Stack Exchange Analysis Framework:

Cloud Based Solutions.

CMIS 4144

Distributed & Cloud Computing

Supervisor

Mr. J.R.K.C. Jayakody

Submitted By

142027

142134

142190

Department of Computing & Information Systems

Faculty of Applied Sciences

Wayamba University of Sri Lanka

9th October 2018

Contents

1. Introduction

This project focuses on the analysis of titles of the questions from Stack Overflow community. The titles are analyzed and judged on the format and also on the sentiment. By this project we tried to find a solution to the endless debate about the format of the title and also checks and determines whether the sentiment of the title plays a key role in the reach and effectiveness of the question. The Stack Overflow community is the largest community among the major Question-and-answer websites and comprises of wide range of "topics" or "tags" related to programming and development such as Java, Python, JavaScript, Android Development, iOS Development, R and many more. Answers are received in 10-15 minutes for questions which fall under the above mentioned "tags" or popular "tags" the above-mentioned advantages might be the sole reason for the incredible popularity of Stack Overflow, but there are downsides to it as well.

In this project, mainly we intend to analyze the activity of the community regarding Data Science category. Furthermore, we are going to do the analysis according to the following topics.

1. No of Posts vs Time
2. No of Posts for selected language vs Time
3. Language wise post trends over time
4. Posts by language
5. Language wise average time to get answer
6. Word cloud
7. Summary

Cloud database means a database that is accessible to clients from the cloud and delivered to users on demand via the internet from the provider's servers. Cloud database can offer significant advantages over their traditional counterparts, including increased accessibility, automatic failover and fast automated recovery from failures, automated on the go scaling and potentially better performance.

2. Related Works

In addition to the study by Treude et al. [2] mentioned above, other works also studied how developers use Q&A sites. Barua et al. [5] proposed a semi-automatic approach to study general topics discussed on Stack Overflow and their trends, and found that web and mobile development are the most popular topics. Bajaj et al. [6] used Stack Overflow data to analyze common challenges and misconceptions among web developers. Rosen and Shihab [4] used Stack Overflow to determine what mobile developers on Stack Overflow ask about. Other researchers have performed studies that examine how Stack Overflow affects developers' activities during software development. For example, Vasilescu et al. [7] analyzed the effect of Stack Overflow activities on the software development process. They established associations between GitHub and Stack Overflow users, and found a correlation between participants' activities in the two platforms. Zagal-sky et al. [8] also investigated the use of Stack Overflow and mailing lists as communication channels for the R project. They found that both resources provide active communication channels where participants are willing to help others. They also observed that Stack Overflow resorts to a crowd-based knowledge construction approach, where participants contribute knowledge independently, whereas for mailing list the focus is on improving specific answers.

In many ways, our work shares similar goals as these prior studies, i.e., to determine what developers use the crowd for during software development. However, our study differs in that we only consider explicit links between Stack Overflow posts and source code commits. Moreover, we use characteristics derived from these posts and commits to understand what knowledge is most helpful and what knowledge is most time consuming to attain.

3. Methodology

We are getting data from stack exchange dump files, which is in XML format. We intend to convert XML data files into CSV files. (https://datascience.stackexchange.com).

