index Terms*—LDA, JSD, topic modeling, clustering, distances, probability distribution.**

I. Introduction

This research looks at the work of Jaimie Murdock, Simon DeDeo and Colin Allen, titled [4] *"Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks"* at Indiana University. The research draws upon Charles Darwin's reading habit extracted based on his hand written notes, maintained by him on what he is reading on a daily basis that spans over 22 years. It finally tries to find an intellectual pattern that empowered him to produce Theory of Evolution by Natural Selection. Natural selection, in the record of science is a pioneering advance.

As a part of their research, they located 669 of his English nonfiction readings and topic modeling was used to the full-text of these readings. They then measured the space travelled by Darwin between and within books for which they applied topic model semantic space. This allowed them to study the two fundamental qualities in his reading pattern: (1) local exploitation by hopping from topic to topic, (2) distant exploration that led Darwin towards the publication of "*The Origin of Species"*

A balancing of factors all of which are not attainable at the same time comes to play when concerning distant exploration and local exploitation. This happens because searching for uncertain distributed resources in a dynamic environment where reader or knowledge seeker switches amid in-depth reading and new domain studying. We examine the choices made by Charles Darwin, who is one of the extremely distinguished modern era scientist. Charles Darwin developed the theory of natural selection partly by combining and creating unrelated functions, fractions and elements of Victorian science. Analyzing Darwin's self-documentation of his readings, shifts are seen, on several time periods, among deciding to stay in the current familiar discipline or to switch and look for amazements in novel domains. Considering the lengthiest time-periods, the switches link to major scholarly times of Darwin's career, this was revealed by using Bayesian epoch assessment. When Darwin's reading pattern was compared with the same text's sequence of publication, he was found out "*more adventurous than the culture as a whole*"[1]

So this project report aims at perform a clustering on the various topic models (k=20,40,60,80) based on the ***Jensen-Shannon divergence*** which is considered superior for a few reasons: (1) its a pure distance measure, (2) it is symmetric, (3) the square root is even a metric. So before we begin to explore this, lets look the underlying concepts of LDA and see how the topic data is prepared and also the coming sections will describe what methodology is used in the machine learning and how the probability distribution function is matched.

II. Human Perspective

We would like to discus how a model is built by keeping a human being in the place of the model and how he perceives information. Hence, we assume this human is an average educated adult and lets say if he/she looks at one of the titles from the corpus *[John Saunders. 1809. "The art of improving the breeds of domestic animals"],* he/she might readily understand what this text is about looking at the title or the topic, and may be he can study this further and begin to answer questions like: What this is about?, What is it related to?, What does it feel like? Etc.

1 Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks, September 9, 2015.

He can do this quiet effortlessly because he has previous knowledge about the world, and he is also good at understanding the change in context on the fly with the right level of abstraction so that the same word occurring at a different context can have a completely different meaning altogether. It applies to a completely objective concept like color, which can change the meaning in any number of ways. So a human understands text with 3 components – Word, Topic and Context. So to build a model, one needs to train the model with these 3 components.

III. Machine Learning and Topic Modeling

In any modeling process, one needs to carry out three major steps: (1) the data processing, (2) training the model, and finally (3) score it on a new document. This section will deal with explaining how these processes are done in the Darwin research.

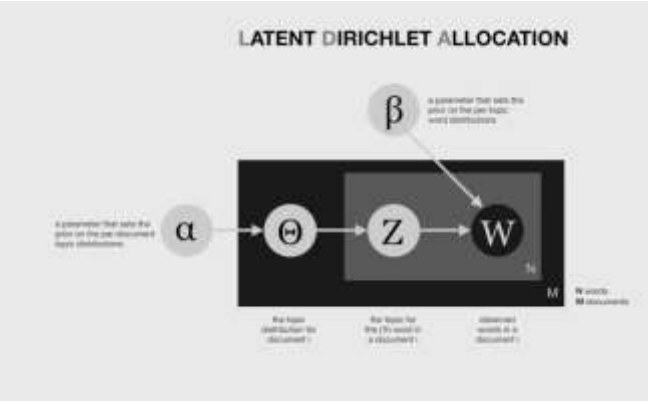
A. The data processing

As mentioned, among all his hand written notes, the two reading notebooks ([1, 2] – transcribed by [3]) contain 1,248 titles that were identified by the Darwin Correspondence Project [3], out of these1248 titles, 915 were checked "read". They reduced the list of 915 readings to 688. They selected non- fiction titles and which were in English. They located 628 of the 688 titles at a digital library named '*HathiTrust*', also found 41more at archives on internet and collected a total of 669 titles.

