Maastricht University Leading in Learning!

TEXT MINING

DOMINIK NERGER

Overview

Overview

Introduction

- Analyze South Park
  - Focus on Season 18
- Answer **who, what, when, why**
- Programming language: R
- Libraries:
  - TM (Text Mining library)
  - NLP/Apache OpenNLP (NER-Tagging)
  - smf (Structural Topic Model)
  - and many more

Overview

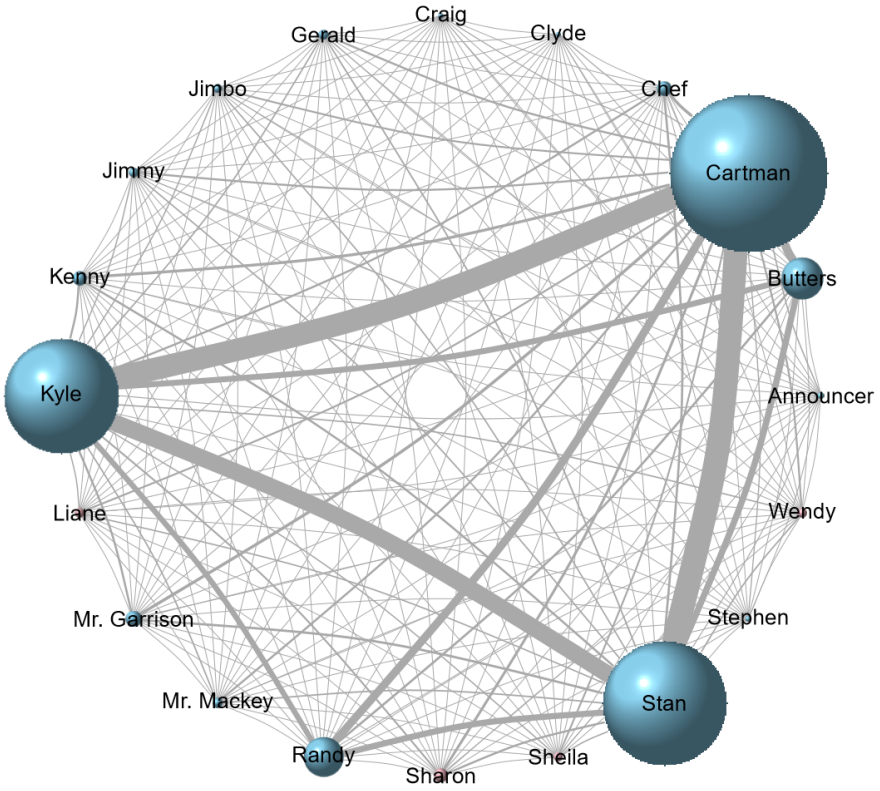
Data set & pre-processing

- 18 seasons, 257 episodes, 70879 Lines
- Reduced size of Corpus by 30% after pre-processing
- Crawled from: http://southpark.wikia.com
- Worked on own crawl, filtering necessary
  - Tables change between seasons
  - Unnecessary information
- Provided by: https://github.com/BobAdamsEE/SouthParkData