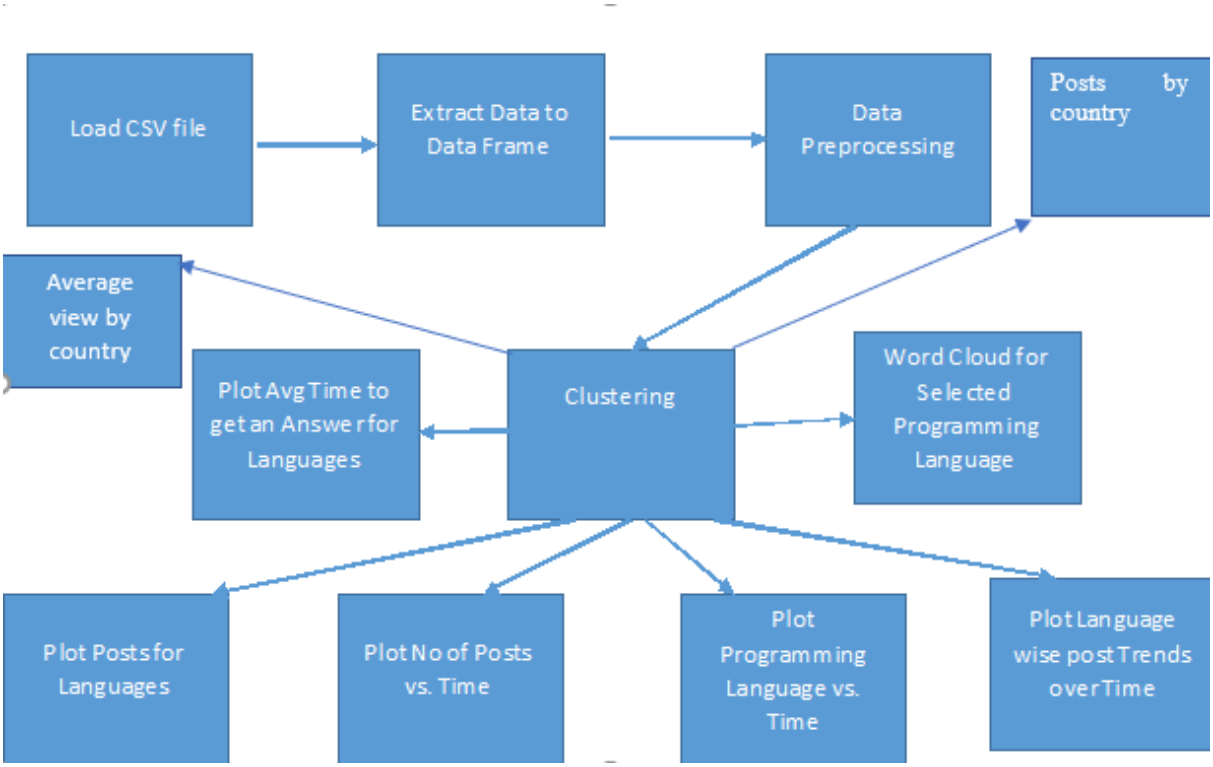


Figure 1: System Architecture

XML to CSV conversion-

```
# utility function to load XML data
loadXMLToDataFrame<- function(xmlFilePath){
  doc <- xmlParse(xmlFilePath)
  xmlList<- xmlToList(doc)

  total<-length(xmlList)

  data<-data.frame()

  for(i in 1: total){
    data <- rbind.fill(data,as.data.frame(as.list( xmlList[[i]])))
  }
  return(data)
}
```

```
PostsDF <- loadXMLToDataFrame(paste0(path,"Posts.xml"))
write.csv(PostsDF, "Posts.csv", row.names=F)
```

Extract location points-

```
coords2country = function(points)
{
  countriesSP <- getMap(resolution='low')
  filteredPostUserDf<-cbind(filteredPostUserDf, points)
  # converting points to a SpatialPoints object
  # setting CRS directly to that from rworldmap
  pointsSP = SpatialPoints(points, proj4string=CRS(proj4string(countriesSP)))


  # use 'over' to get indices of the Polygons object containing each point
  indices = over(pointsSP, countriesSP)

  as.character(indices$ADMIN)  #returns country name
}

# mapping function to set country for given location
postLocation <- function(locationName){
  if(!is.na(locationName)){
    tryCatch(coords2country(geocode(locationName, output = "latlon", source = "dsk")),
             warning = function(w) {

               print("warning");
               # handle warning here
             },
             error = function(e) {
               print("error");
               # handle error here
             })
  }
}
```

Deploy the app in shinyapps.io,

- Create an account in shinyapps.io.
- Get the API key.
- In R console, Run following commands

```
rsconnect::setAccountInfo(name='stackexchangeanalysis',token='5142097A35A2A51C18F38AE
8D223D3A8', secret='<SECRET>')

library(rsconnect)
```

deployApp()