So at this point we would like to thank Mr. Jaimie Murdock and his team for making this data freely available, which is in the .csv format of 600 odd titles along with various versions based on topics (k=20, 40, 60, 80) according to the need.

B. Training the Model

This is where LDA comes into picture, which elegantly combine the Topic, Context and Words. And apart from these, the LDA also involves hyper parameters called **α** and **β.** This is clearly laid out in the below given demographic.



C. Score the Model on a New Document/Title

As of now the existing work follows KL divergence to estimate the difference in probability for any new document or title (p vector) by scoring it on the model value (q vector).

This is the equation for KL divergence: D_{KL}

$$D_{KL}(\overline{p}, \overline{q}) = \sum_{i=1}^{k} p_i \ log_2 \frac{p_i}{q_i}$$

Which is a directed divergence measure and we feel the reason for one to avoid KL divergence is because when Q is used to approximate P, information is lost.

IV. Using Jensen-Shannon Divergence

The big question one might come up with is, how can you simply replace a ***Jensen-Shannon divergence*** metric in place of KL divergence metric.

The answer lies in KL's utility and efficacy, that it is, basically, KL is the basic building element of JS divergence (the symmetric JS divergence is on two distributions and where weights are 1/2 each). Also, if for some reason we are able to calculate only one of the divergence measures then still JS can be calculated from KL whereas vice-versa is not possible i.e. KL can't be calculated from JS. This will be clear once we examine the definition of JSD.

The Jensen-Shannon distance D(a,b) between samples **a** and **b** is defined as:

$$D(a, b) = \sqrt{JSD(Pa, Pb)}$$

Where P_a and P_b are the abundance distributions of samples **a** and **b** and JSD(p,q) is the *Jensen-Shannon divergence* between two probability distributions **P** and **Q** defined as :

$$JSD(p, q) = \frac{1}{2} D_{KL}(p, m) + \frac{1}{2} D_{KL}(q, m)$$

this is symmetric function.

V. Clustering Algorithm (PAM)

For clustering the topics, the clustering algorithm PAM (Partitioning Around Medoids) is used. Partitioning Around Medoids is a very commonly used k-medoid clustering algorithm. Derived although from k-means but finds edge over k-means as it also has the improvement benefit of supporting measurement of arbitrary distance. Given the benefit, it is considered robust as compared. It is an assisted

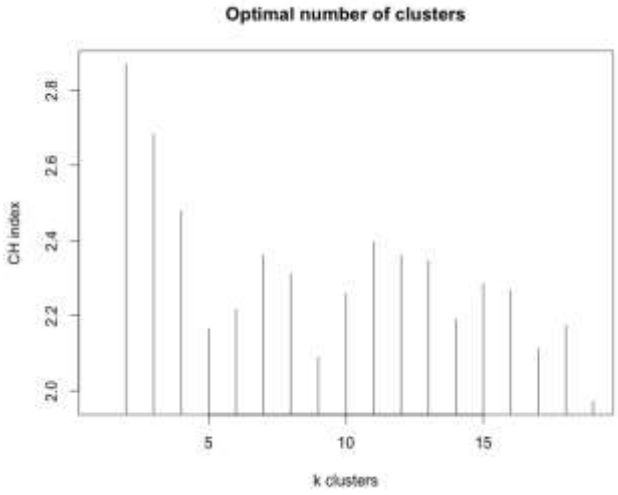
1 Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks, September 9, 2015.

(superwised) algorithm, to input this algorithm a predecided cluster amount is sent. The algorithm makes partition of the data to that number of clusters exact.

We use R to achieve clustering, R has a predefined function pam() in the library 'cluster':

```
data.cluster=pam.clustering(data.dist, k=3)
```

