SEMESTER 2, 2022/2023

STQD6114 – ANALITIK DATA TAK BERSTRUKTUR

Analitik data tak berstruktur

LECTURER:

DR. NOR HAMIZAH BINTI MISWAN

Name	Matric Number
An Hongyu	P120996

CHAPTER I

INTRODUCTION

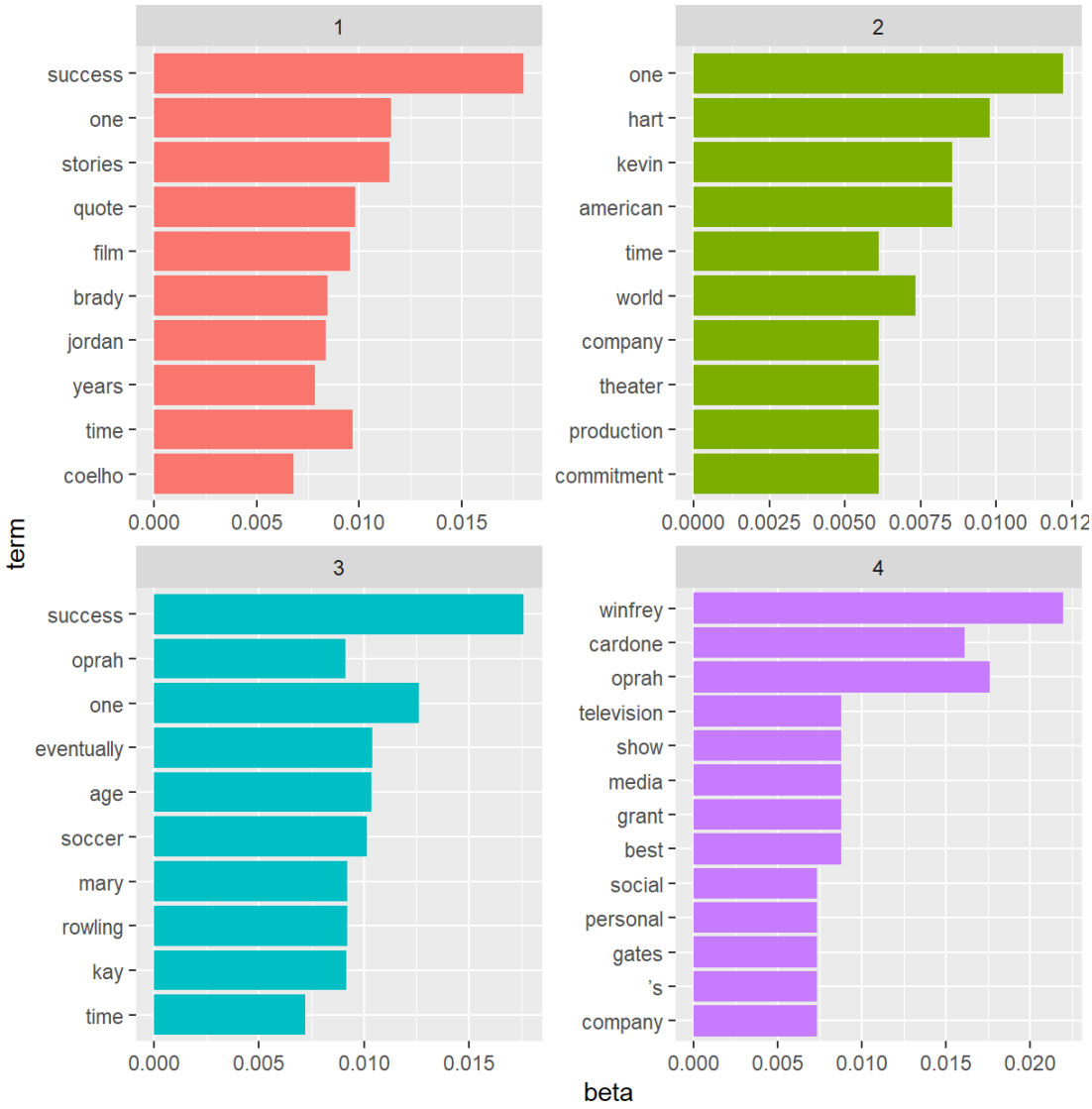
Success stories have always captivated our attention and inspired us to dream big, persevere through challenges, and achieve greatness. These tales of triumph and accomplishment serve as beacons of hope, reminding us that with dedication, determination, and a sprinkle of good fortune, incredible achievements are within our reach.