https://stackexchangeanalysis.shinyapps.io/stack_exchange_analysiz/

4. Results

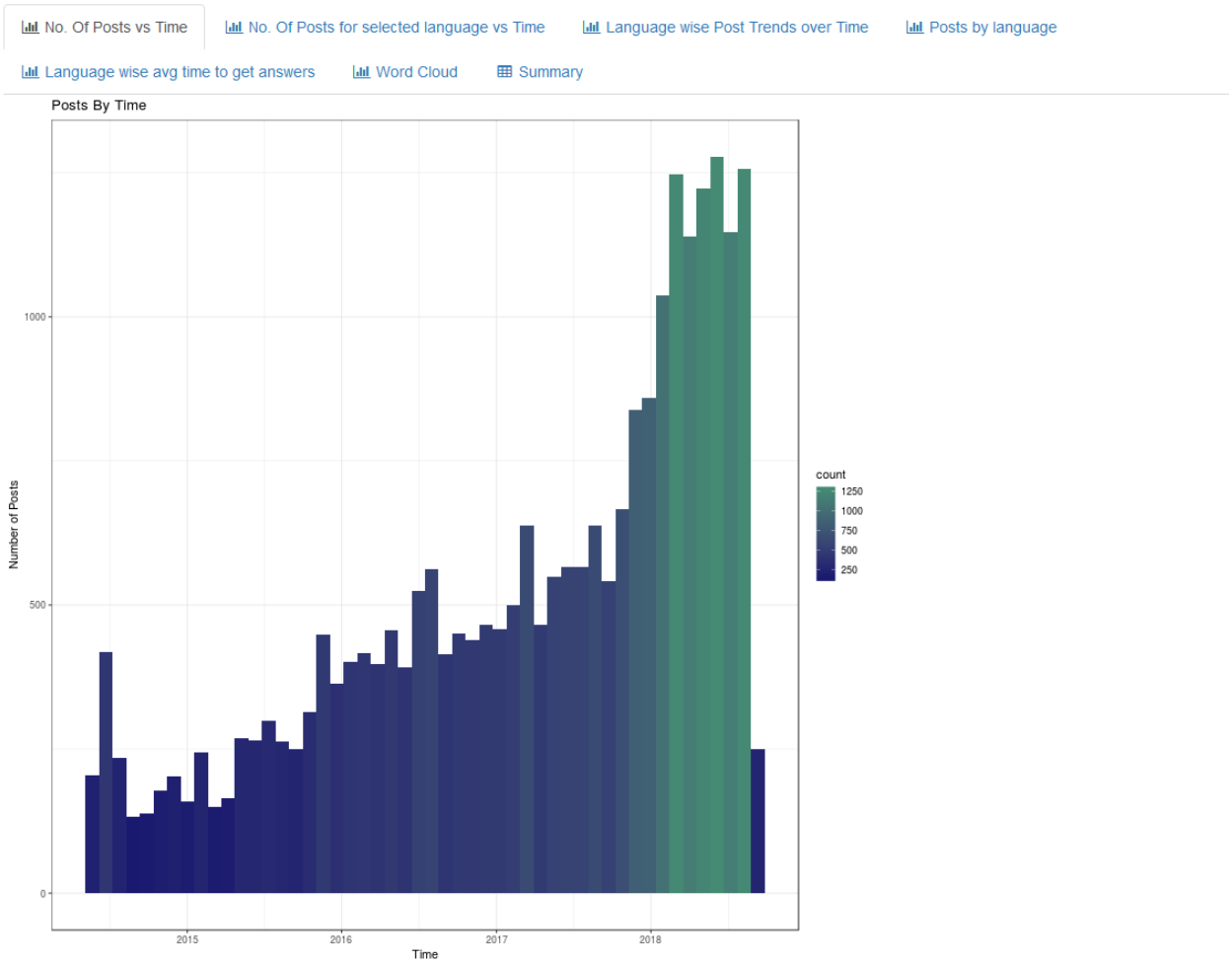


Figure 2. No of posts vs Time

Above figure shows the amount of posts related to a language chosen by the user, posted in a specific time graphically.