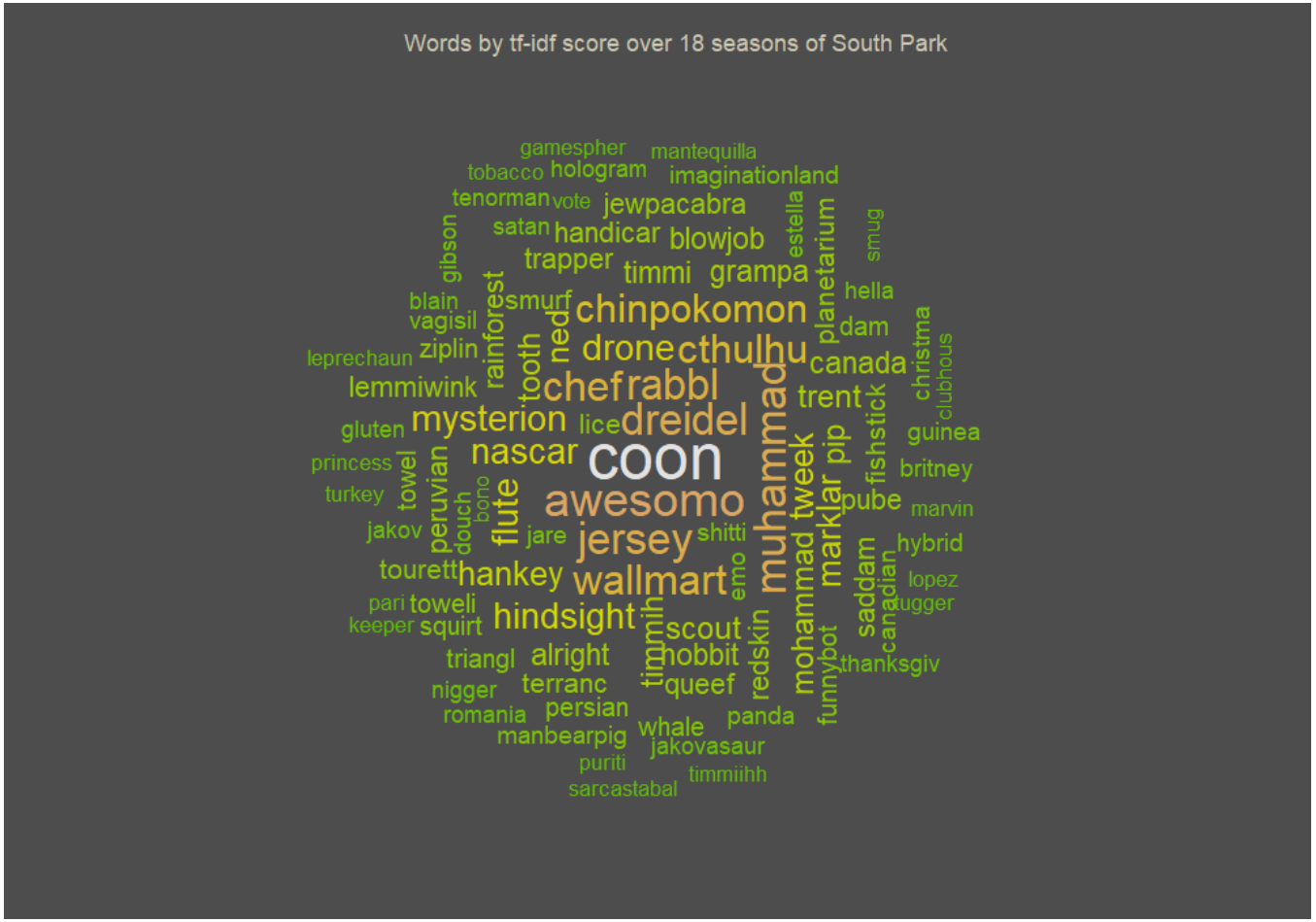
Overview

Overview

Co-occurence of Characters



Wordcloud TF-IDF



Overview

Named Entity Recognition

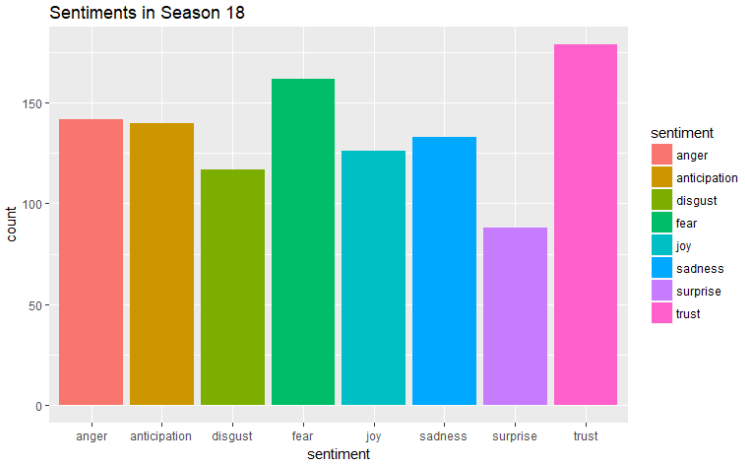
- Annotated Sentiments & Entities (Persons, Locations, Organizations)
- Entities for Season 18 Episode 4:

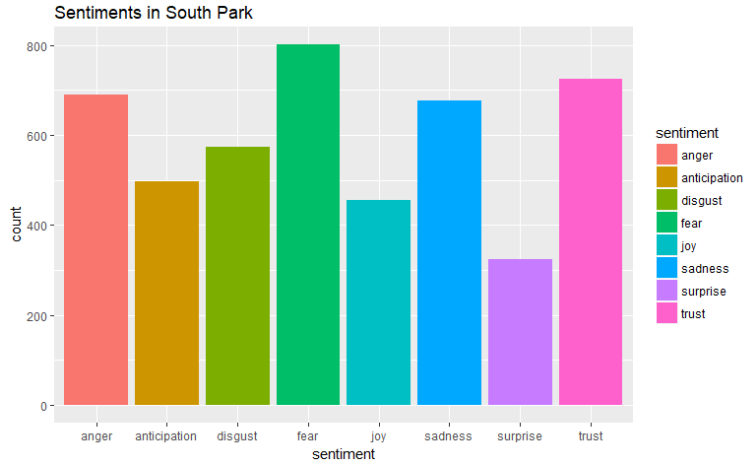
```
 [1] "Jimmy"              "Francis"            "Jimmy Valmer"       "Timmy Burch"
 [5] "Nathan"             "Gerald"             "Gerald"             "Matthew McConaughey"
 [9] "Ah"                 "I"                  "Shinzo Abe"         "Randy"
[13] "Matthew McConaughey" "Neve Campbell"      "Timmy Burch"        "Dick Dasterdly"
[17] "Applegate"          "Wait"               "Matthew McConaughey" "Jimmy Fallon"
[21] "Dick Dastardly"     "Matthew McConaughey" "Matthew McConaughey" "Wait"
[25] "Mom"                "Ohhh"               "Lincoln"            "Colorado"
[29] "Japan"              "Salzburg"           "San Francisco"      "Japan"
[33] "South Park"         "GPS"                "If"                 "Wacky Races"
[37] "Oh"                 "BBC World"          "Canadian"           "Handicar"
```

