Diploma Project

LICENSE USAGE ANALYSIS IN THE R PROGRAMMING LANGUAGE

Alexandros Filippou

UNIVERSITY OF CYPRUS

DEPARTMENT OF COMPUTER SCIENCE

December 2020

UNIVERSITY OF CYPRUS

DEPARTMENT OF COMPUTER SCIENCE

License usage analysis in the R programming language

Alexandros Filippou

Supervising Professor

Kapitsaki Georgia

The Diploma Thesis was submitted for partial fulfilment of the requirements for obtaining a degree in Computer Science from the Department of Computer Science of the University Of Cyprus

December 2020

Summary

The idea that technology has a major effect on our lives is generally acknowledged worldwide. Due to the Covid-19 pandemic our lives have drastically changed for the past few months. In view of that situation a plethora of people turned to alternative ways to earn money. As software developer jobs conquered the market the past year, many of them tried out programming. But do they know the importance of licensing?

Ensuring that it is correctly certified can often be ignored by people or companies when it comes to using computer applications. This is because not enough individuals know the ins-and-outs of what licensing is mandated by law and are unaware of the consequences of using non-licensed software.

The programming language R is one of the most popular among programmers in 2020, as it ranges 8th on the charts. In addition, is the 4th most used programming language for Data Science. Moreover, R is one of the programming languages that other fields of Science, other than Computer Science, use, for example Mathematicians and Biologists. All these reasons make R an ideal programming language to review for the purposes of this Diploma project.

The project is split into multiple components. There are components for fetching and storing data, components that clear and process the data and components that perform topic analysis, all of which will be thoroughly described in this paper.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor Prof. Dr. Kapitsaki Georgia for the guidance and support in all the time of research and writing of this thesis.

I have to also thank all my professors for helping me develop and reinforce my Computer Science related skills.

I would also like to thank Miss Mary Papoutsoglou for offering me assistance in some practical aspects of the project and some datasets.

Finally, I would like to thank my family for supporting me throughout my university years and providing everything for me, so all I had to do was focus on my studies.

Table of Contents

CHAPTER 1

Introduction

1.1 Motive of the Problem

Generally, the fact that technology has a significant influence on our lives is known globally. Our lives have dramatically altered over the last few months due to the Covid-19 pandemic. A plethora of people have turned to alternate means to gain an income in spite of the situation. In the past year, several of them attempted programming as data analysis conquered the market.

The process of choosing a license for an open source project is an important task, as there are many factors on a license decision. But many tend to disregard or neglect this process. This is because not enough individuals know the ins-and-outs of what licensing is mandated by law and are unaware of the consequences of using non-licensed software. Also, a significant amount of code in projects is open source code. Software reuse often can be challenging as the developer must take to consideration the software license that code is under.

In general, the programming languages that are most relevant to Data Science are R and python. We want to test what is being done in this ecosystem regarding permissions, because data is starting to be used more and developers are building new packages based on the old ones. Also, in R there is the concept of tidy data where tidy libraries use a variety of other R libraries and so maybe there is a need to see if the community is geared towards some more open permissions.

The programming language R is one of the most popular among programmers in 2020, as is one of the programming languages that other fields of Science, other than Computer Science, use. Generally, R initially started as a more closed language only for mathematical modeling, compared to python, and it is interesting to see first what they are discussing about licenses in this community and then in the future a comparison with python. This makes R an ideal programming language to review for the purposes of this Diploma project.

1.2 Research Problem

The purpose of this study is to develop an evaluation model for license usage and analysis in the R community. To gain deeper insights on license usage analysis in the R community, general research questions are elaborated and answered within the scope of this thesis.

RQ1: What are the most common used licenses in the R community?

RQ2: What are the most popular licenses throughout the years in the R community?

RQ3: What are the most popular platforms used for questions about licenses throughout the years in the R community?

RQ4: What are the most common questions asked about licenses in the R community?

RQ5: Is license a popular topic among the R community?

1.3 Thesis Structure

In this chapter, the incentive, goals and outline of the project are described.

In the second chapter, the theoretical background will be presented, describing all technologies, platforms and tools that were used to achieve the goals of the project.

In the third chapter, the methodological approach is described. It contains a detailed and procedural description of how the developed algorithms operate and how data is handled at all phases. The development techniques used, and the code structure for all the scripts that the algorithms use will be explained.

The fourth chapter consists of the analysis of the results, and the completion level of the goals that were set for the project.

In the fifth chapter, the final and holistic conclusions regarding the project will be drawn and explained.

CHAPTER 2

Theoretical Background

2.1 Open Source Software and Licensing

"Software licenses are legal instruments to ensure the originators copyright and regulate the distribution and usage. A license agreement between the user of software and its originator governs all legal aspects of usage and redistribution, this can include limitations of liability, warranties and disclaimers. In general, software licenses can be fit into two categories: proprietary software licenses and (free- and) open source software licenses. Licenses for open source software can be divided in permissive and copy-left licenses. Permissive licenses aim to have minimal requirements on the redistribution of the soft- ware, whereas copy-left licenses ensure that all subsequent end-users receive the specified rights." [1]

New technologies and projects are continuously being created by academics at academic universities and distributed in free open source software repositories. Therefore, to decide the further use of his work, a developer will select from one of several licenses. Any of these licenses, based on its intended purpose, has its own characteristics and, therefore, advantages and disadvantages. This encourages a state of affairs that is dynamic. It is the responsibility of the creator to select a correct license for his work properly, which is not well known so far.

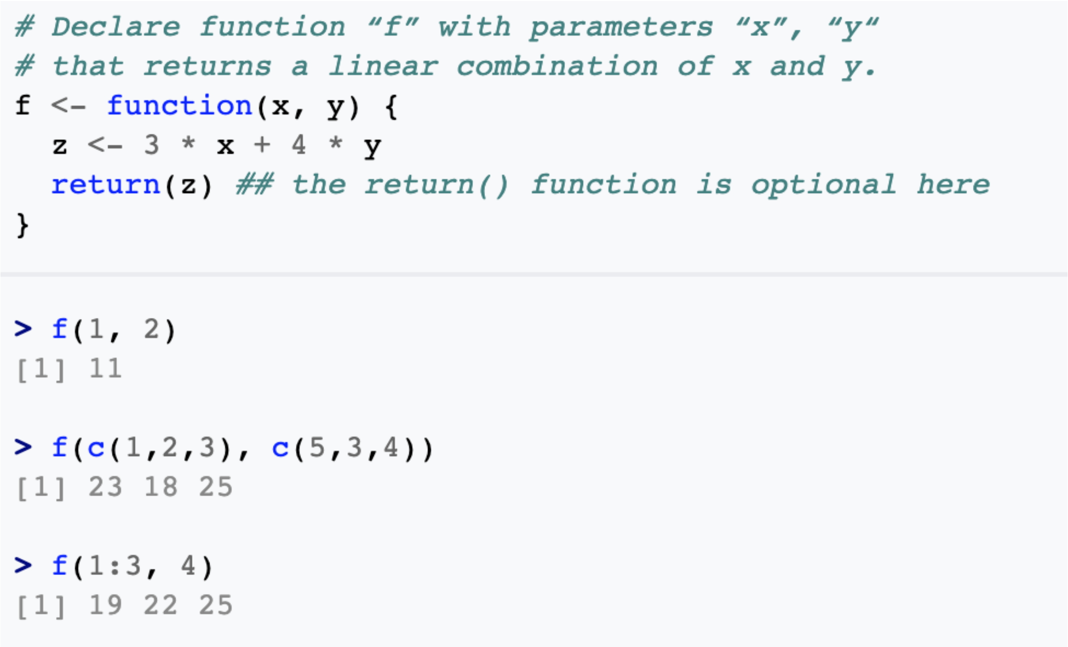
2.2 R (programming language) / Communities / Sources

"R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, data mining surveys, and studies of scholarly literature databases show substantial increases in popularity; as of September 2020, R ranks 9th in the TIOBE index, a measure of popularity of programming languages." [1]
R as a package is licensed under GPL-2 | GPL-3.

2.2.1 Packages

"The capabilities of R are extended through user-created packages, which allow specialised statistical techniques, graphical devices, import/export capabilities, reporting tools , etc. These packages are developed primarily in R, and sometimes in Java, C, C++, and Fortran. The R packaging system is also used by researchers to create compendia to organise research data, code and report files in a systematic way for sharing and public archiving." [1]

One of R's, as an open source programming language, strengths is the ease of creating new functions and providing them as open source packages or libraries for others to use.



```
# Declare function “f” with parameters “x”, “y“
# that returns a linear combination of x and y.
f <- function(x, y) {
  z <- 3 * x + 4 * y
  return(z) ## the return() function is optional here
}
```

```
> f(1, 2)
[1] 11

> f(c(1,2,3), c(5,3,4))
[1] 23 18 25

> f(1:3, 4)
[1] 19 22 25
```

Figure 2-1: Example of Function in R

A simple example as the above in the Figure 2-1, could be installed and used by other users. To install an R package and make its contents available to use is archived with the following commands: *install.packages("<the package's name>")* and *library("<the package's name>").* There are all of sorts of helpful packages in R for free use. Some of the top most downloaded R packages are: [13]

- To load data (DBI, odbc, RMySQL)
- To manipulate data (tidyverse, dplyr, tidyr)
- To visualize data (ggplot2, ggvis,rgl)
- To model data (tidymodels, car, mgcv)
- To report data (shiny, RMarkdown, xtable)
- For spatial data (sp, maps, ggmap)
- For time series and financial data (zoo, xts, quantmod)
- For work with the Web (XML, jsonlite, httr)

For example, the *tidyr* package provides the following functions to the users:

gather() – To convert columns to rows with key and value pairs.
spread() – To convert rows to columns.
separate() – To separate a single column to multiple columns.
unite() – To combine multiple columns into a single column.

2.2.2 r-project.org

"R includes extensive facilities for accessing documentation and searching for help. There are also specialized search engines for accessing information about R on the internet, and general internet search engines can also prove useful." [3]

2.2.2.1 R Help: help() and ?

"The help() function and ? help operator in R provide access to the documentation pages for R functions, data sets, and other objects, both for packages in the standard R distribution and for contributed packages." [3]

2.2.2.2 CRAN

The Comprehensive R Archive Network (CRAN) is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R. [2]

2.2.2.3 R Email Lists

"The R Project maintains a number of subscription-based email lists for posing and answering questions about R, including the general R-help email list, the R-devel list for R code development, and R-package-devel list for developers of CRAN packages; lists for announcements about R and R packages; and a variety of more specialized lists. Before posing a question on one of these lists, please read the R mailing list instructions and the posting guide." [4]

2.2.2.4 R FAQ

R FAQ is an option in R-Help which provides the Frequently Asked Questions on R. [5]

2.2.2.5 R Licenses

R Licenses is an option in R-Help which provides a list of licenses used for R or associated software such as packages.