Throughout history, individuals from various walks of life have embarked on extraordinary journeys that have led them to remarkable success. From visionary entrepreneurs who have transformed industries with their groundbreaking innovations, to talented artists who have touched our hearts and minds with their creativity, and inspiring activists who have fought for justice and equality, success stories come in many forms and carry immense power.

By conducting a comprehensive topic modeling analysis using LDA, we can gain a deeper understanding of the underlying themes and patterns within the acquired dataset. So we need to build an LDA model to analyze these character stories to draw a key conclusion about what success is related to.

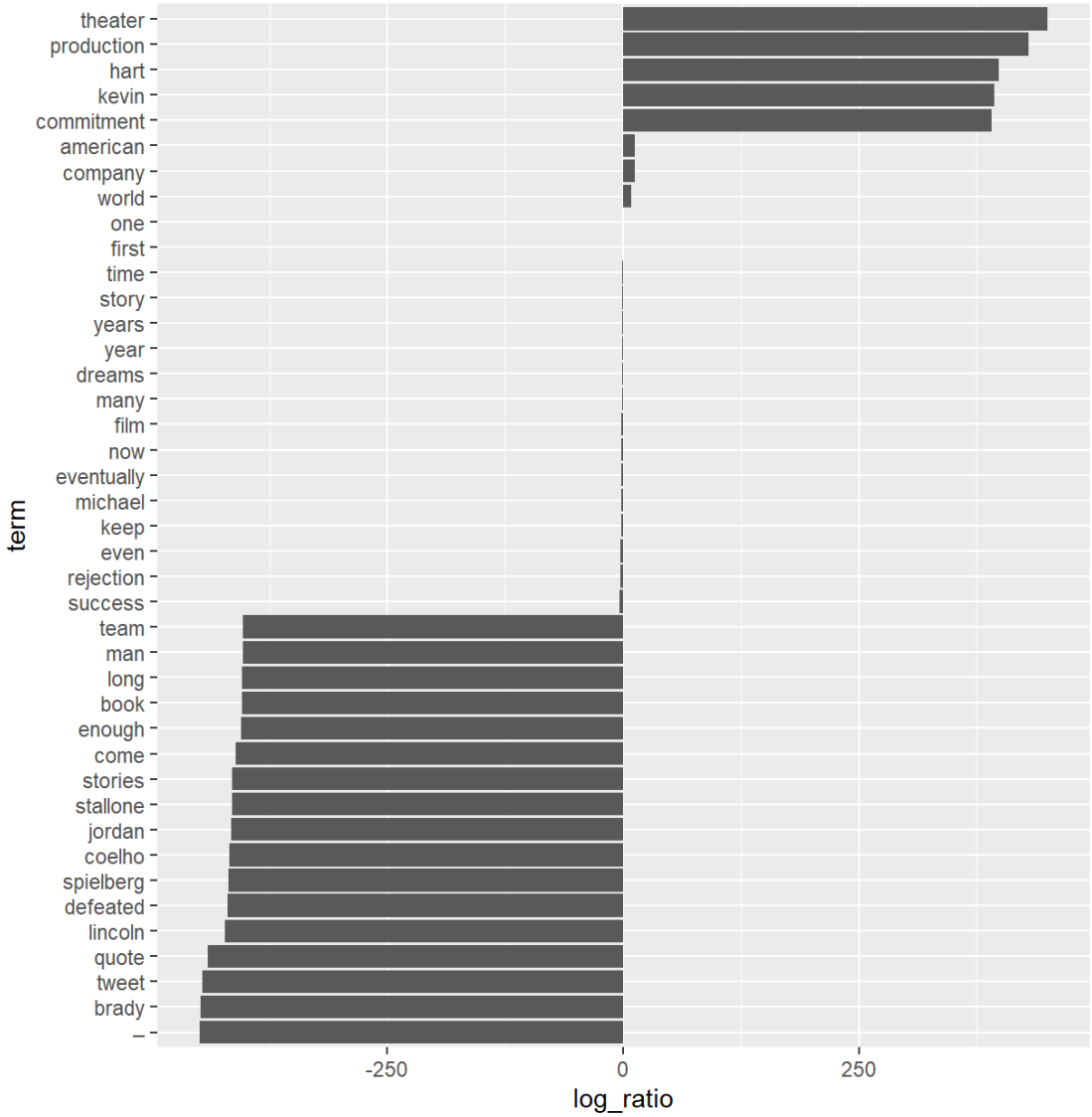
CHAPTER II

CONTENTS



This visualization lets us understand the four topics that were extracted from the articles. The most common words in topic 1 include “success”, “stories”, “film”, and “jordan”, which suggests it may represent film or Success stories news. Those most common in topic 2 include “kevin”, “american”, and “theater”, suggesting that this topic represents