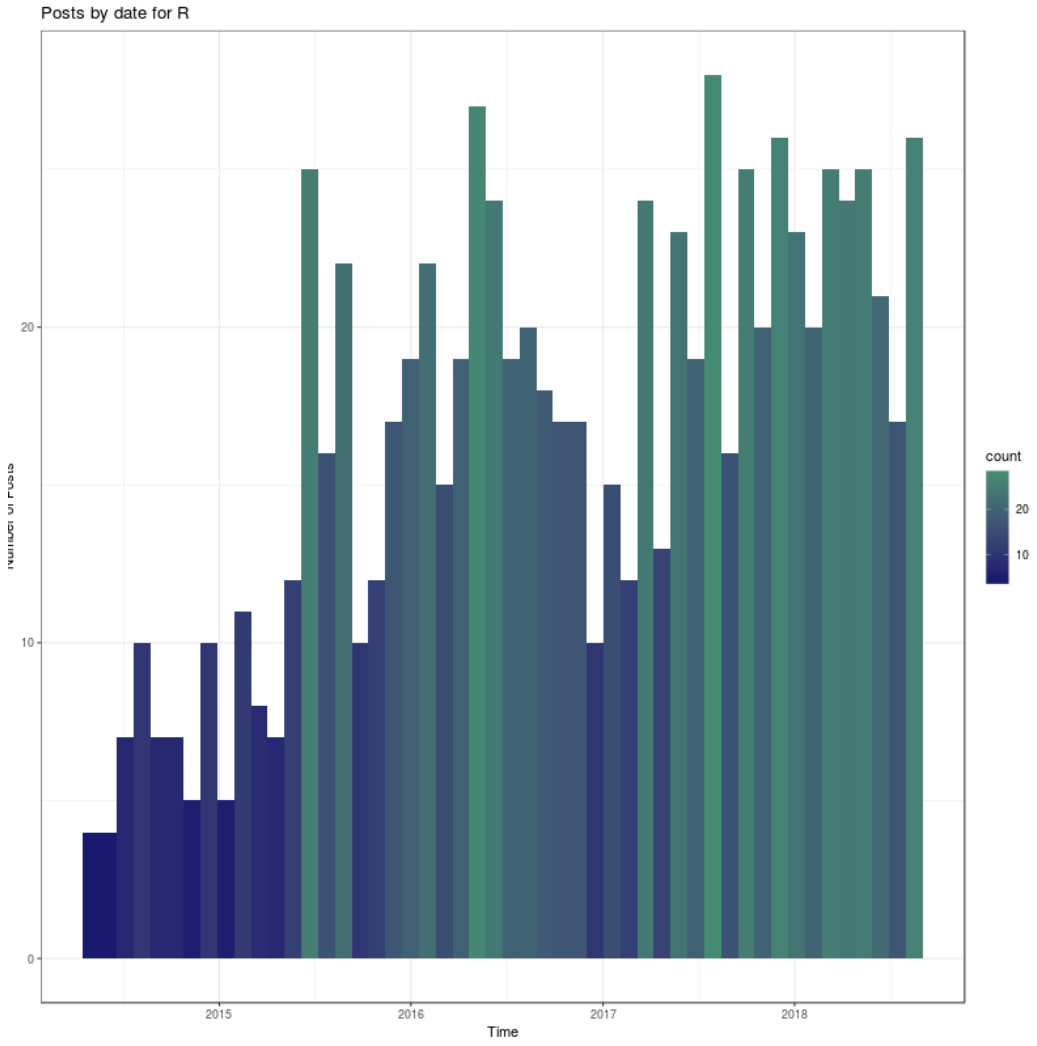


Figure 3. No of posts by date

Above figure shows the amount of posts posted related to a specific language, according to date graphically.

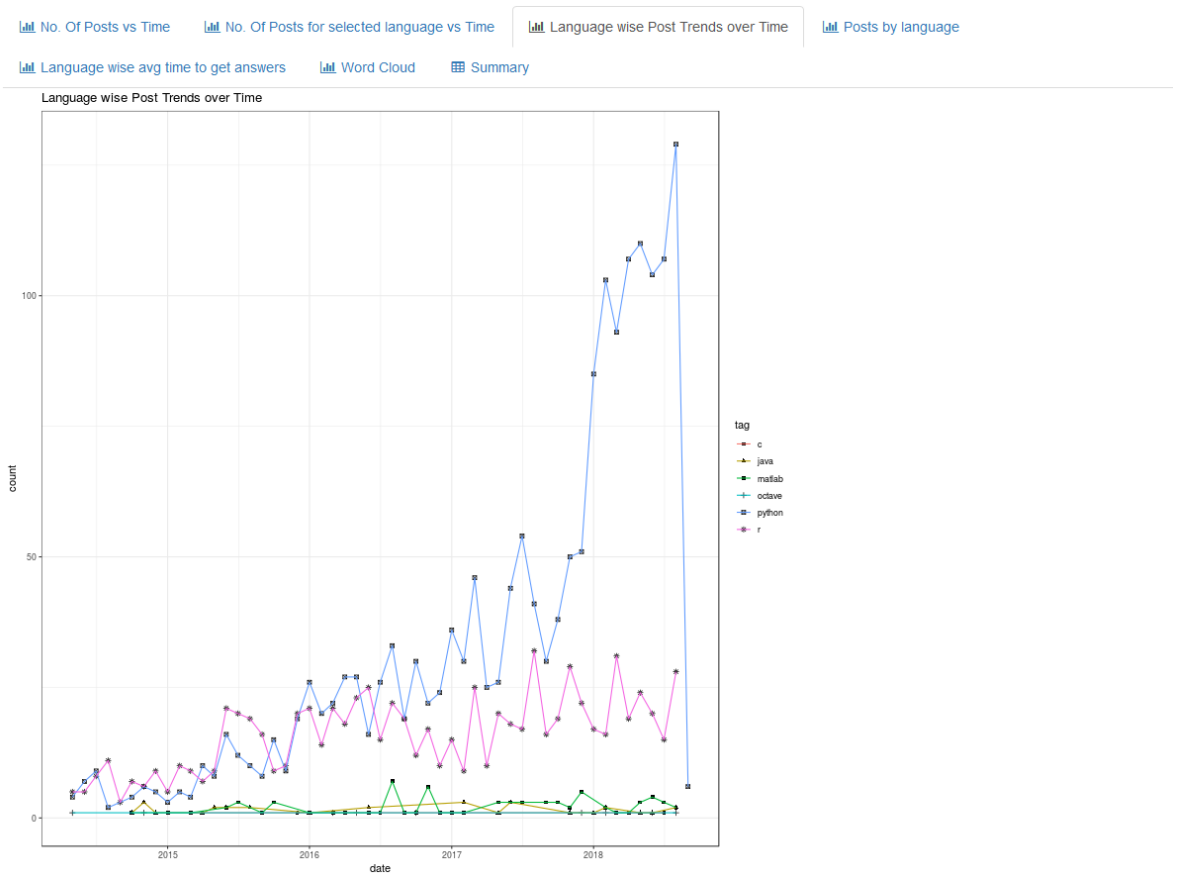


Figure 4. Language vise posts trend over time

Above figure shows how the language related posts trending over time, graphically.

