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, one of the most revolutionary breakthroughs in the history of science.

As a part of their research, they located 669 of his English nonfiction readings and applied topic modeling to the full-text of these readings. They then used the semantic space of the topic models in a novel way to measure the distances that Darwin traveled between books. 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 trade-off between local exploitation and distant exploration comes into play when searching for uncertain distributed resources in a dynamic environment where the reader or knowledge seeker shifts between reading in-depth and studying new domains. We examine the choices made by one of the most celebrated scientists of the modern era: Charles Darwin. Darwin built his theory of natural selection in part by synthesizing disparate parts of Victorian science. When we analyze Darwin's extensively self-documented reading, we find shifts, on multiple timescales, between choosing to remain with familiar topics and seeking cognitive surprise in novel fields. On the longest timescales, these shifts correlate with major intellectual epochs of his career, as detected by Bayesian epoch estimation. When we compare Darwin's reading path with publication order of the same texts, we find Darwin 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.

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

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

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.

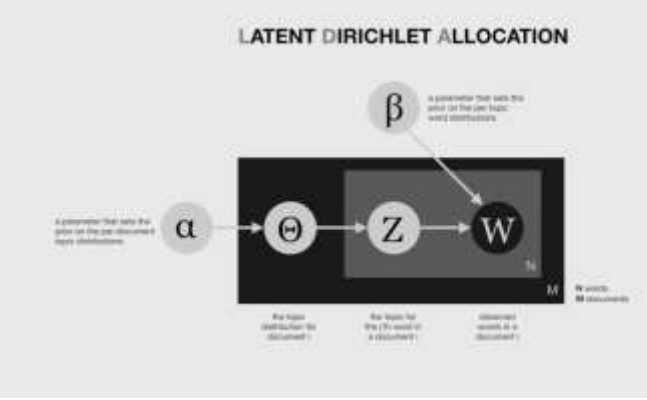
A. The data processing

As mentioned, among all his hand written notes, the two reading notebooks ([1, 2] – transcribed by [3]) contain 1,248 titles identified by the Darwin Correspondence Project [3], of which 915 were marked "read". They reduced this list of 915 to the 688 English-language non- fiction titles, and located 628 of these within the HathiTrust Digital Library and an additional 41 at the Internet Archive for 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 the information lost when Q is used to approximate P.

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 the usefulness of KL, that it is, simply, it is the fundamental building block of the JS divergences (the JS divergence that is symmetric is on two distributions, and the weights are each 1/2). If you were only able to calculate one divergence measure (for whatever reason), you could calculate JS from KL but not the reverse. 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.

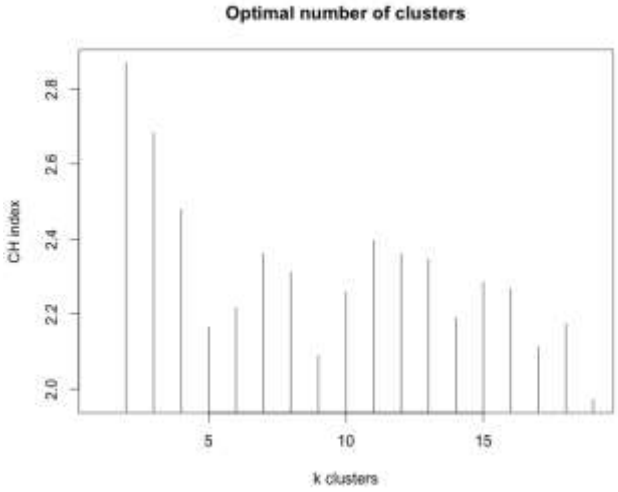
V. Clustering Algorithm (PAM)

We chose to use the Partitioning around medoids (PAM) clustering algorithm to cluster the topics. PAM derives from the basic k-means algorithm, but has the advantage that it supports any arbitrary distance measure and is more robust than k-means. It is a supervised procedure, where the predetermined number of clusters is given as input to the procedure, which then partitions the data into exactly that many clusters.