American famous successful news. Those most common in topic 3 include “success”, “age”, and “soccer”, suggesting that this topic represents sporter famous successful news. Those most common in topic 4 include “winfrey”, “television”, and “media”, suggesting that this topic represents American famous successful televison news. One important observation about the words in each topic is that some words, such as “success” and “one”, are common within both topics.



From this figure, we can see that the biggest difference is the name and field of different people, such as movies, books, and sports. The smallest difference is success. So we can conclude that the topic is about the success stories of celebrities in different fields.

```
   document topic     gamma
   <chr>    <int>     <dbl>
 1 1.txt        1 1.00
 2 10.txt       1 1.00
 3 11.txt       1 1.00
 4 12.txt       1 0.568
 5 13.txt       1 0.000146
 6 14.txt       1 0.0000811
 7 15.txt       1 0.000175
 8 16.txt       1 0.000142
 9 17.txt       1 0.000479
10 18.txt       1 0.000552
# … with 130 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```
> findAssocs(tdm, terms = c("success"), corlimit = 0.75)
$success
       stories          quote            abe       accolades        achieve"      achieving        actually
          0.99           0.99           0.99            0.99            0.99           0.99            0.99
        adlib            afc          agree           ahead       alldecade      ambition"             arm
          0.99           0.99           0.99            0.99            0.99           0.99            0.99
      arrived        athlete       athletic           avoid           begin      believing       biproduct
          0.99           0.99           0.99            0.99            0.99           0.99            0.99
         bowl          brady         brady's           build            case         chance           chart
          0.99           0.99           0.99            0.99            0.99           0.99            0.99
        coach       combined          come"          common     controlling     culmination           debut
          0.99           0.99           0.99            0.99            0.99           0.99            0.99
  definitions          depth       discover           draft          easily      encourage         exposed
          0.99           0.99           0.99            0.99            0.99           0.99            0.99
   extinguish      failures"       failure"         fashion        favorite          field          follow
          0.99           0.99           0.99            0.99            0.99           0.99            0.99
         grid        groomed        grossly           heart          heroic     highlighting          indeed
          0.99           0.99           0.99            0.99            0.99           0.99            0.99
interception       interest           iron          junior         knocked          lacks        laughter
          0.99           0.99           0.99            0.99            0.99           0.99            0.99
      league         little         living          long"       mechanics       michigan        mobility
          0.99           0.99           0.99            0.99            0.99           0.99            0.99
         nfc            nfl           nope         numbers            onto         passes            path
          0.99           0.99           0.99            0.99            0.99           0.99            0.99
  persisting       physical           pick          plenty        position        prepare     progressive
          0.99           0.99           0.99            0.99            0.99           0.99            0.99
        pros         proved        purpose          pushed       qualities        quarter     quarterback
          0.99           0.99           0.99            0.99            0.99           0.99            0.99
quarterbacks         quotes          rated           ratio           reach    realization         related
          0.99           0.99           0.99            0.99            0.99           0.99            0.99
      report     requesting          round            rush           seems        sharing           sheer
          0.99           0.99           0.99            0.99            0.99           0.99            0.99
      skinny          sleep         sleep"         someday         special      stallone's        starting
```

As can be seen from the figure above, the word with the highest frequency of success has failure, indicating that failure is the mother of success. Failure is inevitable on the road to success.

CHAPTER III

CONCLUSION

In this analysis, we performed topic modeling using Latent Dirichlet Allocation (LDA) on a dataset consisting of text data. The objective was to identify and explore four distinct topics within the dataset. In conclusion, the topic modeling analysis using LDA allowed us to uncover hidden themes and patterns within the dataset. Through the extraction of per-topic per-word probabilities, visualization of common terms, beta spread analysis, and per-document per-topic probabilities, we gained a comprehensive understanding of the topics and their relationships to the documents. In the end we concluded that the word most closely associated with success is failure.