The following licenses are in use for R or associated software such as packages. [6]

- The “GNU Affero General Public License” version 3
- The “Artistic License” version 2.0
- The “BSD 2-clause License”
- The “BSD 3-clause License”
- The “GNU General Public License” version 2
- The “GNU General Public License” version 3
- The “GNU Library General Public License” version 2
- The “GNU Lesser General Public License” version 2.1
- The “GNU Lesser General Public License” version 3
- The “MIT License”
- The “Creative Commons Attribution-ShareAlike International License” version 4.0

2.2.2.6 r-help archives

R-help archives is an option in R-Help which provides a list of all questions ever asked in R-Help, sorted either by Month/Year, Thread, Subject, Author or Date. [7]

2.2.2.7 The Mail Archive

The Mail Archive is one of the R Email Lists which provides a search engine for the archived lists. [8]

2.2.3 Stack Exchange

“Stack Exchange is a network of question-and-answer (Q&A) websites on topics in diverse fields, each site covering a specific topic, where questions, answers, and users are subject to a reputation award process. It has excellent searchability, especially as topics are tagged.”

“The Stack Exchange network comprises 173 Q&A communities including Stack Overflow, the largest, most trusted online community for developers to learn, share their knowledge, and build their careers.” [9]

The past year the Stack Exchange Network had 423.5M Monthly Visits, 3.2M Questions Asked and 3.5M Answers Submitted.

2.3 Latent Dirichlet Allocation (LDA)

“In natural language processing, the latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. LDA is an example of a topic model and belongs to the machine learning toolbox and in a wider sense to the artificial intelligence toolbox.” [1]

Latent Dirichlet Allocation (LDA) model is a statistical model that follows the principles of k-means clustering to produce its results. The LDA algorithm is used mostly in two areas, Biology and Natural Language Processing. In Biology, LDA is used to diagnose the occurrence of organized genetic variation in a population of people. In Natural Language Processing, as in this paper, LDA is used more specifically for Individual Topic Detection from a given text input.

Some applications of LDA in Natural Language Processing are:

- Topic categorization on social media data (i.e. tweets)
- Topic prediction
- Surveys
- Communication research
- Extraction of linguistic phrases
- Sentiment extraction

CHAPTER 3

Methodological Approach

3.1 Methodology

The method of this study is achieved by splitting it into three sequential stages: the initial study, the data collection, and the analysis, and this is shown in Figure 3-1.

Initial Study
Data Collection
Analysis

Figure 3-1: Visualization of the approach of the thesis

Initial study

The Initial Study has been an iterative process which has mainly consisted of understanding the problem, studying relative papers and finding the data to be collected.

Data Collection

The Data Collection was the process which mainly consisted of collecting the data and performing different filters on them, in order to get analysable data. The process of collecting data is roughly precented as follows:

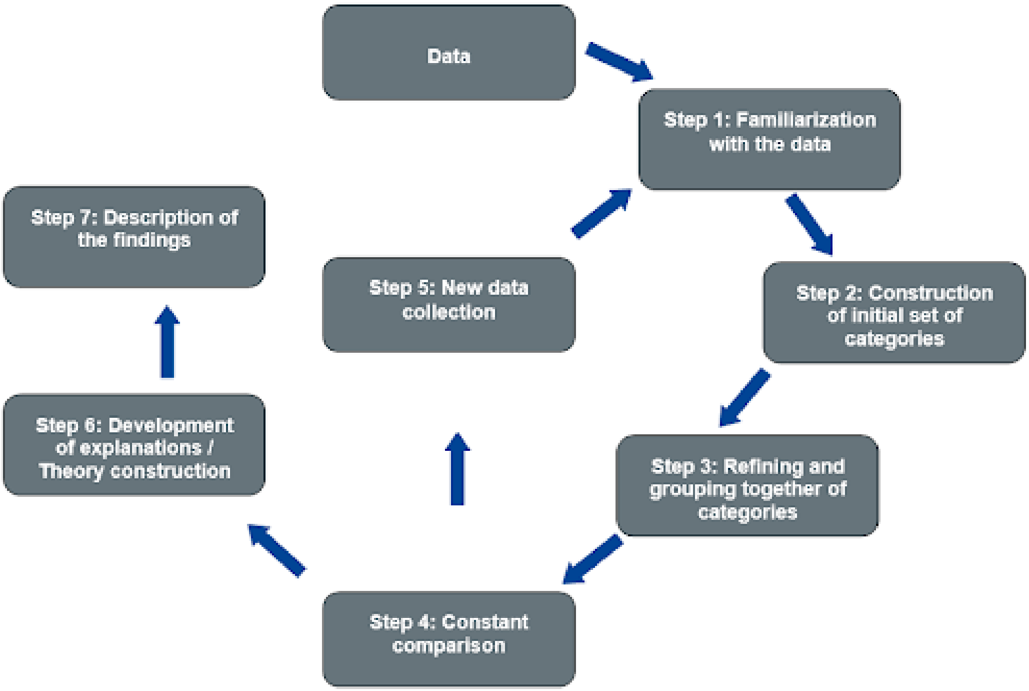

Data
Step 1: Familiarization with the data
Step 2: Construction of initial set of categories
Step 3: Refining and grouping together of categories
Step 4: Constant comparison
Step 5: New data collection
Step 6: Development of explanations / Theory construction
Step 7: Description of the findings

Figure 3-2: Data Collection

Analysis

The Analysis stage was the process which all the analysis and the conclusions about the results were deliberated. The analysis stage was achieved by splitting it into three sequential stages: Statistical analysis and common licenses, Topic analysis and Implications for practitioners.

For the stage of Statistical analysis and common licenses, the goal was to extract graphs and tables with data to indicate the most common and popular licenses throughout the years. Also, to gather data to determine the most popular platforms used for questions about licenses throughout the years in the R community.

For the stage of Topic analysis, the goal was to use the data produced by the LDA algorithm, to estimate the most common questions asked about licenses in the R community.

For the stage of Implications for practitioners, the goal was the extraction of the final results overall and conclude if license a popular topic among the R community.

As mentioned above the project is splitted into three sequential stages: the initial study, the data collection, and the analysis. A thorough research led to four sources of data: c-ran packages, R-help archive, Stack Overflow and Stack Exchange.

3.2 Data Fetch from Sources

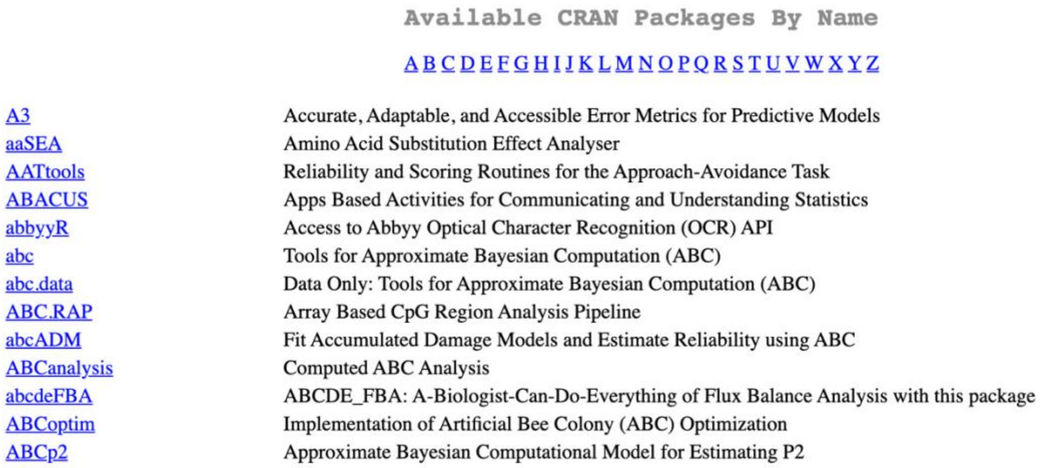
The constructs of each source are different, thus the way the data was collected varies.

[1] c-ran packages:

This source is a web page of CRAN, which provides a list of all the repositories of available packages in R. Each repository provides a number of information about the package, such as name, author, version and licence.

The task was to make a csv file, through java, with all the fields of data the repository provides for every package.

The figures below showcase the format of the CRAN packages list from the website and the details provided by the website for each package.

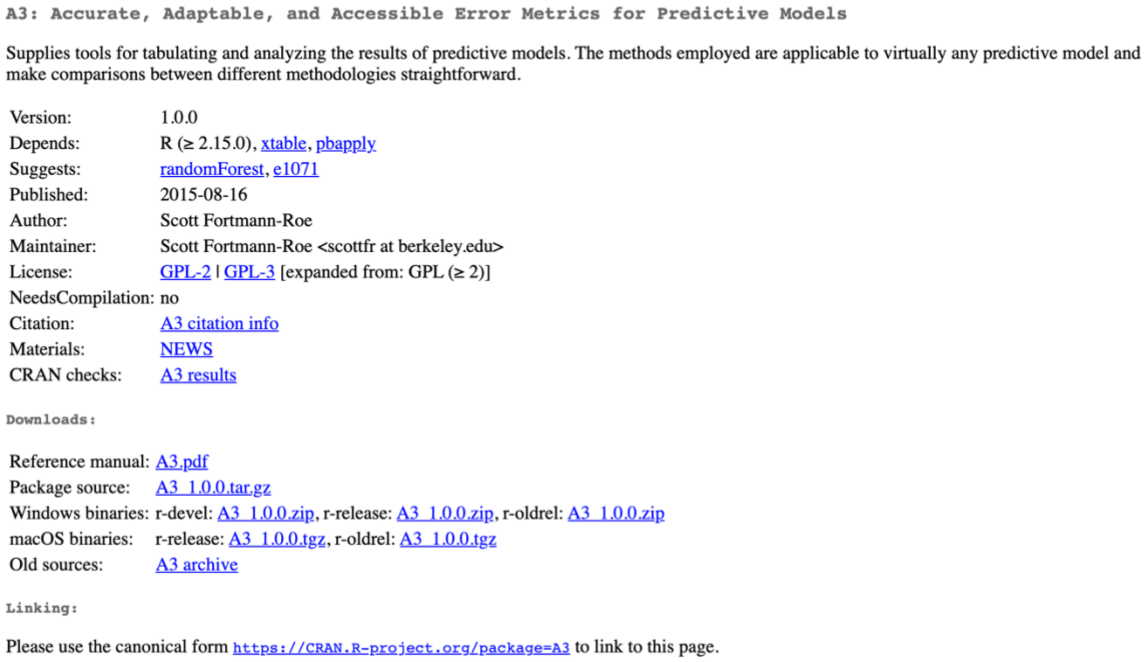


Available CRAN Packages By Name

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

A3	Accurate, Adaptable, and Accessible Error Metrics for Predictive Models
aaSEA	Amino Acid Substitution Effect Analyser
AATtools	Reliability and Scoring Routines for the Approach-Avoidance Task
ABACUS	Apps Based Activities for Communicating and Understanding Statistics
abbyyR	Access to Abbyy Optical Character Recognition (OCR) API
abc	Tools for Approximate Bayesian Computation (ABC)
abc.data	Data Only: Tools for Approximate Bayesian Computation (ABC)
ABC.RAP	Array Based CpG Region Analysis Pipeline
abcADM	Fit Accumulated Damage Models and Estimate Reliability using ABC
ABCanalysis	Computed ABC Analysis
abcdeFBA	ABCDE_FBA: A-Biologist-Can-Do-Everything of Flux Balance Analysis with this package
ABCoptim	Implementation of Artificial Bee Colony (ABC) Optimization
ABCp2	Approximate Bayesian Computational Model for Estimating P2

Figure 3-3: CRAN packages list



A3: Accurate, Adaptable, and Accessible Error Metrics for Predictive Models

Supplies tools for tabulating and analyzing the results of predictive models. The methods employed are applicable to virtually any predictive model and make comparisons between different methodologies straightforward.