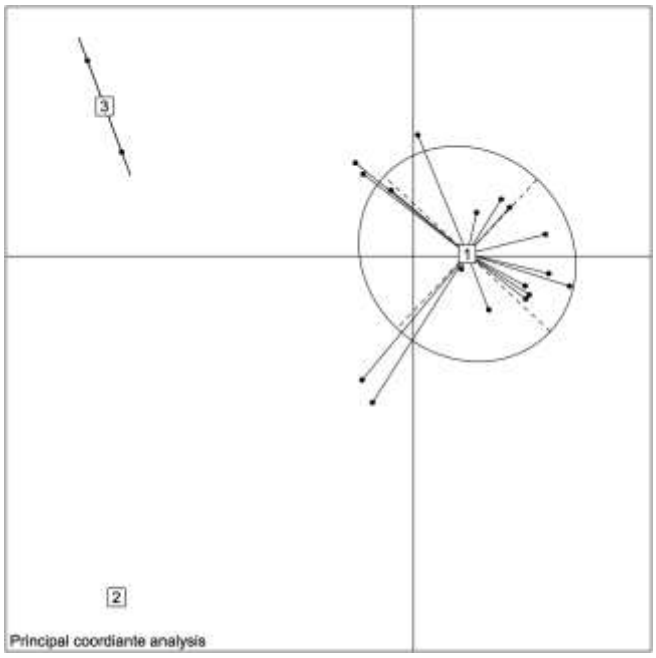
*Calinski-Harabasz (**CH**) Index* is employed next to establish k value to be 3. For recovering number of clusters, *Calinski-Harabasz (**CH**) Index* delivers good performance and hence was chosen.



Showing optimal clusters number for this dataset = 3(k is 3)

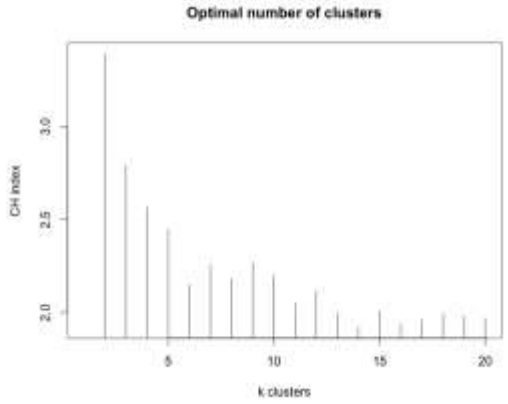
VI. Principal Coordinates Analysis (PCoA)

We chose to use the Partitioning around Medoids (PAM) clustering algorithm to cluster the topics. Partitioning Around Medoids is a very commonly used k-medoid clustering algorithm derived from k-means For **PCoA,** is performed through **R** using the package **ade4.** The s.class() function was used to plot the result achieved from PCoA which is shown in the plot.

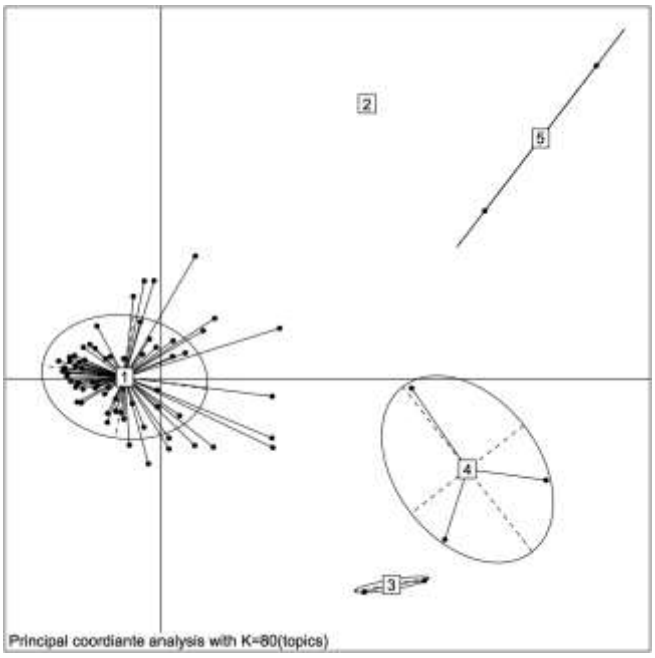


This plot shows the cluster formation within the topics K=20, and the clustering algorithm allocates Cluster-labels (1, 2, 3) arbitrarily.

Now we can similarly plot the K=80 topic distribution with probably more clusters, may be 5 with cluster labels (1,2,3,4,5) after checking the CH index for the 80 topic model as well.



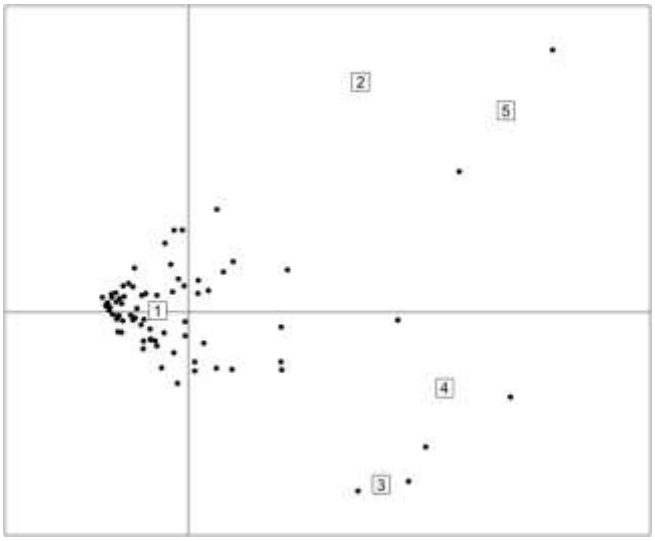
And we can see that, a 5 cluster would depict the k=80 topic model better due to higher number of data points.