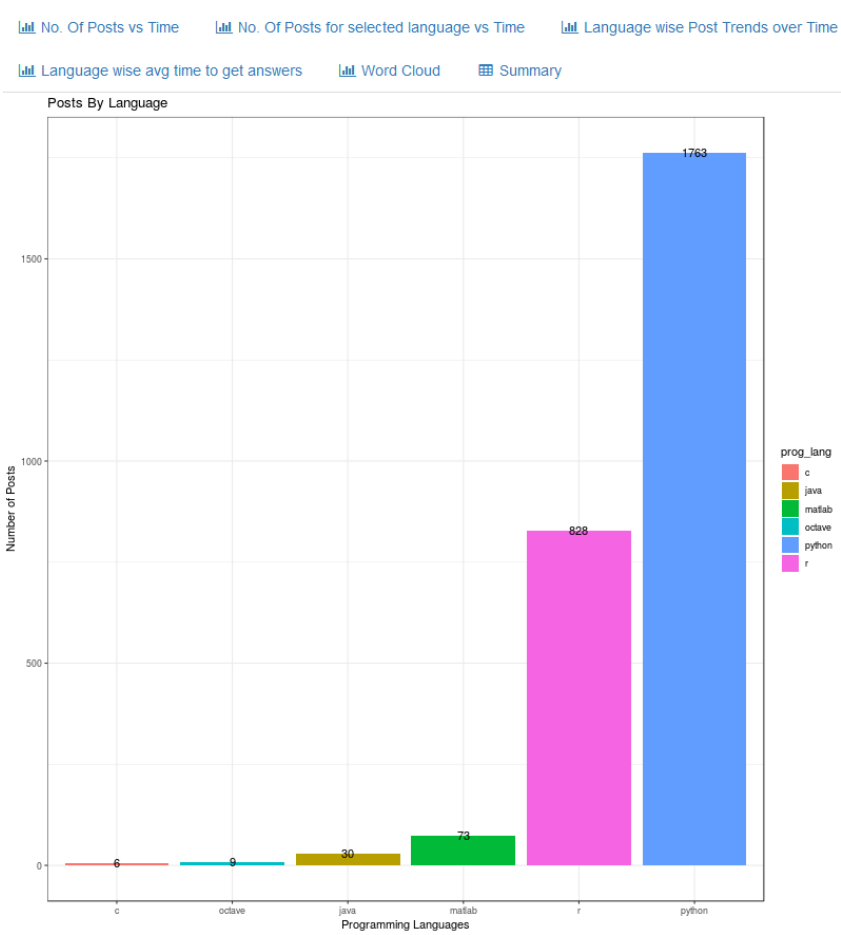


Figure 5: Posts by language

Above figure shows the amount of posts according to related languages graphically.