Version:	1.0.0
Depends:	R (≥ 2.15.0), xtable, pbapply
Suggests:	randomForest, e1071
Published:	2015-08-16
Author:	Scott Fortmann-Roe
Maintainer:	Scott Fortmann-Roe <scottfr at berkeley.edu>
License:	GPL-2 | GPL-3 [expanded from: GPL (≥ 2)]
NeedsCompilation:	no
Citation:	A3 citation info
Materials:	NEWS
CRAN checks:	A3 results

Downloads:

Reference manual:	A3.pdf
Package source:	A3_1.0.0.tar.gz
Windows binaries:	r-devel: A3_1.0.0.zip, r-release: A3_1.0.0.zip, r-oldrel: A3_1.0.0.zip
macOS binaries:	r-release: A3_1.0.0.tgz, r-oldrel: A3_1.0.0.tgz
Old sources:	A3 archive

Linking:

Please use the canonical form `https://CRAN.R-project.org/package=A3` to link to this page.

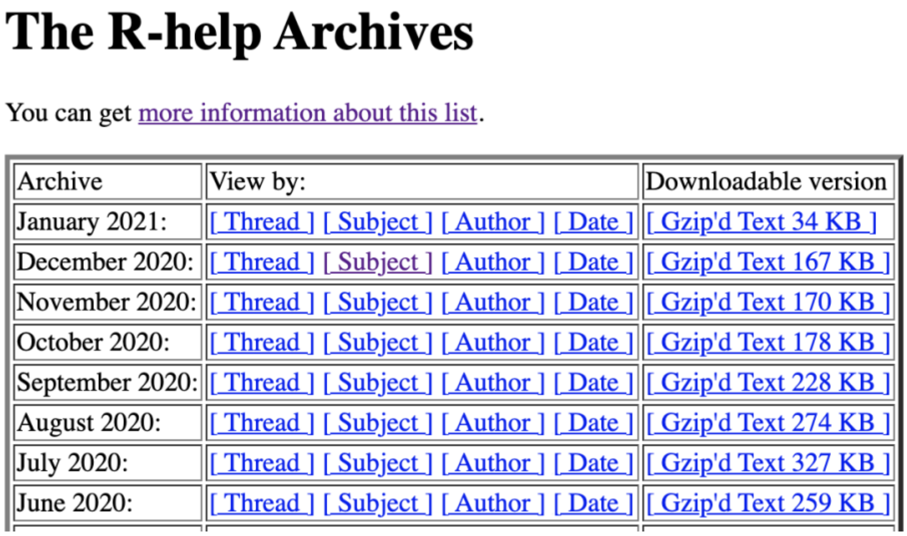
Figure 3-4: CRAN package example

[2] R-help archive:

This source is a web page of CRAN, which provides a list of all the questions and answers made to the R-help mailing list. The data is in the form of a text file, one text file for every month from April 1997 until today.

The task was to make, through java, respectively text files with only the questions/answers containing a keyword, such as “licence”, “MIT” or/and “GPL”.

The figures below showcase the format of the R-help Archive website and the txt files extracted.

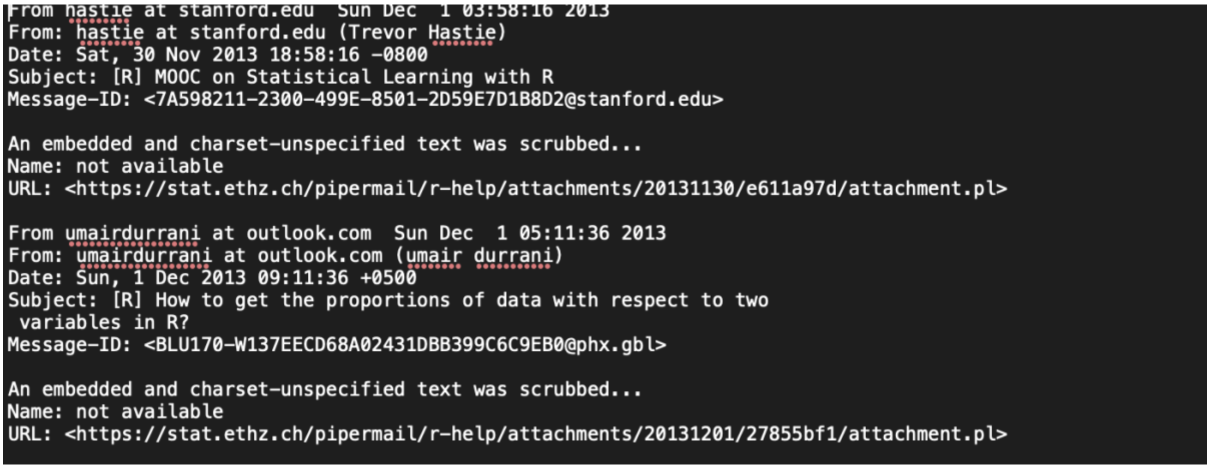


The R-help Archives

You can get more information about this list.

Archive	View by:	Downloadable version
January 2021:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 34 KB]
December 2020:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 167 KB]
November 2020:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 170 KB]
October 2020:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 178 KB]
September 2020:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 228 KB]
August 2020:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 274 KB]
July 2020:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 327 KB]
June 2020:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 259 KB]

Figure 3-5: The R-help Archives



```
From hastie at stanford.edu  Sun Dec  1 03:58:16 2013
From: hastie at stanford.edu (Trevor Hastie)
Date: Sat, 30 Nov 2013 18:58:16 -0800
Subject: [R] MOOC on Statistical Learning with R
Message-ID: <7A598211-2300-499E-8501-2D59E7D1B8D2@stanford.edu>

An embedded and charset-unspecified text was scrubbed...
Name: not available
URL: <https://stat.ethz.ch/pipermail/r-help/attachments/20131130/e611a97d/attachment.pl>

From umairdurrani at outlook.com  Sun Dec  1 05:11:36 2013
From: umairdurrani at outlook.com (umair durrani)
Date: Sun, 1 Dec 2013 09:11:36 +0500
Subject: [R] How to get the proportions of data with respect to two
 variables in R?
Message-ID: <BLU170-W137EECD68A02431DBB399C6C9EB0@phx.gbl>

An embedded and charset-unspecified text was scrubbed...
Name: not available
URL: <https://stat.ethz.ch/pipermail/r-help/attachments/20131201/27855bf1/attachment.pl>
```

Figure 3-6: txt example from R-help

[3] Stack Overflow and Stack Exchange:

The datasets were provided by the university in a csv file for each. Specifically, the data from Stack Exchange sites were collected in the framework of a previous research work and they were provided for use in the framework of the current thesis. [17]

The following table showcase the collective of data from sources:

	From	To	Data	Final
CRAN	Mar-06	Mar-20	15716	15716
RHELP	Apr-97	May-20	162797692	377
SE	Jun-15	Mar-20	2418	15
SO	Sep-21	Oct-20	694	146

Table 3-1: Data from sources

3.3 Data Cleansing and Pre-processing

To filter the data a table with keywords was used. The table was created by a list with the names of the most popular software licences and words like “license” and “licensing”. The keywords table consists of 122 keywords. After several processing, a new keywords table was created with 11 of the most common keywords. Also, a table with error words was created. The error table consisted of words that after pre-processing and manual checking were fitted to be irrelevant data.

For creating the Error table below, much pre-processing and manual checking was due. Most of these words relate to installation or updating of packages and libraries, where the license is stated. It is also worth noting that the “MIT” keyword was more often used as a reference to the Massachusetts Institute of Technology rather than the MIT licence.

Error table

[install,linux,installation,debug,library,compiling,package,version,error]

Keywords table

[license,licence,licencing,licensing,MIT,GPL,GPL 3.0,GPL 2.0,LGPL,BSD,Apache 2.0]

[1] c-ran packages:

The c-ran packages csv file did not require any pre-processing or data cleansing as the licence of each package was provided on data fetching.

[2] R-help archive:

The first task was to clear the questions/answers from symbols, whitespaces and convert the text to lowercase letters only. The second task was to create a list of java objects with the subject, the paragraph, the date and the number Of Keys. With the help of the keywords table the irrelevant data were removed. Then, all the duplicate entries and entries with only one key were removed, as after manual research concluded that single key entries were irrelevant. The third and final task was to make a csv file, through java, with the fields of data that was collected, such as date, subject and key-count.

[3] Stack Overflow and Stack Exchange:

For clearing the data of Stack Exchange and Stack Overflow, python scripts were used. The first task was to clear the questions/answers from symbols, whitespaces and convert the text to lowercase letters only. After that, the data was cleared for entries without the tag "R" and with the use of table keywords, the irrelevant data were removed. In addition, entries containing error words in the Subject and single hits were removed. The third and final task was to make a csv file with the fields of data that was collected, such as date, subject and key-count.

3.4 Topic Modelling with LDA

The LDA topic modelling algorithm was implemented in python with the use of several libraries.

A csv file was created with the Subjects from all the entries of the three datasets: R-help archive, Stack Exchange and Stack Overflow.

First of all, with the help of the pandas library all the data were converted to lowercase. The second step was to remove any of the stopwords using the python library of nltk. Next, symbols were removed and the nodes were tokenized with python library nltk. "*Tokenization is the process of substituting a sensitive data element with a non-sensitive equivalent, referred to as a token, that has no extrinsic or exploitable meaning or value.*" [1] After that, all the data were lemmatized with an English dictionary. Lemmatization is used to build a dictionary where for each subject, each word has its own id. The next step was to build the corpus, i.e. vectors with the number of occurrence of each word per subject. Next, the coherence is calculated. "*Topic Coherence measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic.*" [11]

For the Coherence Model the following parameters were used:

CoherenceModel*(model= lda, corpus= sub_corpus, dictionary= sub_dictionary, texts= new_subs.lemmatized, coherence = 'c_v')*

C_v *measure is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity.* [11]

For the next step the LDA model is used with the following parameters:

LdaModel*(sub_corpus, num_topics=k, id2word=sub_dictionary, passes=10)*

The parameter num_topics, "*the number of requested latent topics to be extracted from the training corpus*" [12] , was set to 5 as after multiple tests was deemed appropriate. The number of topics was set to 5, since most of the trial runs showed that Coherence is high on 5 topics, as shown at the figures bellow, and the overall data is a small set, 538 subjects.

The final step was to visualise the results of the LDA and the Coherence. As the Coherence graph suggests, the LDA should return topics with 10 words for better accuracy and coherence. The above LDA model is built with 10 different topics where each topic is a combination of keywords and each keyword contributes a certain weightage to the topic.

Find below some of the manual testing made for determining the best suitable parameters for num_topics (k) and nb_words, words per topic.

