Report on science and technology in Space Science

By: Syed Norman Daniel Bin Syed Mahadir

No matrices :P125754

Space science, or science done from vehicles that travel into Earth's upper atmosphere or beyond, encompasses a wide range of fields, including meteorology and geology, lunar, solar, and planetary science, astronomy and astrophysics, and life sciences. Researchers and scientists hope that one day space science and exploration through space will benefit us for the upcoming future. This dataset contains four types of topics which are black holes, JWST known as the James Webb Space Telescope, planets, and asteroids. Black hole, cosmic body of extremely intense gravity from which nothing, not even light, can escape.James Webb Telescope (JWST) is the newest technology from NASA that can capture scenery from a distance with high resolution. A planet is a celestial body that is in orbit around the sun and has sufficient mass for itself-gravity to overcome a rigid body so that it assumes a hydrostatic equilibrium shape. Lastly, Asteroids are stony bodies that circle the Sun and although asteroids circle the Sun in the same way as planets do, they are far smaller.

The visualization graph below shows four selected topics based on the science articles about black holes, JWST, planets, and asteroids. However, there is something interesting about those four topics, the appears word for topic 1 which has the highest term “impact” and “asteroid”. The word “impact” represents the collision between two or more things that can cause destruction or damage and the word “asteroid” is also related to topic 1. Both words are correlated with each other because the word “asteroid” itself has always give “impact” which mean “asteroid” are dangerous and would cause disaster to humankind. Next, the most common words in topic 2 are “planet” and “asteroid”, the reason why the common words “planet” and “asteroid” are high in the topic is that the article maybe explains life before or after an asteroid hits “earth” during Paleozoic and Mesozoic which around 66 million ago. Besides that, topic 3 shows the words “stars” and “telescope” It might represent space exploration since there are other words like “univers” and “image” which James Webb Space Telescope has managed to capture “image” name as Carina Nebula on 12 July 2022 from NASA which has brought us a new era of astronomy. Finally, the last topic which has high terms is the words “black” and “hole” which represent the dying stars known as a black hole. The last topic also appears as “star” and “supermass” which the word “supermass” itself is represent how big the black holes were in this article.

Figure 2 shows the beta of topic 1 and topic 2 which we can see that the different Beta in topic 1 has less spread than topic 2 due the dominance of topic 2 in topic 1. The log ratio in "asteroid" between Topic 1 and Topic 2 has a very small spread means that both has same ratio beta. However, the words "sun" and "impact" both have different high ratio for topic 1 and topic 2.

Moving on figure 3 and 4, based on the graph with different algorithms, we can see that the result of the hierarchical clustering graph has the best result compared to Density-Based Spatial Clustering since the DBSCAN can only detect one cluster area only . The DBSCAN shows poor results and the Hierarchical clustering can perform well with small data sets due to advantages of clustering structure through dendrogram. The K-means result shows quite promising results but compared to hierarchical clustering, the graph is much better because it can compare the Topic with k=4 meanwhile K-means can detect four clusters but one of them detects as outlier.

Finally, space science and technology play critical roles in improving our understanding of the cosmos and propelling breakthroughs in a variety of sectors. Black holes, the James Webb Space Telescope (JWST), planets, and asteroids are all intriguing topics that provide light on cosmic events and our study of space. We have gotten a better grasp of these issues and their linkages using word analysis and clustering methods. Moving forward, continuing research and use of space science and technology will likely provide new discoveries and advances, altering our view of the universe and helping mankind in a variety of ways.