In R, we used the **pam()** function in **cluster** library to achieve the clustering :

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

Now to determine the k value as 3, we used the *Calinski-Harabasz (**CH**) Index* that shows good performance in recovering the number of clusters

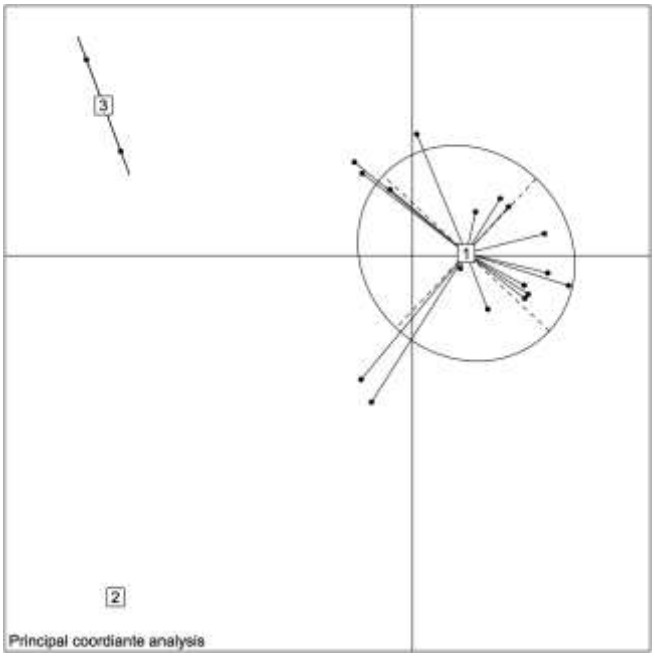


This has shown that the optimal number of clusters for this particular dataset is 3 (**k=3**).

VI. Principal Coordinates Analysis (PCoA)

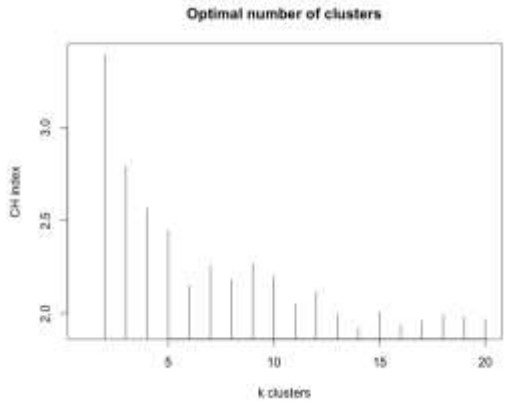
We chose to use the Partitioning around medoids (PAM) clustering algorithm to cluster the topics. PAM derives from the basic k-means algorithm.

PCoA is done using **R** with the **ade4** package and we can perform the analysis and plot the result using the **s.class()** function to get the following plot.

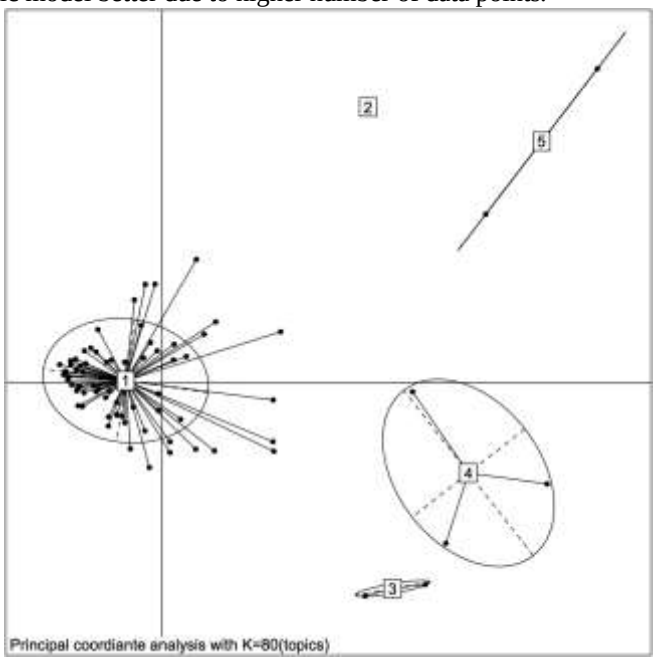


This plot shows the cluster formation within the topics K=20, and the Cluster labels (1,2,3) are randomly assigned by the clustering procedure.