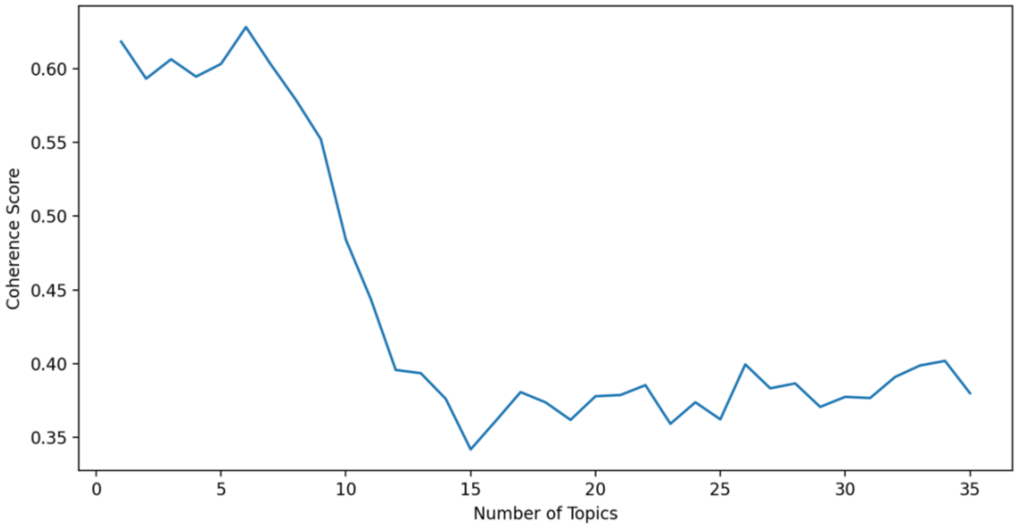


Figure 3-7: Coherence for Stack Overflow with np = 8 k = 5

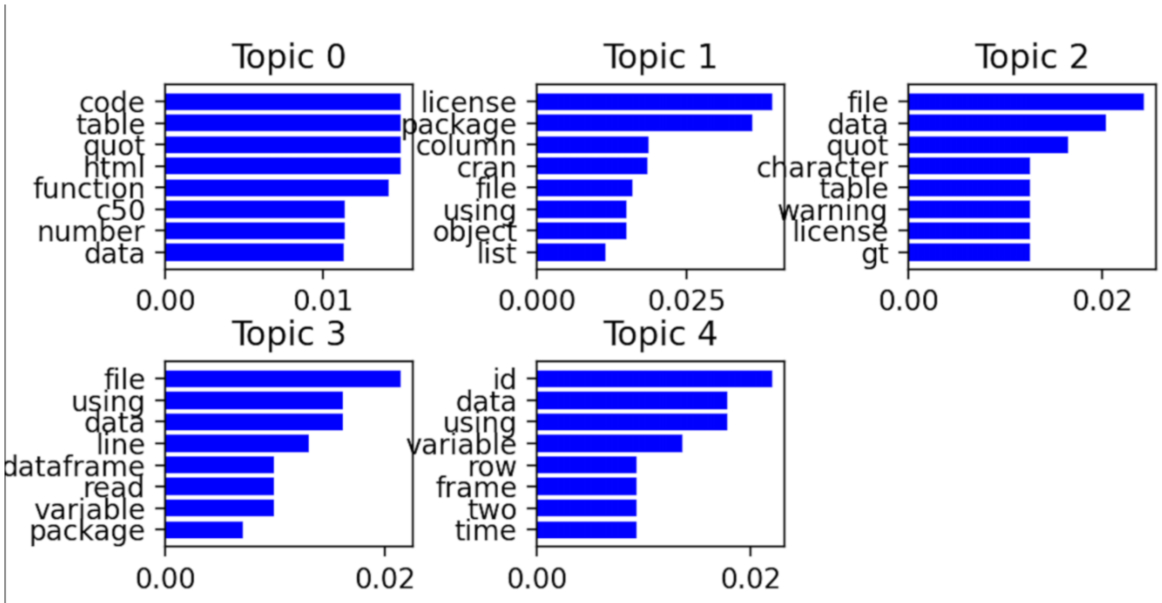


Figure 3-8: LDA topics for Stack Overflow with np = 8 k = 5

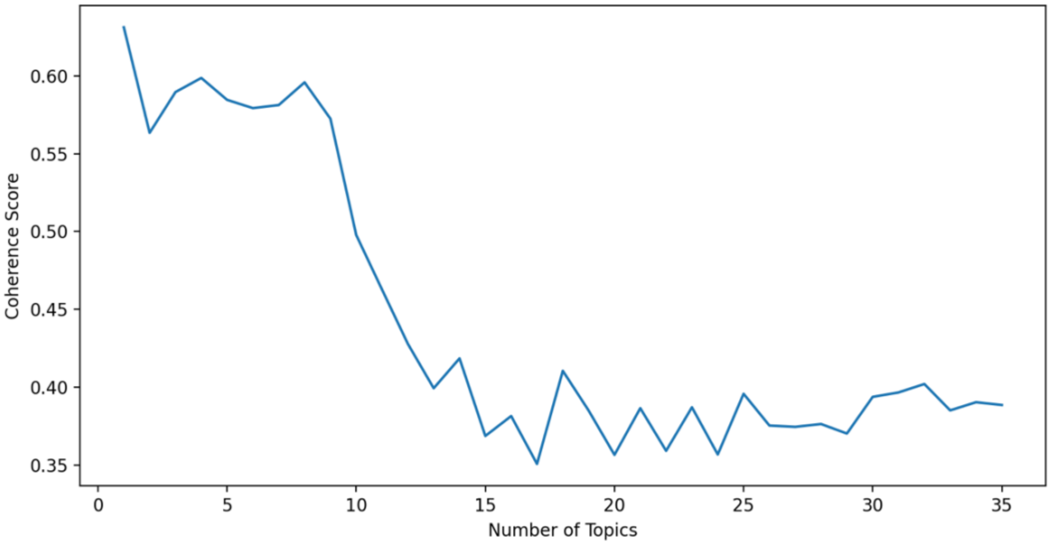


Figure 3-9: Coherence for Stack Overflow with np = 10 k = 10

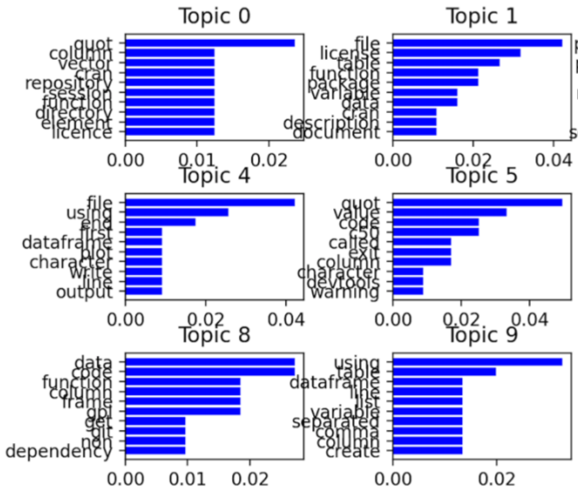


Figure 3-10: LDA topics for Stack Overflow with np = 10 k = 10

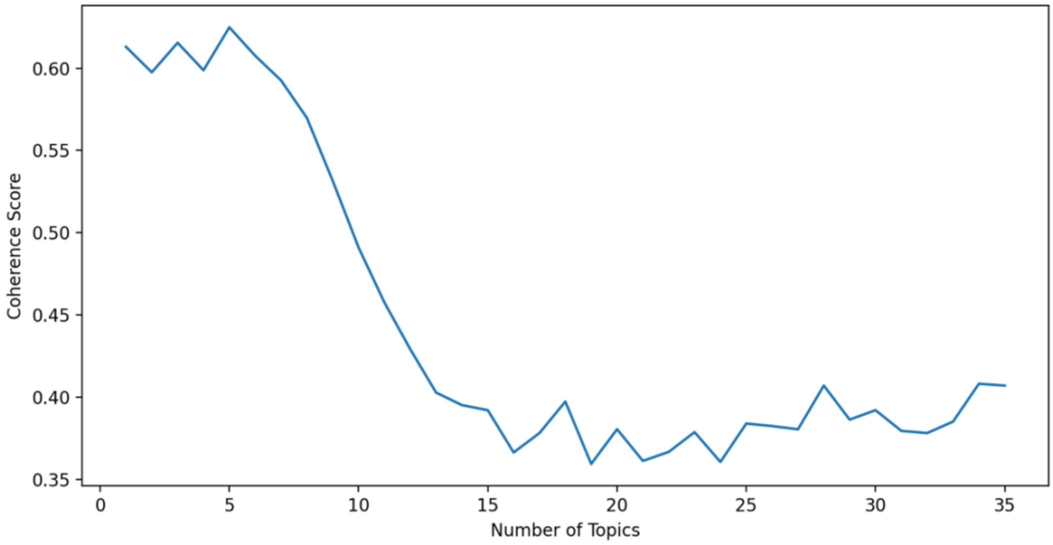


Figure 3-11: Coherence for Stack Overflow with np = 12 k = 5

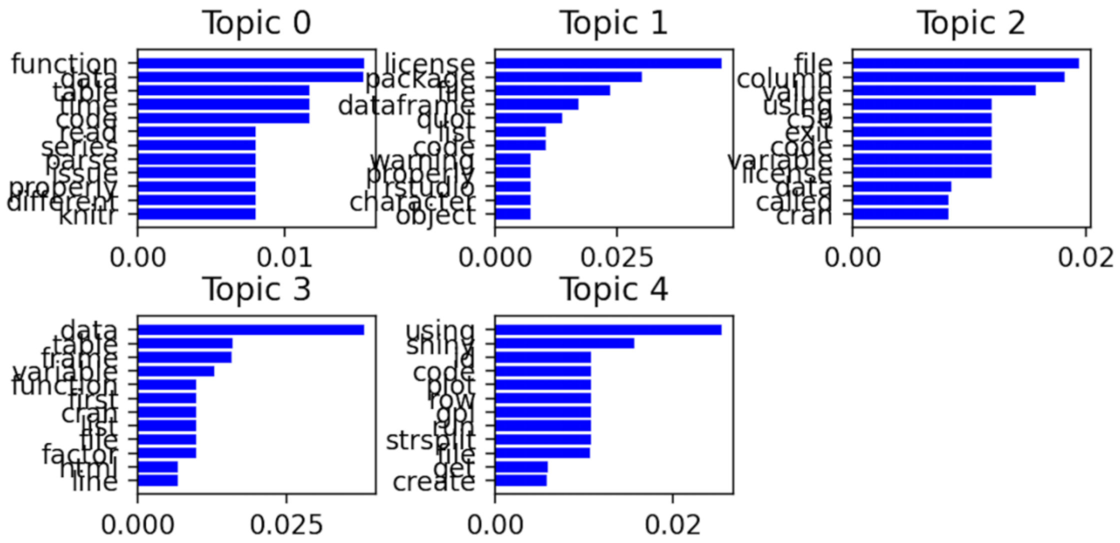


Figure 3-12: LDA topics for Stack Overflow with np = 12 k = 5

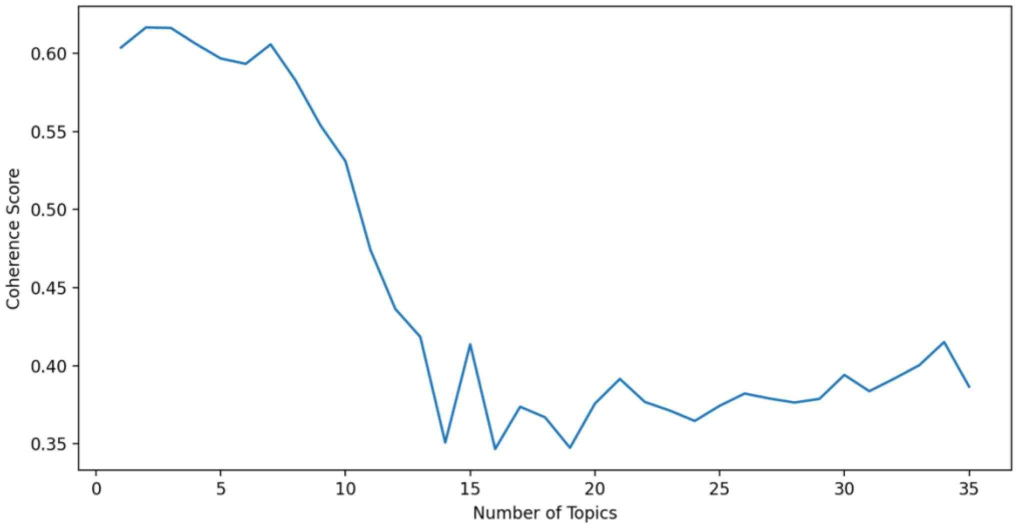


Figure 3-13: Coherence for Stack Overflow with np = 15 k = 5

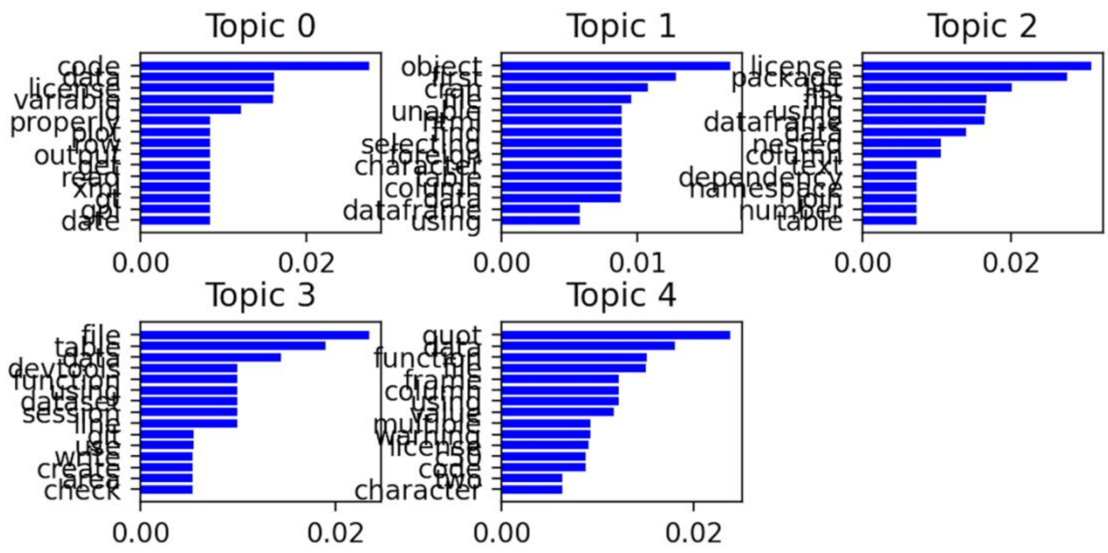


Figure 3-14: LDA topics for Stack Overflow with np = 15 k = 5

3.5 Implementation Details

For all three datasets, R-help archive, Stack Exchange and Stack Overflow, four files were created.

A text file containing the references of licences by every year. A text file containing specific details for the dataset, such as the total references of each license. A csv file with all the entries and the details of each entry. And a csv file containing only the Subject field of the entries.

CHAPTER 4

Analysis and Results

4.1 Statistical analysis and common licenses