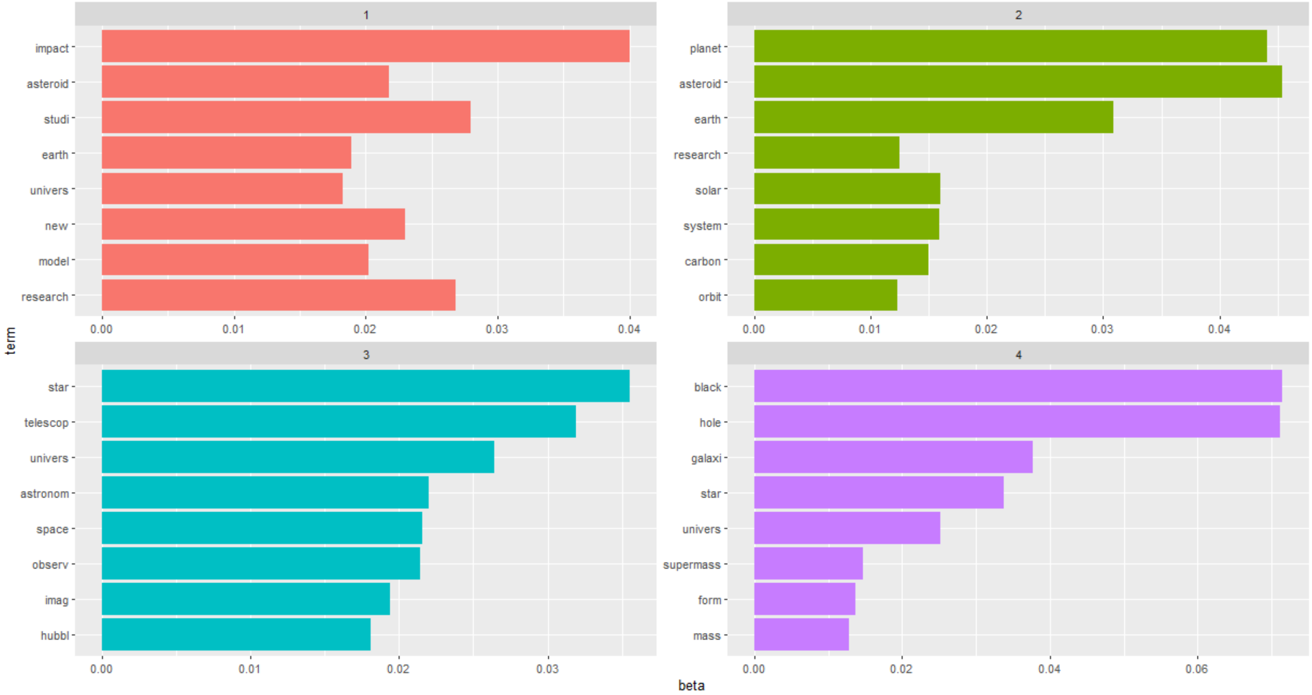


Figure 1 shows the most top 8 common words in four topics

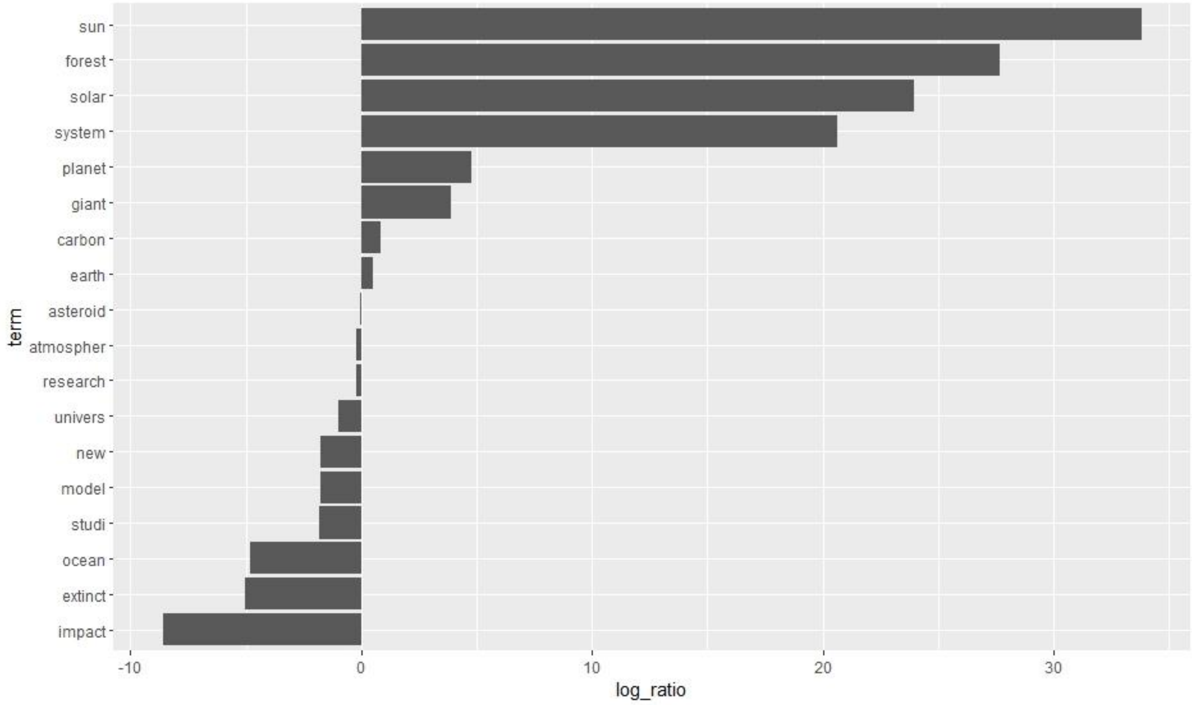


Figure 2 shows the difference in Beta between topic 1 and topic 2

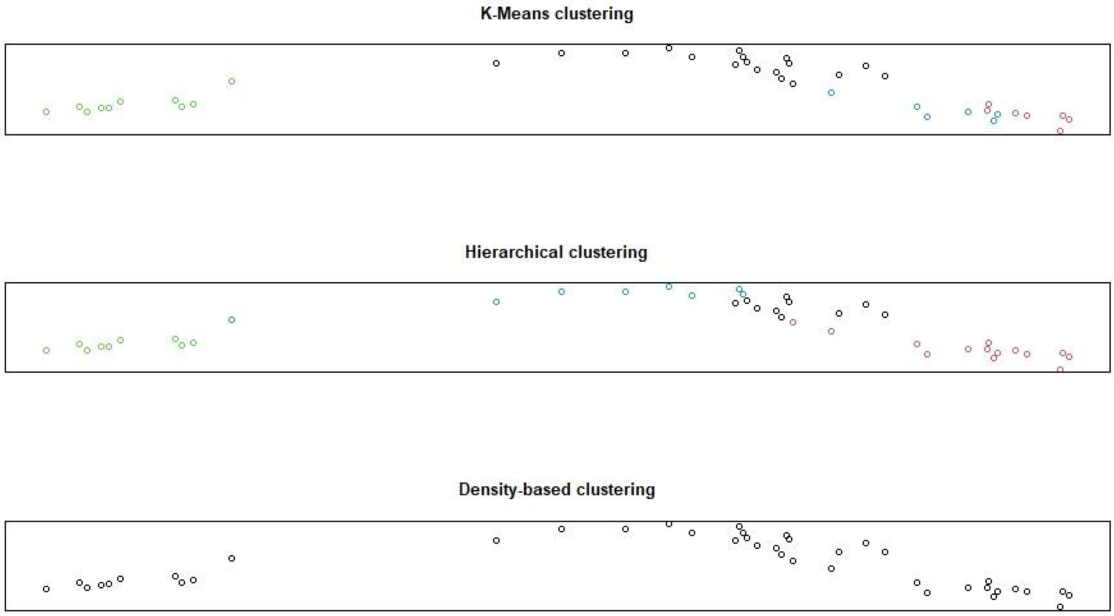
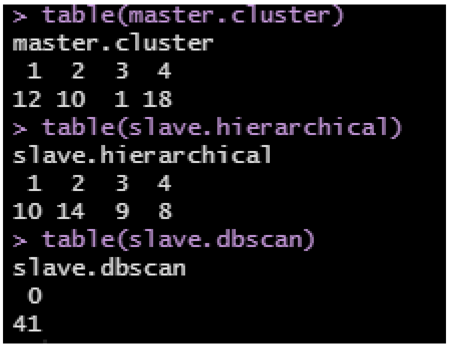


Figure 3 shows three different types of clustering algorithms graph, K-Means clustering, Hierarchical clustering and Density based clustering



```
> table(master.cluster)
master.cluster
 1  2  3  4 
12 10  1 18 
> table(slave.hierarchical)
slave.hierarchical
 1  2  3  4 
10 14  9  8 
> table(slave.dbscan)
slave.dbscan
 0 
41 
```