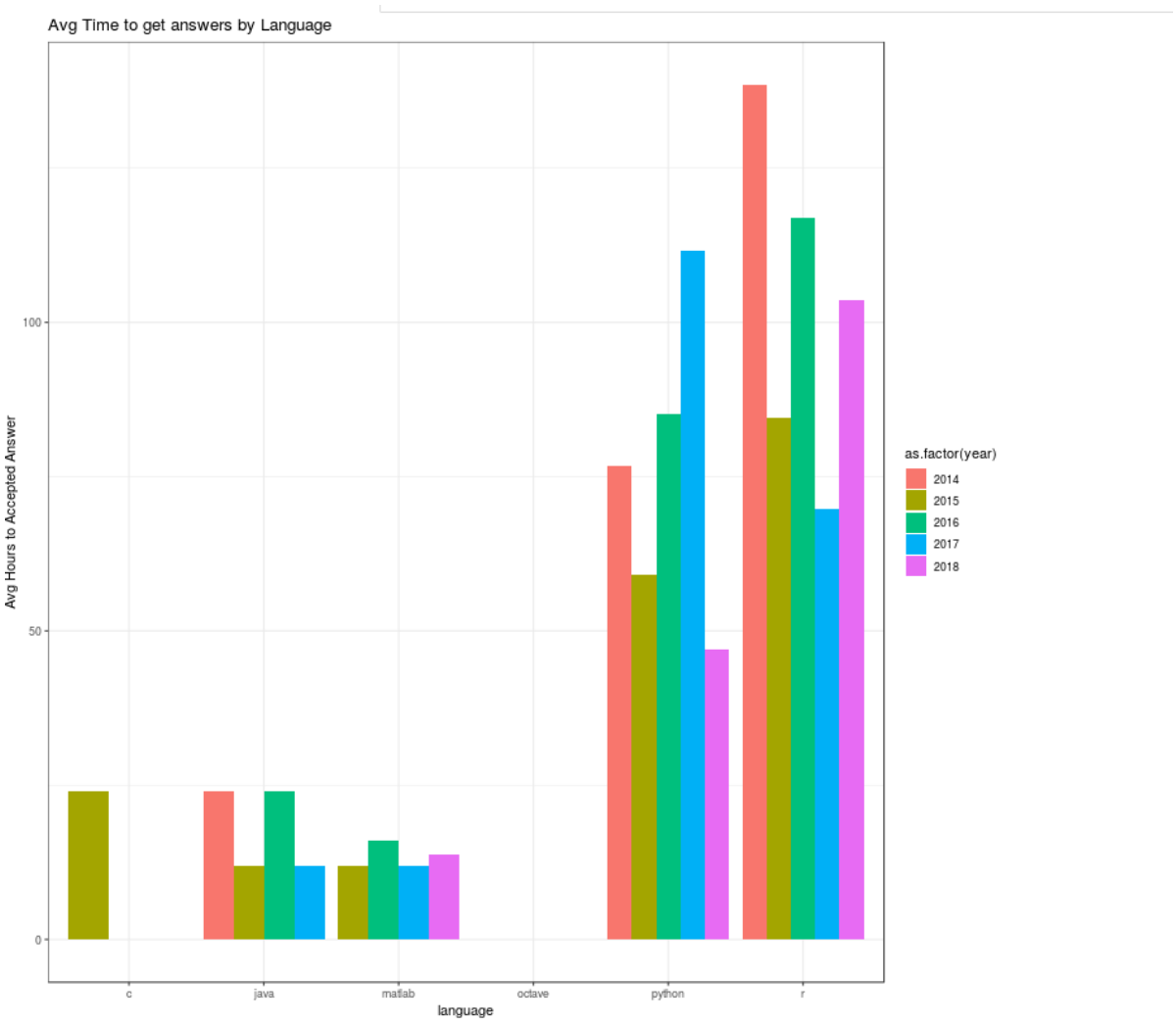
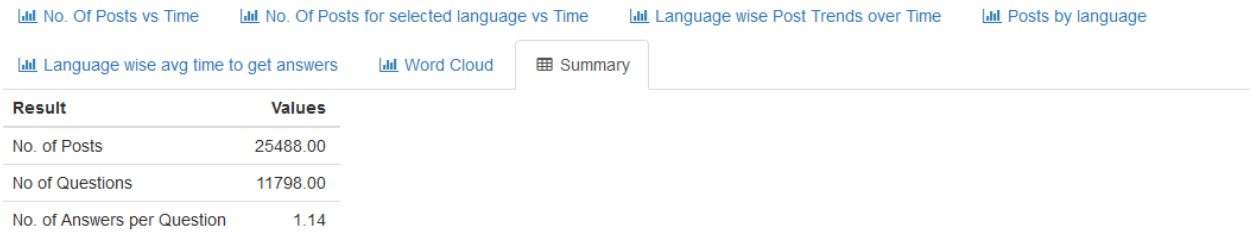


Figure 6: Average time to get answers by Language

This figure shows the average time it take a post to get answered by, related to specific languages graphically.

Figure 7: Word Cloud

This figure shows the word cloud function. It shows the keywords and most used words related to a specific language chosen by the user.



Result	Values
No. of Posts	25488.00
No of Questions	11798.00
No. of Answers per Question	1.14

Figure 8: Summary

This figure shows the summary of Number of posts, Number of questions and Number of answers per question.