Respective to the described methodology and implementation, the experimental approach led to the following results for the first research question (RQ 1).

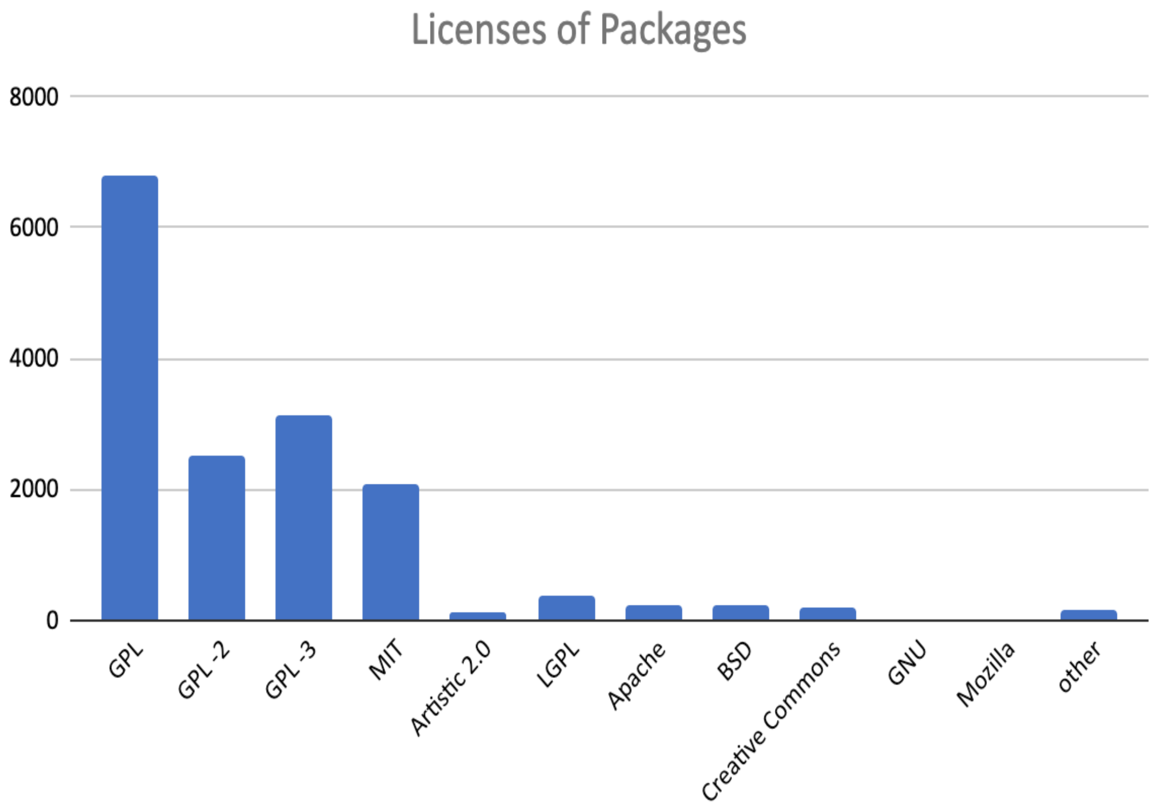


Figure 4-1: Licenses of Packages

License	Count	Percentage
GPL	6814	43.357
GPL -2	2503	15.926
GPL -3	3140	19.980
MIT	2074	13.197
Artistic 2.0	102	0.649
LGPL	378	2.405
Apache	235	1.495
BSD	243	1.546
Creative Commons	202	1.285
GNU	14	0.089
Mozilla	11	0.070
other	163	1.037

Table 4-1: Licenses of Packages

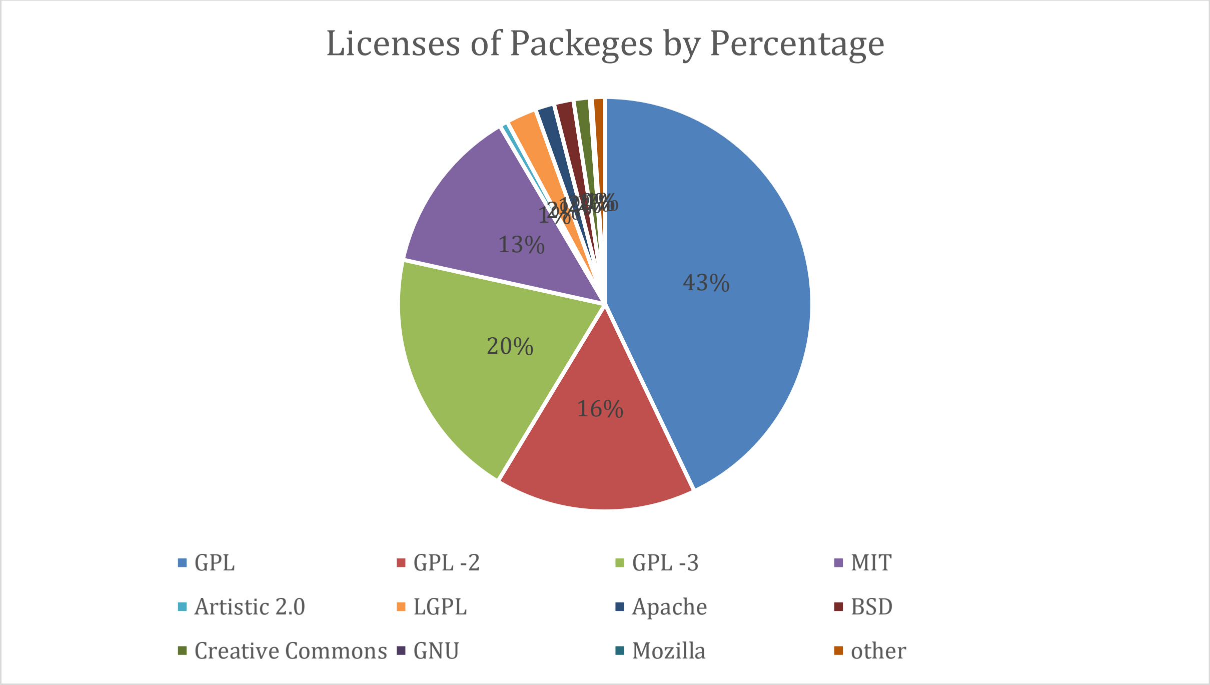


Figure 4-2: Licenses of Packages by Percentage

From the pre-processing on the packages C-ran provides the graphs of Figure 4-1, Figure 4-2 and the data of table 4-1 arise. The “GNU GENERAL PUBLIC LICENSE" (GPL) seems to be the most commonly used licenses (43%) among the developers in the R community, with the “MIT” license (13%) worth mentioning. Unfortunately, it is difficult to find if the developers shifted to another license throughout the versions of the package.

Is it easy to identify that the most popular licenses throughout the years in the R community (RQ 2) are the GPL, GPL-2, GPL-3 and MIT. Although throughout the years the demands and the licenses in the community might have changed, the GPL licence was mostly the number one option for the developers.

