Text Mining South Park

Dominik Nerger (i6146759)

June 2, 2017

Contents

1 Introduction **1**

2 Data set **1**

3 Libraries **6**

4 Pre-processing **6**

5 Implementation and results **6**

6 Conclusion **6**

7 Future work **6**

1 Introduction

The repository is available on GitHub[1].

2 Data set

The data set spans from seasons **1** to **18**, adding up to 257 episodes overall with a file size of 5.41MB. It contains 70896 rows, with each row possessing information about the **season**, **episode**, **character** and **line**.

The data set has been crawled by Bob Adams and is available to download on GitHub[2]. It has been assembled by crawling the South Park Archives[3].

The code of GitHub repository is not available to the public. An attempt at a crawl is available in the GitHub repository.

[1] `https://github.com/dnerger/South-Park-Text-Mining`
[2] `https://github.com/BobAdamsEE/SouthParkData`
[3] `http://southpark.wikia.com/wiki/South_Park_Archives`

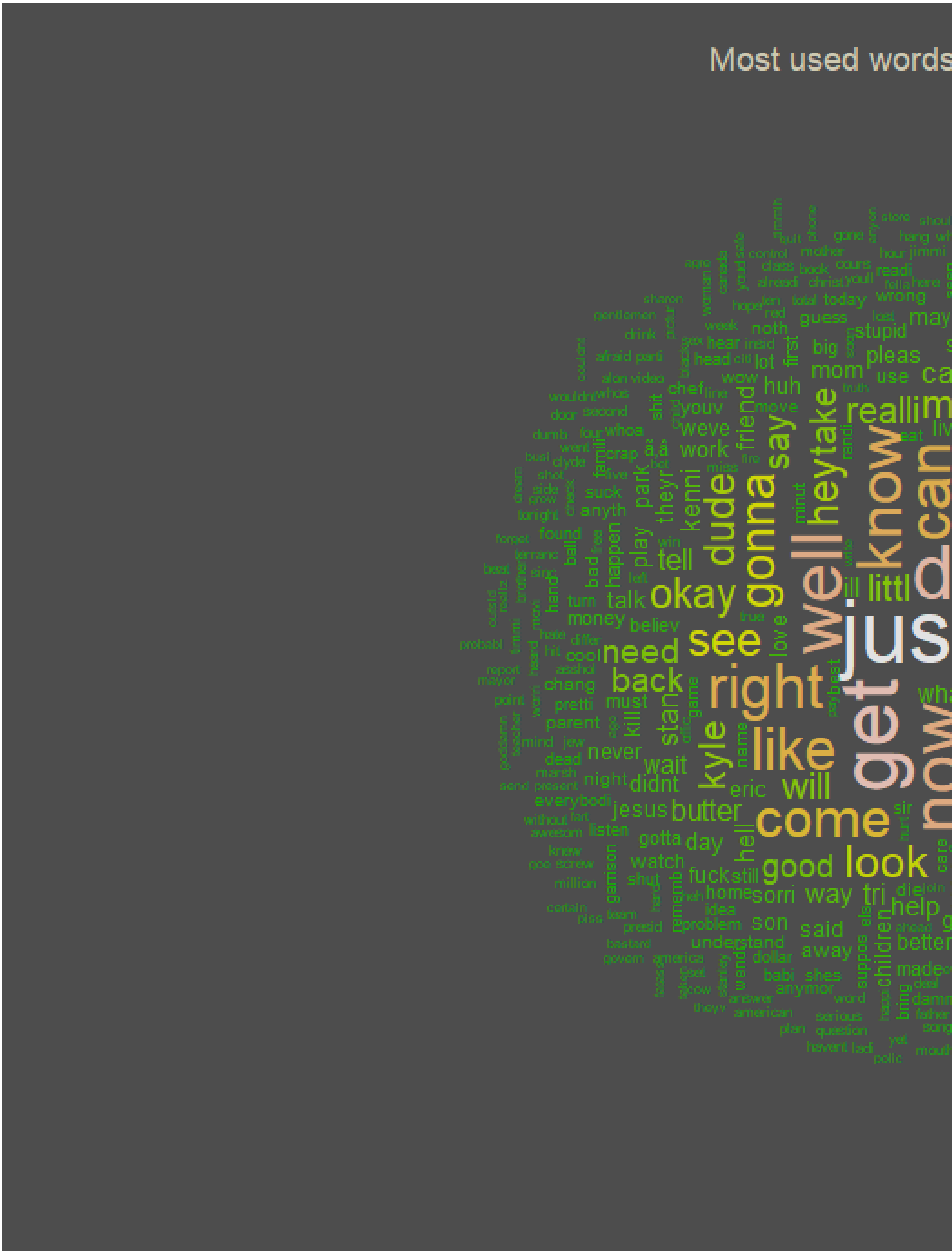


Figure 1: General wordcloud, containing terms by frequency

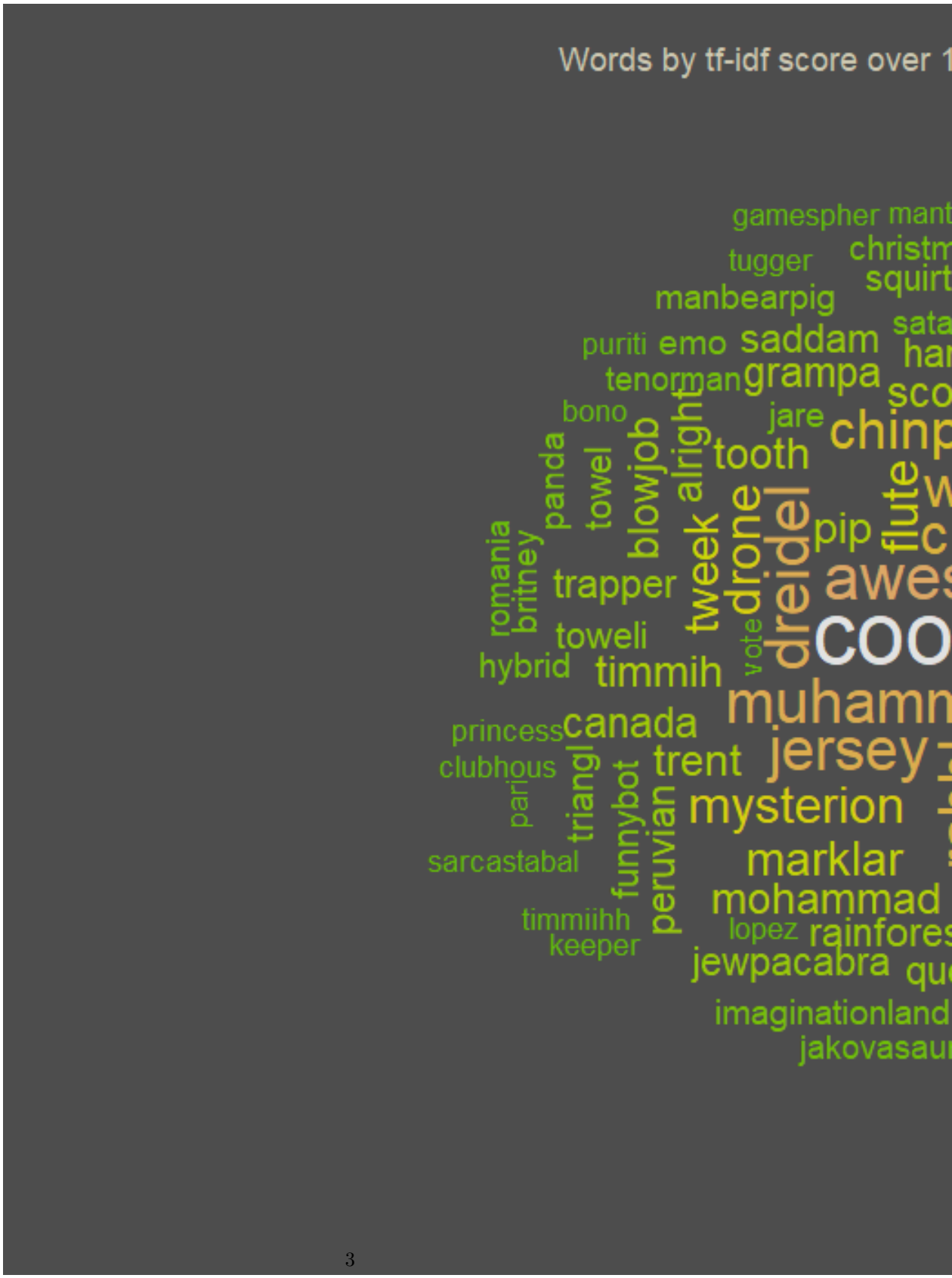


Figure 2: General wordcloud, containing terms by TF-IDF score

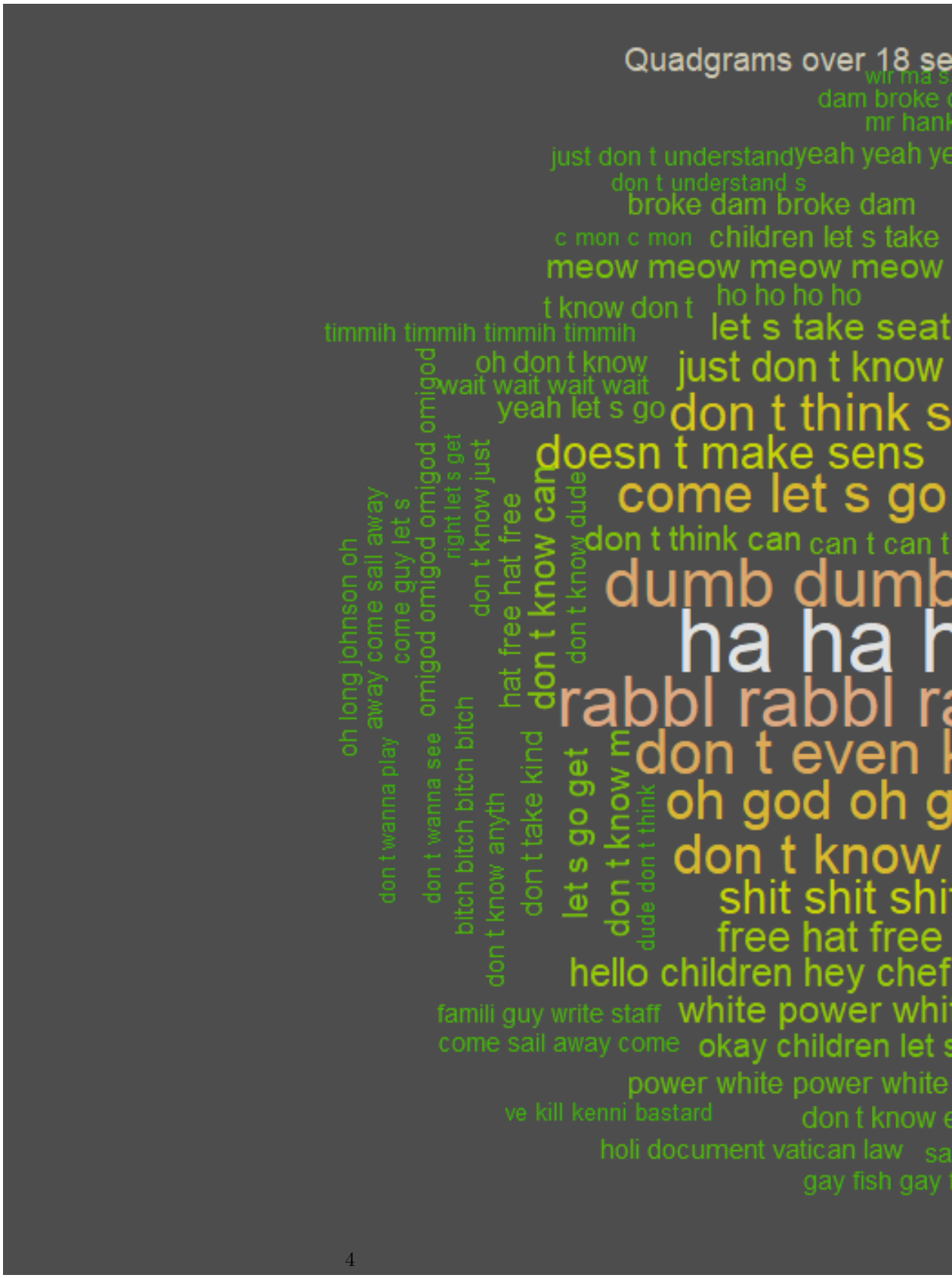
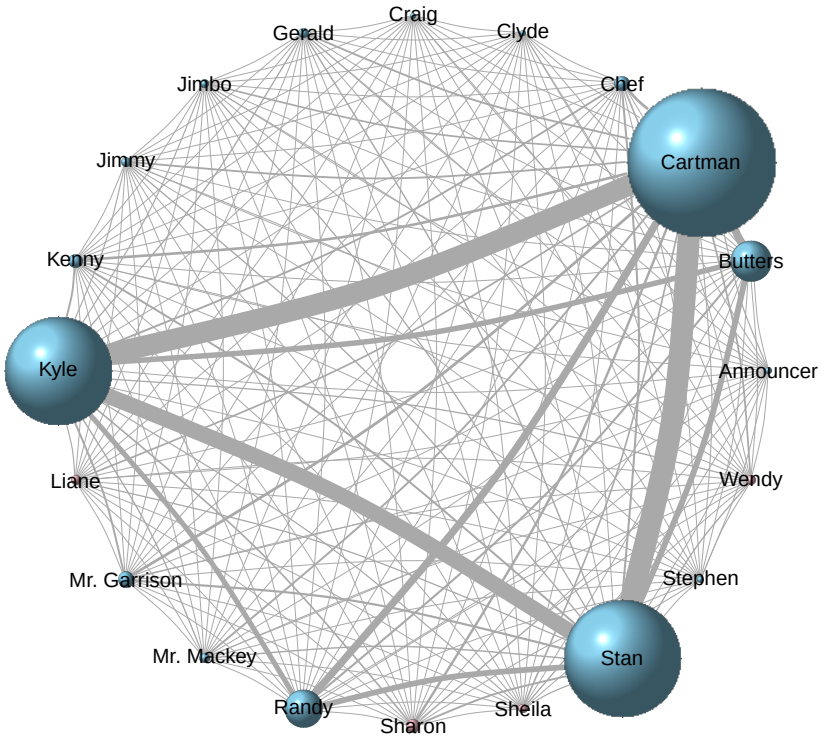