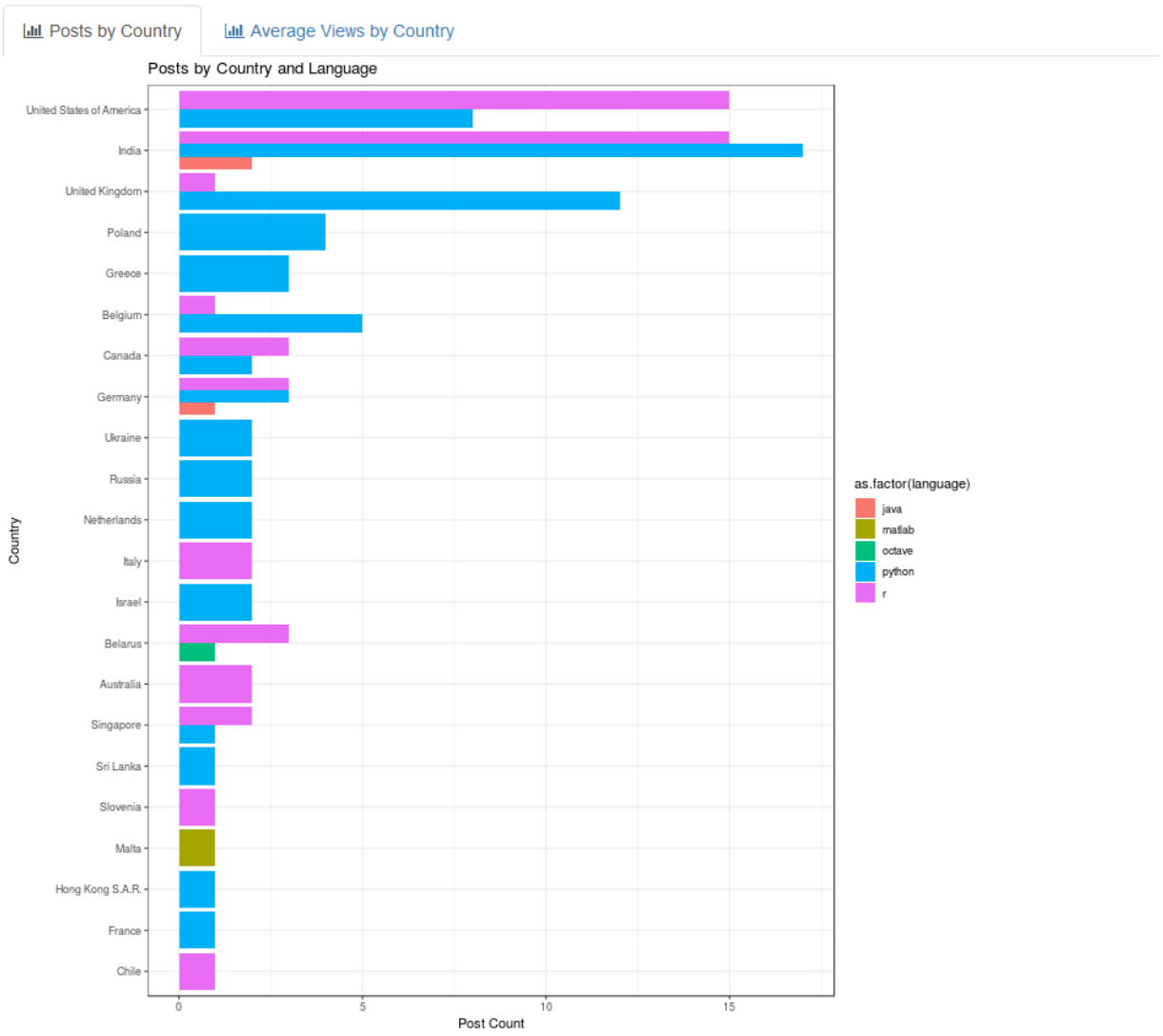


Figure 9: Posts by country

This figure shows the posts related to specific languages countrywise.