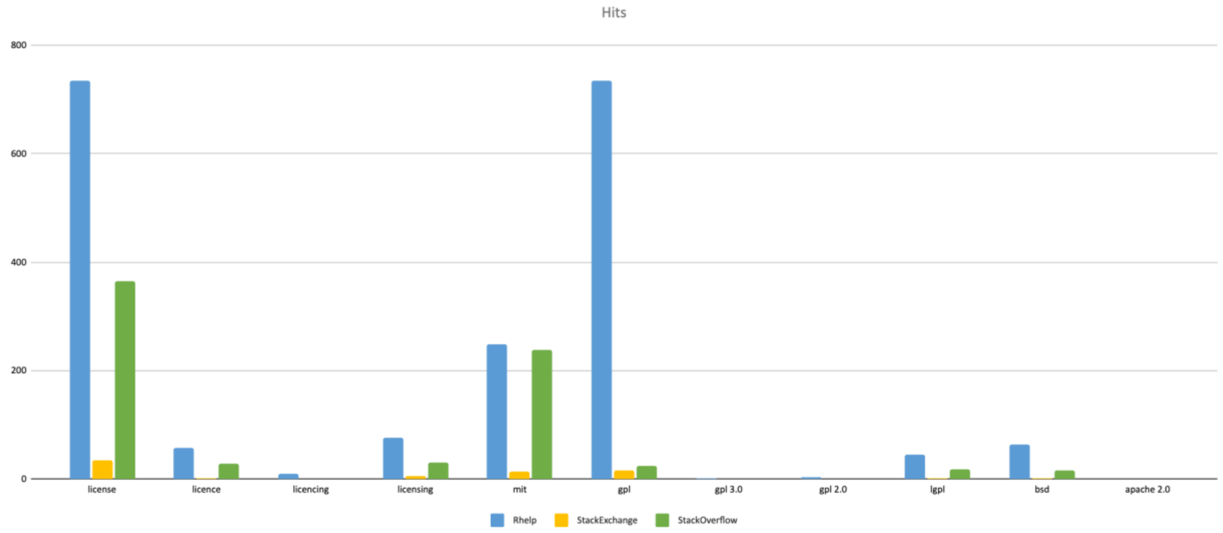


Figure 4-3: Datasets Hits

Keys	Rhelp	StackExchange	StackOverflow
license	734	35	364
licence	58	2	28
licencing	9	0	0
licensing	76	5	30
mit	249	13	237
gpl	735	16	24
gpl 3.0	1	0	0
gpl 2.0	4	0	0
lgpl	45	1	17
bsd	63	2	15
apache 2.0	0	0	0

Table 4-2: Datasets Hits

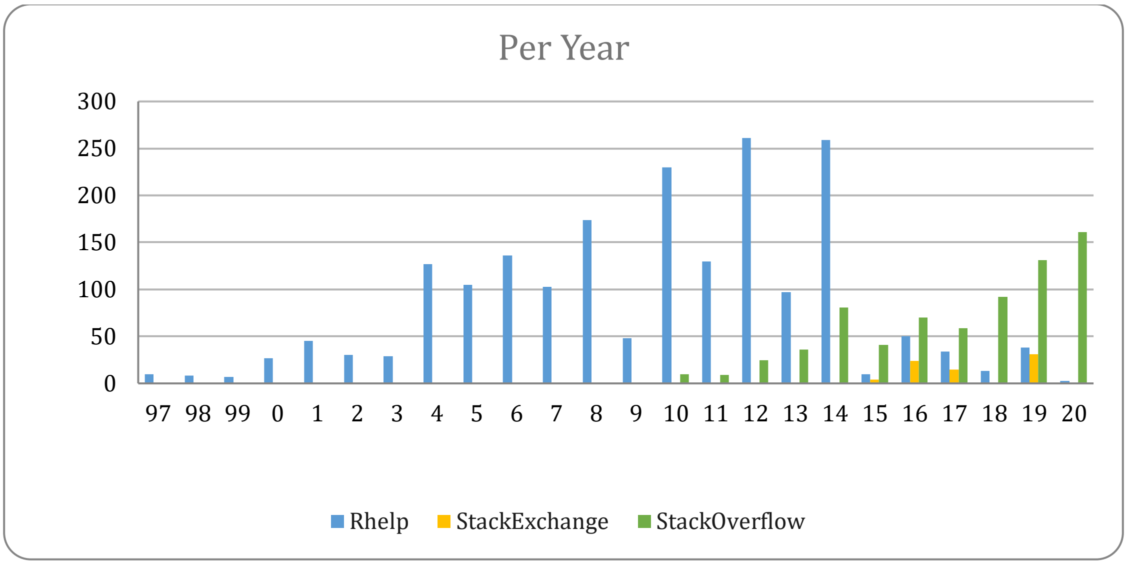
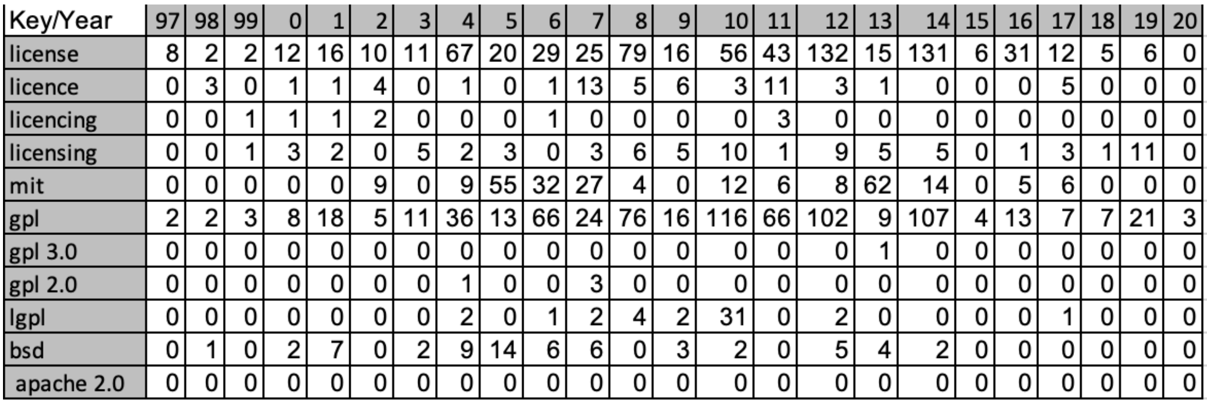


Figure 4-4: Per Year

Year	Rhelp	Stack Exchange	Stack Overflow
97	10	0	0
98	8	0	0
99	7	0	0
0	27	0	0
1	45	0	0
2	30	0	0
3	29	0	0
4	127	0	0
5	105	0	0
6	136	0	0
7	103	0	0
8	174	0	0
9	48	0	0
10	230	0	10
11	130	0	9
12	261	0	25
13	97	0	36
14	259	0	81
15	10	4	41
16	50	24	70
17	34	15	59
18	13	0	92
19	38	31	131
20	3	0	161

Table 4-3: Per Year



Key/Year	97	98	99	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
license	8	2	2	12	16	10	11	67	20	29	25	79	16	56	43	132	15	131	6	31	12	5	6	0
licence	0	3	0	1	1	4	0	1	0	1	13	5	6	3	11	3	1	0	0	0	5	0	0	0
licencing	0	0	1	1	1	2	0	0	0	1	0	0	0	0	3	0	0	0	0	0	0	0	0	0
licensing	0	0	1	3	2	0	5	2	3	0	3	6	5	10	1	9	5	5	0	1	3	1	11	0
mit	0	0	0	0	0	9	0	9	55	32	27	4	0	12	6	8	62	14	0	5	6	0	0	0
gpl	2	2	3	8	18	5	11	36	13	66	24	76	16	116	66	102	9	107	4	13	7	7	21	3
gpl 3.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
gpl 2.0	0	0	0	0	0	0	0	1	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
lgpl	0	0	0	0	0	0	0	2	0	1	2	4	2	31	0	2	0	0	0	0	1	0	0	0
bsd	0	1	0	2	7	0	2	9	14	6	6	0	3	2	0	5	4	2	0	0	0	0	0	0
apache 2.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 4-4: Licences Per Year - Rhelp

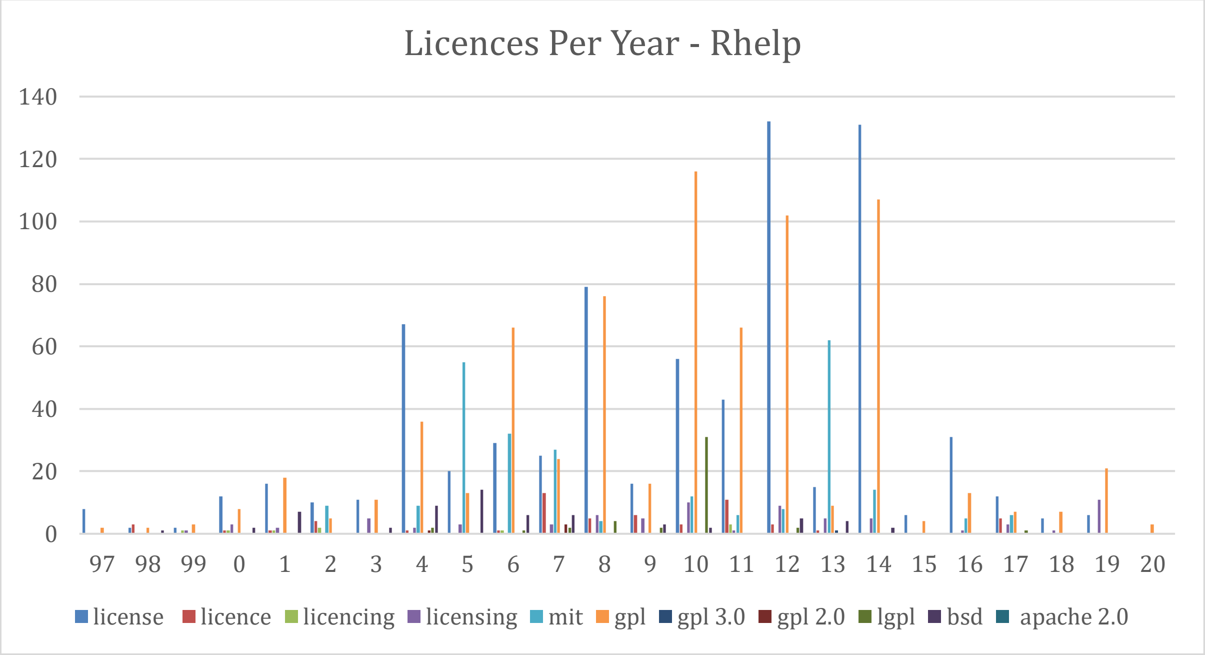


Figure 4-5: Licences Per Year - Rhelp

Key/Year	10	11	12	13	14	15	16	17	18	19	20
license	0	0	0	0	0	2	12	8	0	13	0
Licence	0	0	0	0	0	0	0	0	0	2	0
Licencing	0	0	0	0	0	0	3	2	0	0	0
Licensing	0	0	0	0	0	0	0	0	0	0	0
Mit	0	0	0	0	0	2	4	2	0	5	0
Gpl	0	0	0	0	0	0	4	1	0	11	0
gpl 3.0	0	0	0	0	0	0	0	0	0	0	0
gpl 2.0	0	0	0	0	0	0	0	0	0	0	0
lgpl	0	0	0	0	0	0	1	0	0	0	0
Bsd	0	0	0	0	0	0	0	2	0	0	0
apache 2.0	0	0	0	0	0	0	0	0	0	0	0