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.

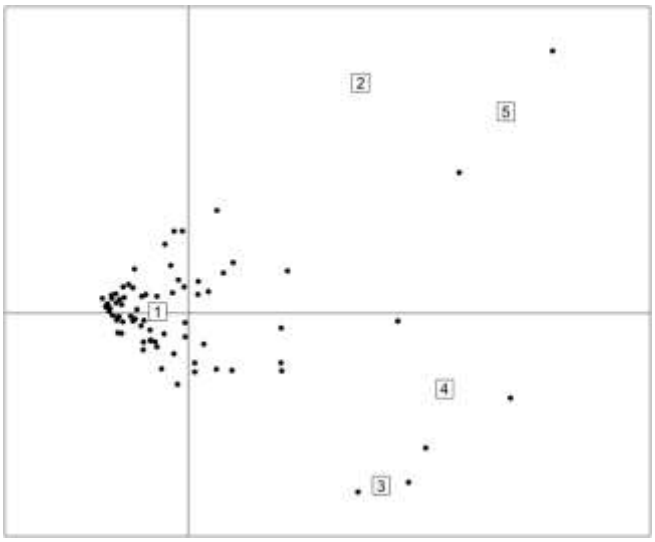


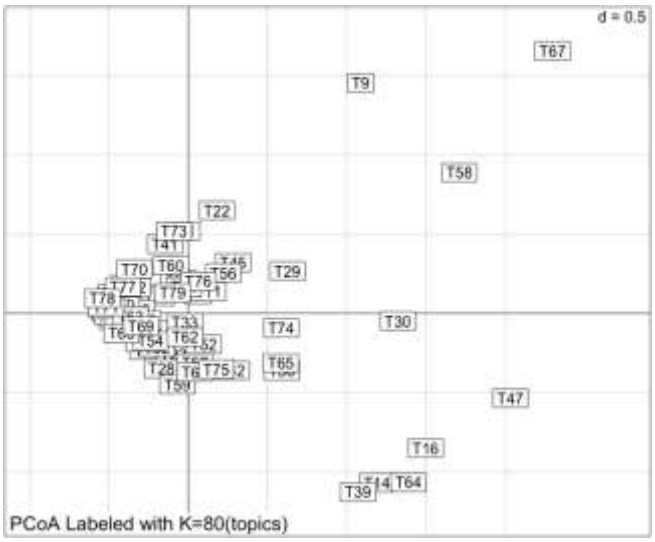
And we can see that, a 5 cluster would depict the k=80 topic model better due to higher number of data points.



Here is the plot above without the ellipses representing the five clusters using the following r command for the 80 topic model.

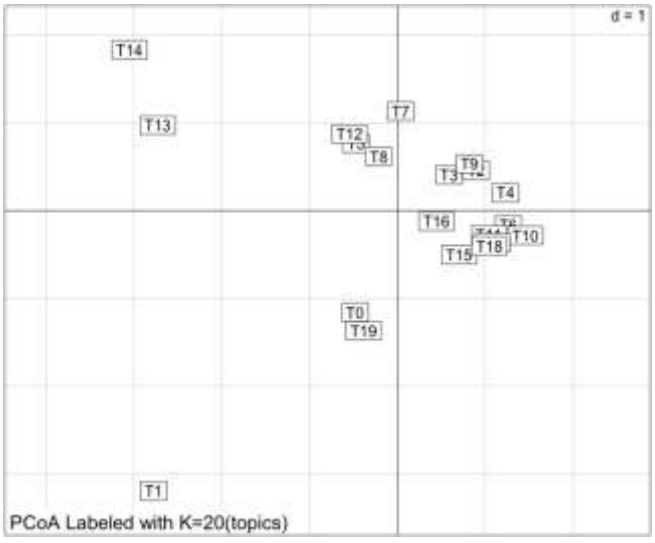
```
s.class(obs.pcoa$li,   fac=as.factor(data.cluster),
grid=F, cell=0, cstar=0)
```





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.