The above plot, without ellipses, represents the five clusters.

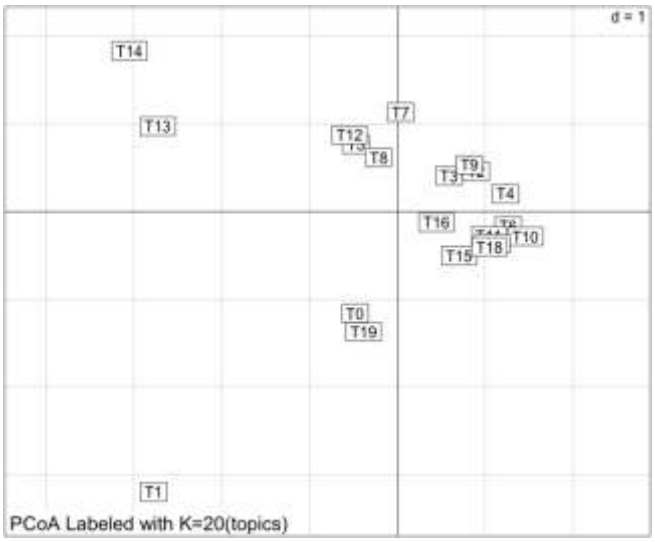
R command used for the 80 topic model is as follows.

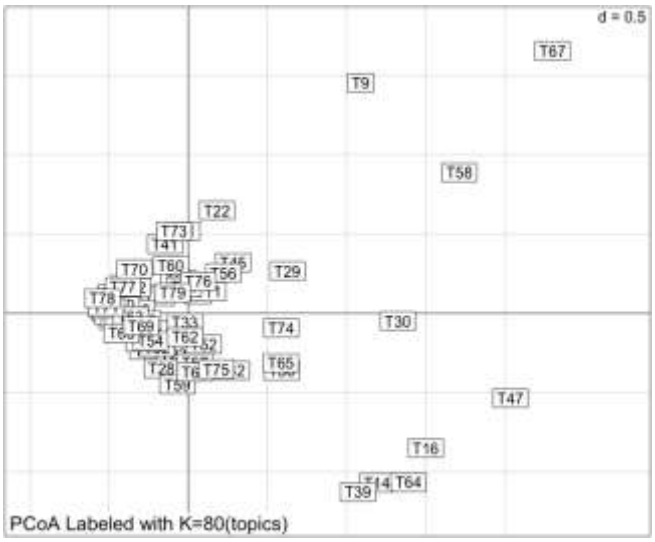
```
s.class(obs.pcoa$li,   fac=as.factor(data.cluster),
grid=F, cell=0, cstar=0)
```



VII. PCoA Clusters Labeled

Adding labels for the samples will make it very interesting when compared to the scatter plots. This can be achieved using an **s.label** function in the same package.





VIII. Conclusion

These JSD based topic plots provide an interesting platform for us to explore into the Darwin's reading patterns. And with further involvement to this project in the future, I am sure we can contribute something worthwhile to this exciting research area and we once again like to thank every one who is responsible for making these datasets freely available for our study.

TOOLS USED

1. R- R is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing.
2. Packages required:
   - install.packages("cluster")
   - install.packages("clusterSim")
   - library(cluster)
   - install.packages("rgl")
   - library(rgl)
   - library(clusterSim)

REFERENCES

List and number all bibliographical references is given belows.

[1] CharlesDarwin. *'Bookstoberead' and 'BooksRead' notebook*.1838-1851.CUL-DAR119.- Transcribed by Kees Rookmaaker.

[2] CharlesDarwin. *'Bookstoberead' and 'BooksRead' notebook*.1852-1860.CUL-DAR128.- Transcribed by Kees Rookmaaker.

[3] Peter J. Vorzimmer. The Darwin reading notebooks (18381860). *Journal of the History of Biology*, 10(1):107–153, 1977.

[4] Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks, September 9, 2015.