Table 4-5: Licences Per Year – Stack Exchange

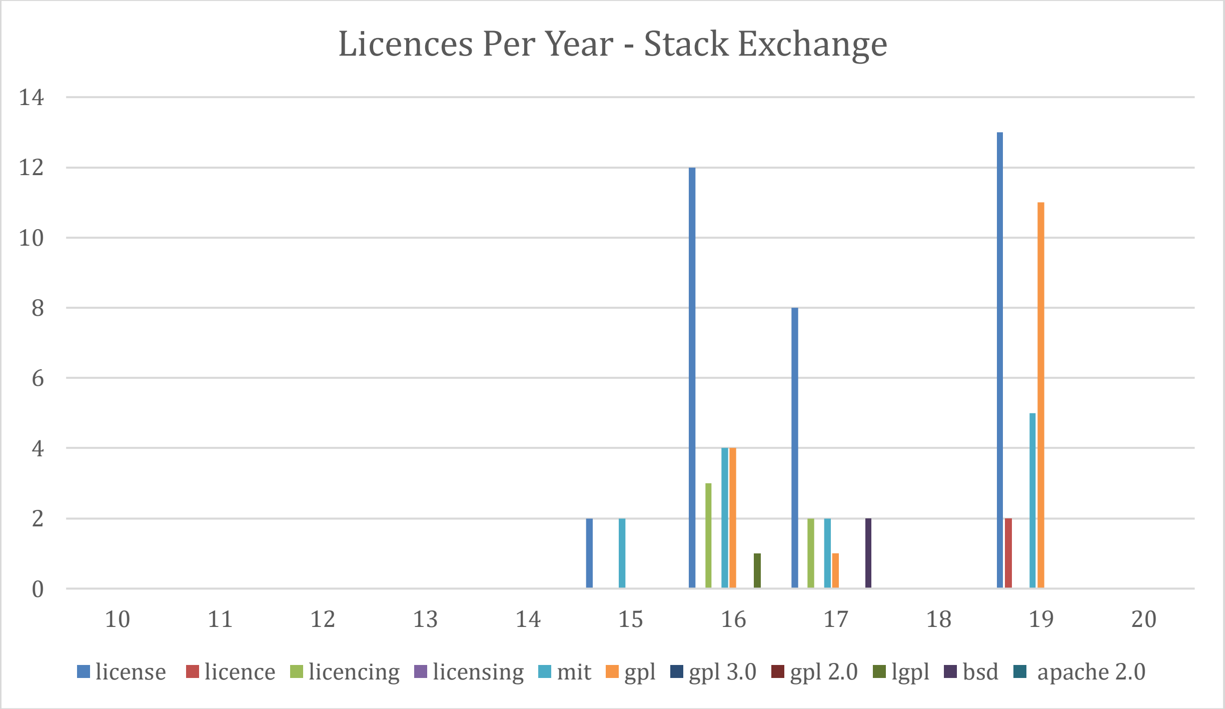


Figure 4-6: Licences Per Year – Stack Exchange

Key/Year	10	11	12	13	14	15	16	17	18	19	20
license	0	3	13	9	41	23	47	27	44	75	82
licence	6	0	1	4	3	0	1	8	0	2	3
licencing	0	0	1	2	2	0	1	3	2	4	15
licensing	0	0	0	0	0	0	0	0	0	0	0
mit	0	4	9	11	14	17	20	21	42	44	55
gpl	1	2	1	4	2	1	1	0	3	3	6
gpl 3.0	0	0	0	0	0	0	0	0	0	0	0
gpl 2.0	0	0	0	0	0	0	0	0	0	0	0
lgpl	1	0	0	3	10	0	0	0	1	2	0
bsd	2	0	0	3	9	0	0	0	0	1	0
apache 2.0	0	0	0	0	0	0	0	0	0	0	0

Table 4-6: Licences Per Year – Stack Overflow

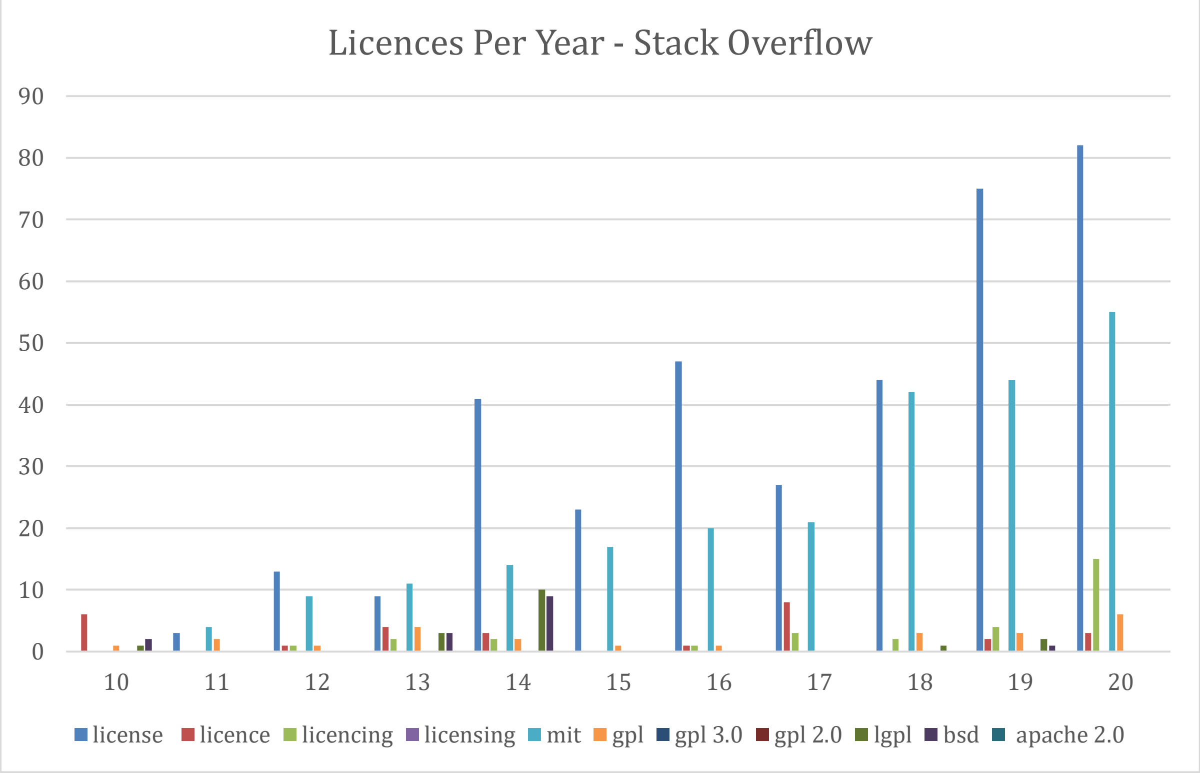


Figure 4-7: Licences Per Year – Stack Overflow

The above figures and tables provide valuable information for the research problem. Firstly, a significant shift in the popularity of the available platforms used for questions from the developers is noted. Until 2014, the most popular platform was the R-help mailing list. After 2014, software development, big data and therefore analysis had a huge growth. Web forums, such as Stack Exchange and Stack Overflow conquered the field of networks of question-and-answer (Q&A) websites. This change is observed on the above graph and table, as Stack Overflow has nearly the same questions about licensing in R as the R-help mailing list the past 5 years. (RQ 3)

4.2 Topic analysis

The following figures were produced by visualising the results of the LDA Topic Modelling algorithm.

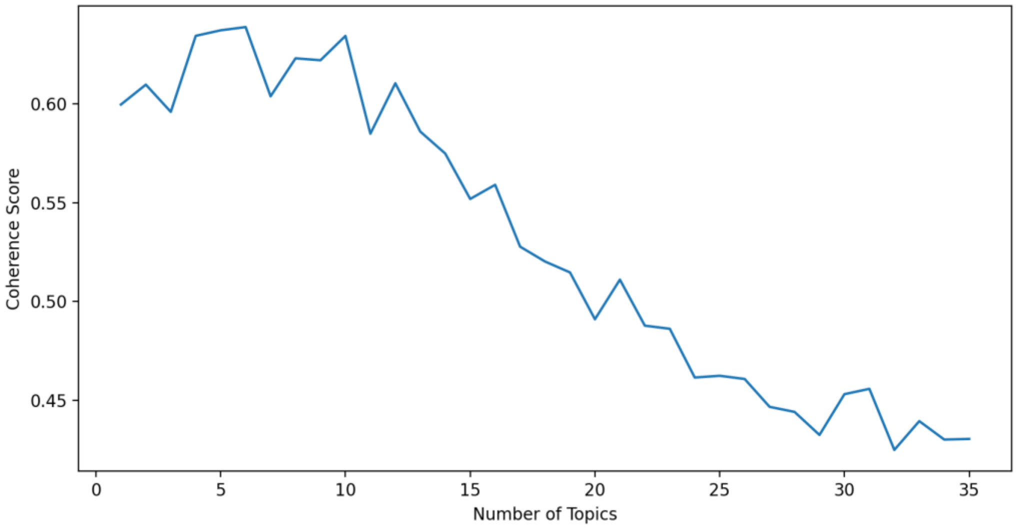


Figure 4-8: LDA Coherence

The Figure 4-8: LDA Coherence represents the coherence score over the number of topics. As the graph shows 10 topics has the highest coherence score with 0.60.