Figure 4 shows the table cluster

APPENDIX

```
### LDA
library(reshape2)
library(tm)
library(SnowballC)
library(RColorBrewer)
library(syuzhet)
library(ggplot2)
library(tidytext)
library(topicmodels)
library(tidyr)
library(dplyr)
library(tibble)
library(ggthemes)
library(wordcloud)
#import text data to r
document<-Corpus(DirSource("C:/Users/LENOVO/Desktop/textmining"))
#Data cleaning process
toSpace<-content_transformer(function(x,pattern)gsub(pattern,"",x))
document<-tm_map(document,toSpace,":")
document<-tm_map(document,toSpace,",")
document<-tm_map(document,toSpace,""")
document<-tm_map(document,toSpace,"?")
document<-tm_map(document,content_transformer(tolower))
document<-tm_map(document, removeWords,stopwords("english"))
document<-tm_map(document,stripWhitespace)
document<-tm_map(document,removePunctuation)
document<-tm_map(document,removeNumbers)


myStopwords <- c("can", "say","one","way","use",
          "also","howev","tell","will",
          "much","need","take","tend","even",
          "like","particular","rather","said",
            "get","well","make","ask","come","end",
            "first","two","help","often","may",
            "might","see","someth","thing","point",
            "post","look","right","now","think","'ve ",
            "re ")
document<-tm_map(document,removeWords,myStopwords)
inspect(document)
#word Stemming process
document2<-tm_map(document,stemDocument)
#term matrix
minimumFrequency <- 5
DTM <- DocumentTermMatrix(document2, control = list(bounds = list(global =
c(minimumFrequency, Inf))))
# have a look at the number of documents and terms in the matrix
dim(DTM)

#create 4 topic and seed(1234)
ap_lda<-LDA(DTM,k=4,control=list(seed=1234))
ap_topics<-tidy(ap_lda,matrix="beta")
#Find terms that are most common within each topics
ap_top_terms <- ap_topics %>% group_by(topic) %>% top_n(8,beta) %>% ungroup () %>%
arrange (topic, -beta)
ap_top_terms%>% mutate(term=reorder(term,beta))%>%
  ggplot(aes(term,beta,fill=factor(topic)))+geom_col(show.legend=FALSE)+
  facet_wrap(~topic,scales="free")+coord_flip() #visualize the above

beta_spread1<- ap_topics %>% mutate (topic=paste0("topic",topic)) %>% spread(topic,beta)
%>%
  filter (topic1>0.010 | topic2 > 0.010) %>% mutate(log_ratio = log2(topic2/topic1))
beta_spread1%>% mutate(term=reorder(term,log_ratio))%>%
  ggplot(aes(term,log_ratio))+geom_col(show.legend=FALSE)+coord_flip()

#Textclustering
#clustering analysis
```

```
tdm.tfidf <- weightTfIdf(DTM)
tdm.tfidf <- removeSparseTerms(tdm.tfidf, 0.999)
tfidf.matrix <- as.matrix(tdm.tfidf)

# Cosine distance matrix (useful for specific clustering algorithms)
library(proxy)
dist.matrix <- dist(tfidf.matrix, method = "cosine")

truth.K=4
#Perform clustering
library(dbscan)
clustering.kmeans <- kmeans(tfidf.matrix, truth.K)
clustering.hierarchical <- hclust(dist.matrix, method = "ward.D2")
clustering.dbscan <- hdbscan(dist.matrix, minPts = 10)

#Combine results
master.cluster <- clustering.kmeans$cluster
slave.hierarchical <- cutree(clustering.hierarchical, k = truth.K)
slave.dbscan <- clustering.dbscan$cluster
#plotting combined results
library(colorspace)
points <- cmdscale(dist.matrix, k = 4)
palette <- diverge_hcl(truth.K) # Creating a color palette, need library(colorspace)
layout(matrix(1:3,ncol=1))

plot(points, main = 'K-Means clustering', col = as.factor(master.cluster),
    mai = c(0, 0, 0, 0), mar = c(0, 0, 0, 0),
    xaxt = 'n', yaxt = 'n', xlab = '', ylab = '')
plot(points, main = 'Hierarchical clustering', col = as.factor(slave.hierarchical),
    mai = c(0, 0, 0, 0), mar = c(0, 0, 0, 0),
    xaxt = 'n', yaxt = 'n', xlab = '', ylab = '')
plot(points, main = 'Density-based clustering', col = as.factor(slave.dbscan),
    mai = c(0, 0, 0, 0), mar = c(0, 0, 0, 0),
    xaxt = 'n', yaxt = 'n', xlab = '', ylab = '')
```

```
table(master.cluster)
table(slave.hierarchical)
table(slave.dbscan)
```