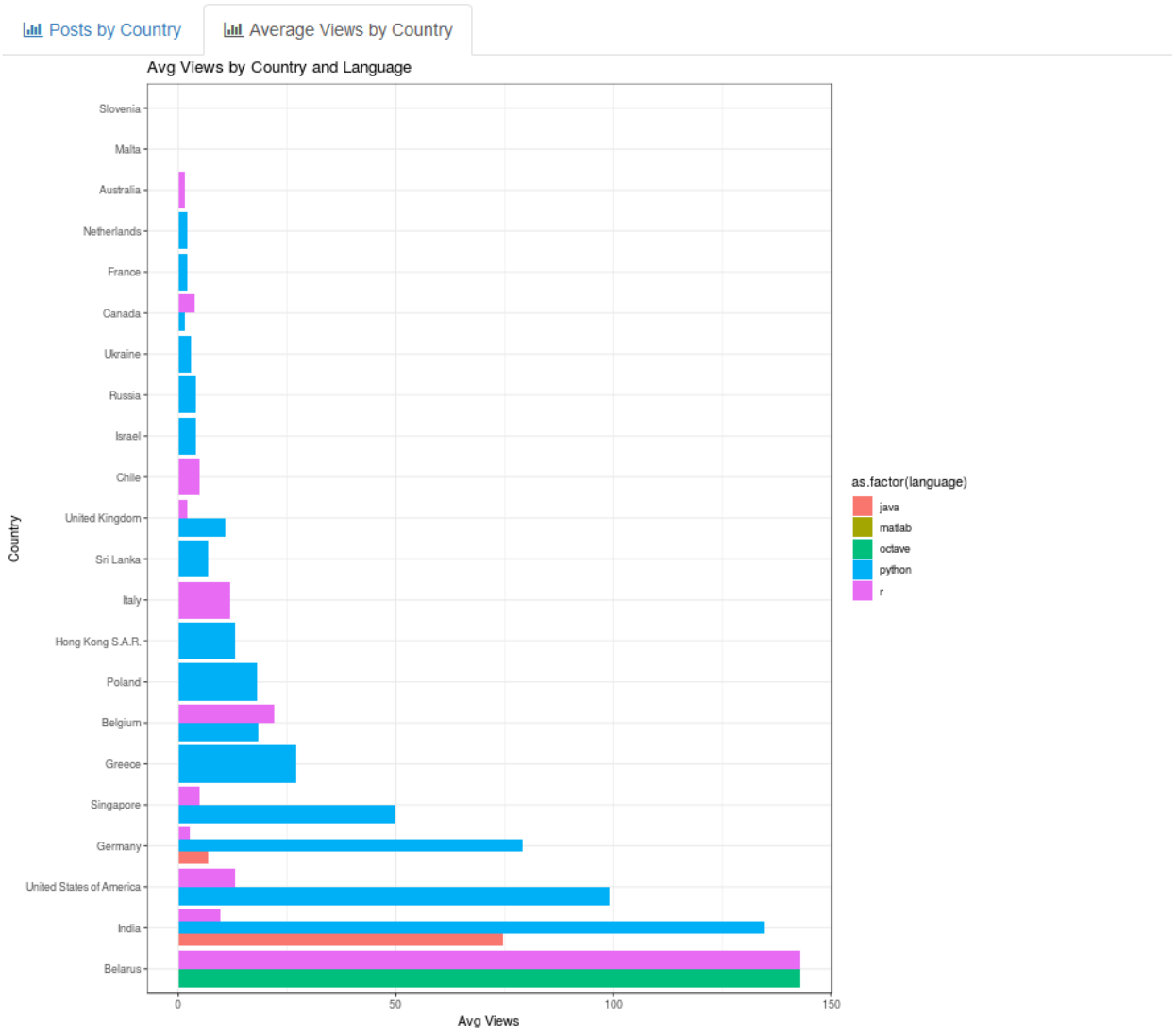


Figure 10: Average views by country

Above figure shows the average views for posts related to languages, countrywise.

5. Discussion

In this project, we use CSV file for get data set. Its take time to load data from CSV file to Data frame, so we need higher processing speed. We use cloud services for that. For save processing time we convert XML file to CSV. We did not able to use Xml to Data frame method because some values are N/A in XML file. We have to filter Programming Languages from tags column, and some of programming languages were not listed. so, some results are not available. Because of privacy concerns some details were not provided in the data set. We are able to understand that results are very much accurate for python, java and R.

6. Conclusion

In this project, mainly we intend to analyze the activity of the community regarding Data Science category. Furthermore, we were going to do the analysis according to the following parts. First, we can analyze no of posts vs. time, no of post for selected language vs. time, language wise post trends over time, posts by language, Language wise average time to get answer, Posts vs. Countries, Average views vs. countries, word cloud and Summary with post analysis and according to date range. In here we analyze data for all the posts and for R language. We can see that number of posts in stack exchange getting increase with time (Figure 2) and period of 2018 shows the highest numbers. For the R language the highest numbers are shown in mid-2017(Figure 3). In Figure 4 we can see the trends of languages over the time and we can see that python is the most trending language, R is the second trending language. Figure 5 shows number of posts for each language and it shows. We can get an idea about the time that takes to get accepted answer for questions (Figure 6) .In Figure 7 we get a word cloud for R language. We can get an idea about programming languages that uses in countries (Figure 10) here we get data for 200 posts.

7. References

[1] T. D. LaToza and A. van der Hoek, “Crowdsourcing in software engineering: Models, motivations, and challenges,” IEEE Software, vol. 33, no. 1, pp. 74–80, 2016.

[2] C. Treude, O. Barzilay, and M.-A. Storey, “How do programmers ask and answer questions on the web? (nier track),” in Proceedings of the 33rd International Conference on Software Engineering, ser. ICSE’11. ACM, 2011, pp. 804–807.

[3] A. S. Badashian, A. Hindle, and E. Stroulia, “Crowdsourced bug triaging,” in Proceedings of the IEEE International Conference on Software Maintenance and Evolution, ser. ICSME’15. IEEE, 2015, pp. 506–510.

[4] C. Rosen and E. Shihab, “What are mobile developers asking about? a large scale study using stack overflow,” Empirical Software Engineering (EMSE), vol. 21, no. 3, pp. 1192–1223, 2016.

[5] A. Barua, S. W. Thomas, and A.E. Hassan, “What are developers talking about? an analysis of topics and trends in stack overflow,” Empirical Software Engineering (EMSE), vol. 19, no. 3, pp. 619–654, 2014.

[6] K. Bajaj, K. Pattabiraman, and A. Mesbah, “Mining questions asked by web developers,” in Proceedings of the 11th Working Conference on Mining Software Repositories, ser. MSR, 2014, pp. 112–121.