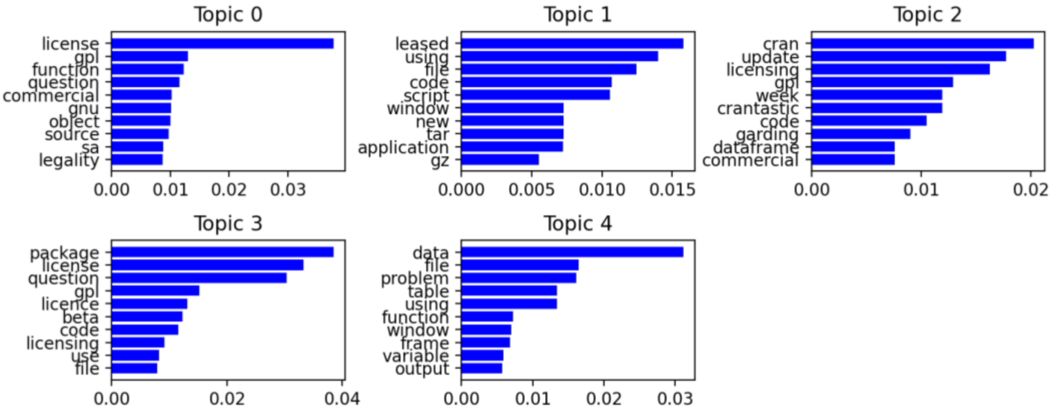
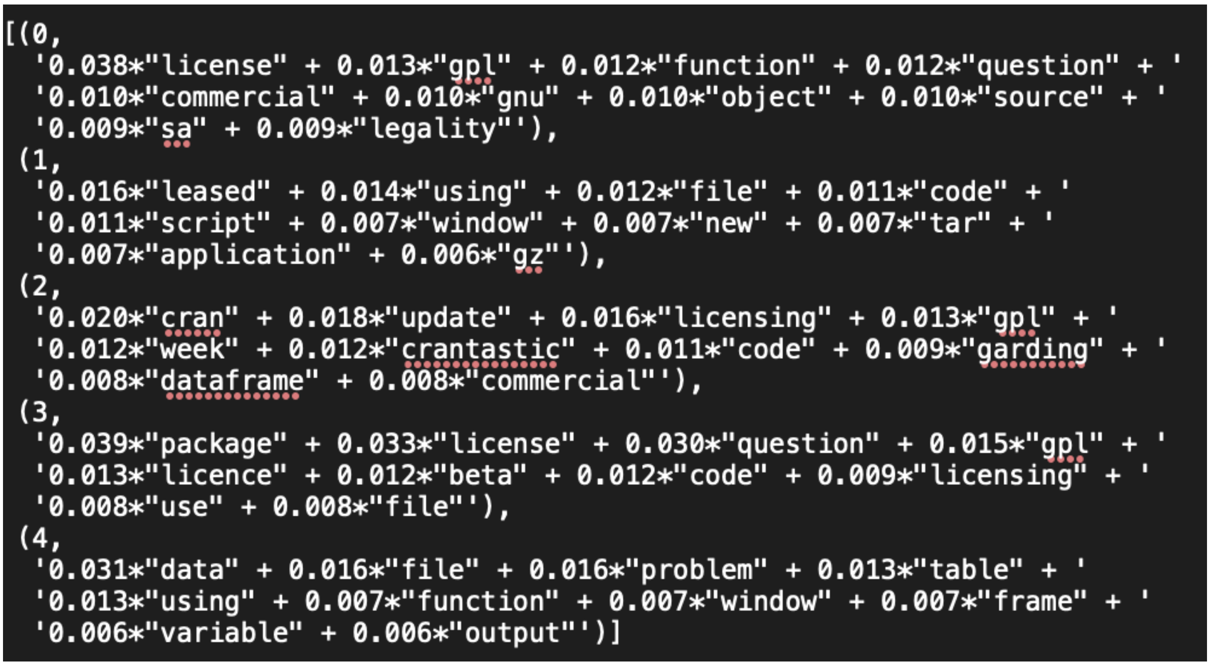


Figure 4-9: LDA topics



```
[(0,
  '0.038*"license" + 0.013*"gpl" + 0.012*"function" + 0.012*"question" + '
  '0.010*"commercial" + 0.010*"gnu" + 0.010*"object" + 0.010*"source" + '
  '0.009*"sa" + 0.009*"legality"'),
 (1,
  '0.016*"leased" + 0.014*"using" + 0.012*"file" + 0.011*"code" + '
  '0.011*"script" + 0.007*"window" + 0.007*"new" + 0.007*"tar" + '
  '0.007*"application" + 0.006*"gz"'),
 (2,
  '0.020*"cran" + 0.018*"update" + 0.016*"licensing" + 0.013*"gpl" + '
  '0.012*"week" + 0.012*"crantastic" + 0.011*"code" + 0.009*"garding" + '
  '0.008*"dataframe" + 0.008*"commercial"'),
 (3,
  '0.039*"package" + 0.033*"license" + 0.030*"question" + 0.015*"gpl" + '
  '0.013*"licence" + 0.012*"beta" + 0.012*"code" + 0.009*"licensing" + '
  '0.008*"use" + 0.008*"file"'),
 (4,
  '0.031*"data" + 0.016*"file" + 0.016*"problem" + 0.013*"table" + '
  '0.013*"using" + 0.007*"function" + 0.007*"window" + 0.007*"frame" + '
  '0.006*"variable" + 0.006*"output"')]
```

Figure 4-10: LDA topics specifics

The figure 4-9: LDA topics displays the 5 topics that the LDA algorithm produced. As the figure 4-10: LDA topics specifics shows the most commonly asked questions and references throughout the sources regards packages, updating libraries, new versions of packages and errors occurring at installation or use of packages. It is important to notice that the words “licence”, “license” and “GPL” are common on most of the topics, confirming that the GPL licence is the most popular. (RQ 4)

Although the plausibility of the topics is uncertain, we can see the substantive match that follows. Topic 0 deals with questions about license and legality, such as the licenses GPL and GNU. Topic 1 deals with questions about files, code and scripts. Topic 2 deals with questions about updating CRAN packages and licensing. Topic 3 deals with questions about licensing packages. Topic 4 deals with questions about problems with data and tables.

More specifically, the topics 0-4 can be split into two categories. The first category is Topic 0, Topic 2 and Topic 3, where the subject is around “license”, more notably “GPL” and “GNU” licenses. Also, the word “legality” in Topic 0 has 0.009 relevance metric, that suggest it is a frequent discussion. The second category is Topic 1 and Topic 4, where the subject is around problems in coding and scripts.

4.3 Implications for practitioners

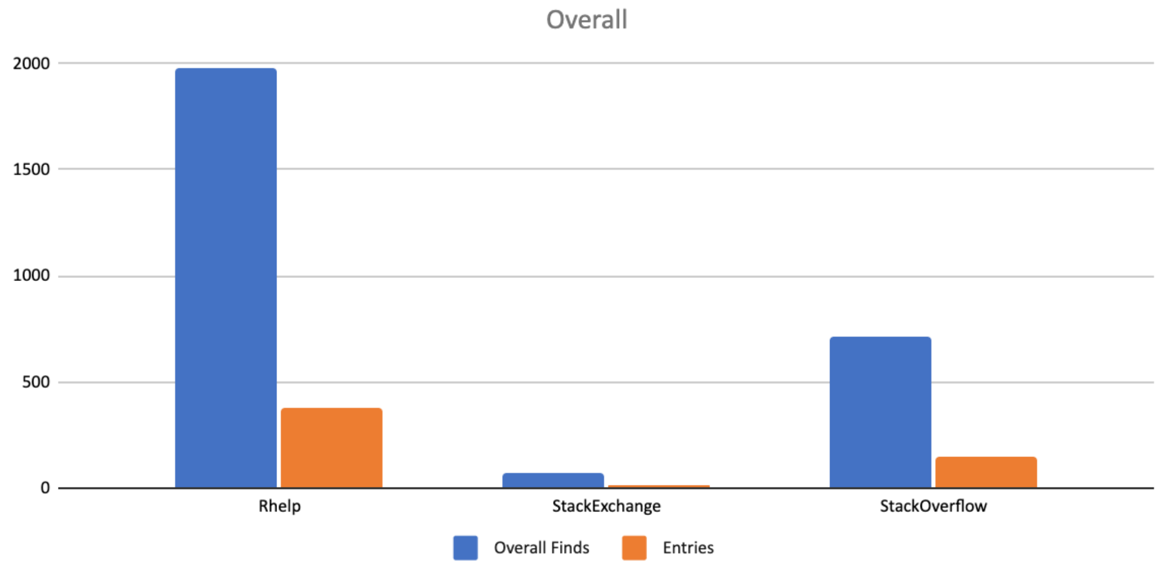


Figure 4-11: Overall Finds

	Rhelp	StackExchange	StackOverflow
Overall References	1974	74	715
Entries	377	15	146

Table 4-7: Overall Finds

Overall, as the above graph and table suggest, licensing as a topic among the R community is not very popular. From hundreds of thousands of questions asked throughout the years and the platforms, only approximately 500 of them were related to licencing and licences. (RQ 5)

CHAPTER 5

Conclusions and Future Work

5.1 Conclusions

The process of reviewing different publications in the field of Big Data and Data Analysis, as well as the process of designing the different algorithms needed for the project's objectives to be accomplished, has led me to draw several conclusions about the field, the features implemented and the results from them.

The system that was developed has successfully fulfilled the goals of the project. The system indicated the most common licenses throughout the years in the R community. It also indicated how the community shifted to other platforms throughout the years and if the “licensing” as a topic is popular among the R community.

Throughout the dissertation, the immense importance of time and space efficiency when having to deal with such large quantities of data and multiple datasets with various structures was discussed many times. In order to produce as good performance and as precise results as possible, many parts of the code and many techniques used were changed several times.

It is important to notice that every single component was developed in such a way so that it can be executed individually, or it can be used as a standalone script in any other Python or Java project.

In this project, I got the opportunity to experience the science of Data Analysis at its base form. I got to fetch, clear, process and analyse the data from scratch. I also had the opportunity to interact with many powerful and state-of-the-art tools. It has helped me to build an important skill set that will certainly be of great benefit in my career by working with vast amounts of real data and gaining valuable experience in this area of Computer Science that has skyrocketed in the last few years.

5.2 Future Work

The outcome of this project is satisfying and the goals that were set have been met. But this project can be expanded and improve in many ways.

First of all, there are many other sources to examine. Some of them are other mailing lists the C-RAN provides (https://www.mail-archive.com/r-help@r-project.org/) . Also, it could be fruitful to examine the answers posted at the different sources as well, as some users might refer to licensing topics.

In addition, many NLP techniques are still only available in English (like the Lemmatization). This has forced us to translate text written in foreign languages, but that is not an optimal solution, as even the best translating systems (like Google Translate) many times provide very inaccurate translations.

Last but not least, due to the small number of topics related to licensing, the results of the LDA Topic Model Algorithm were not as successful. With the addition of data from other sources the results will be improved.

Bibliography

[1] Definitions, https://en.wikipedia.org/wiki/R_(programming_language)#Communities
https://en.wikipedia.org/wiki/Software_license
https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation
https://en.wikipedia.org/wiki/Tokenization_(data_security)

[2] The Comprehensive R Archive Network (CRAN), https://cran.r-project.org/

[3] The R Project for Statistical Computing, https://www.r-project.org/help.html

[4] CRAN Mailing Lists, https://www.r-project.org/mail.html

[5] Frequently Asked Questions on R (R FAQ), https://cran.r-project.org/doc/FAQ/R-FAQ.html

[6] R Licenses, https://www.r-project.org/Licenses/

[7] The R-help Archives, https://stat.ethz.ch/pipermail/r-help/

[8] The Mail Archive, https://www.mail-archive.com/r-help@r-project.org/.

[9] Stack Exchange, https://stackexchange.com/

[10] Topic Modeling Tutorial with Latent Dirichlet Allocation (LDA),
https://towardsdatascience.com/topic-modeling-with-latent-dirichlet-allocation-by-example-3b22cd10c835

[11] Evaluate Topic Models: Latent Dirichlet Allocation (LDA),
https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

[12] Documentation of sklearn LDA algorithm, https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

[13] Quick list of useful R packages, https://support.rstudio.com/hc/en-us/articles/201057987-Quick-list-of-useful-R-packages

[14] Analyzing Measurements of the R Statistical Open Source Software, https://ieeexplore.ieee.org/document/6479797

[15] License usage and changes: a large-scale study on gitHub, https://link.springer.com/article/10.1007/s10664-016-9438-4

[16] How the R community creates and curates knowledge: an extended study of stack overflow and mailing lists, https://link.springer.com/article/10.1007%2Fs10664-017-9536-y

[17] Georgia M. Kapitsaki, Maria Papoutsoglou, Daniel M. Germán, Lefteris Angelis: What do developers talk about open source software licensing? SEAA 2020: 72-79

List of Figures

List of Tables