Figure 3: Wordcloud containing ngrams (n=4) by frequency



Matrix.pdf

Figure 4: Co-occurences of characters over 257 episodes, size of vertex in relation to amount of lines

3 Libraries

All scripts have been programmed in *R*. To execute the scripts, R needs to be installed. To view temporary files that are executed during runtime, e.g. the corpus or a TermDocumentMatrix, it is advised to install RStudio. The libraries necessary for each script are imported at the top of each script, if they are not installed they can be installed by executing:

```
install.packages("library-name")
```

In the following, all libraries that are related to Text Mining techniques will be introduced.

The library **tm** is the Text Mining package of R, which enables pre-processing of data sets and allows to build the corpus. **RWeka** is a collection of machine learning algorithms for data mining tasks. **NMF** introduces the Non-negative Matrix Factorization to R. **NLP** and **OpenNLP** are libraries that provide Natural Language Processing techniques and are used for NER-Tagging. **syuzhet** extracts sentiments from text and contains the three sentiment dictionaries *bing*, *afinn* and *nrc*. The package **stm** is used for Structural Topic Modeling which is LDA with additional met-data and can be visualized using the package **LDAvis**. Libraries used for visualization include **igraph**,**ggplot2**, **ggraph**, **viridis** and **wordcloud**.

4 Pre-processing

5 Implementation and results

6 Conclusion

In conclusion,

7 Future work