Overview

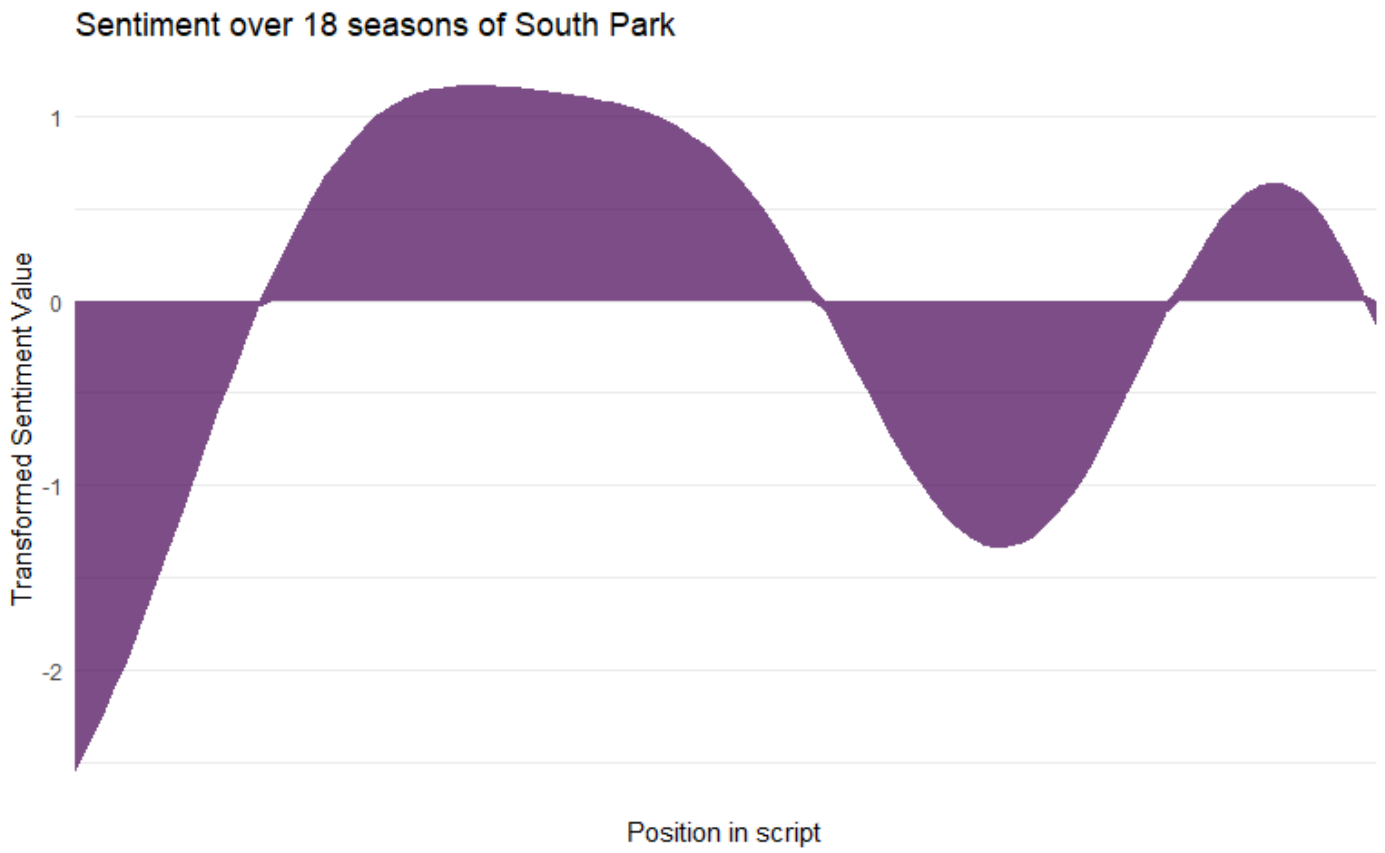
Sentiment Mining

- NRC Lexicon
- Positive/negative sentiment timeline
- Aggregation of other sentiments





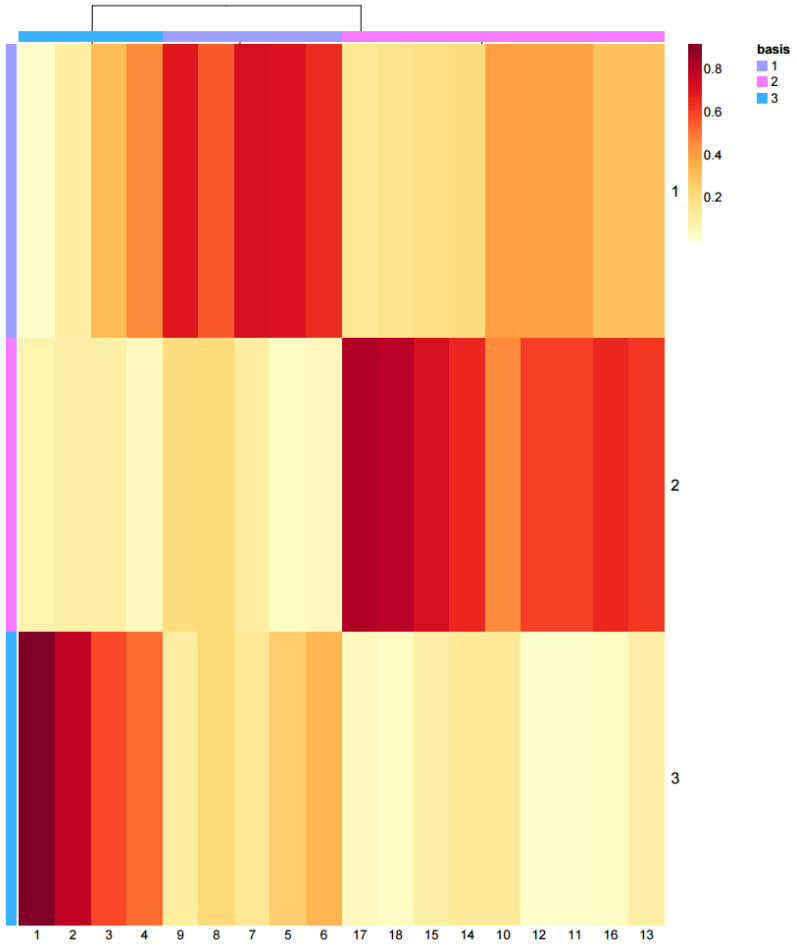
Sentiment Mining



Overview

Non-negative Matrix Factorization

- Data cluster of related seasons
- 18 Seasons, 3 Topics
- Clusters:
  - [1:4]
  - [5:9]
  - [10:18]



Non-negative Matrix Factorization

- Major changes in writing (2001/2002) and producing (2006)
- Correlates to different clusters from NMF
  - Season 5 in 2001
  - Season 10 in 2006
- Conclusion: change of writing style

Series Writing Credits

Trey Parker	...	(creator) (278 episodes, 1997-2016)
Matt Stone	...	(creator) (278 episodes, 1997-2016)
Brian Graden	...	(developer) (238 episodes, 1997-2016)
David R. Goodman	...	(88 episodes, 1997-2002)
Nancy Pimental	...	(62 episodes, 1998-2001)
Kyle McCulloch	...	(56 episodes, 1999-2002)

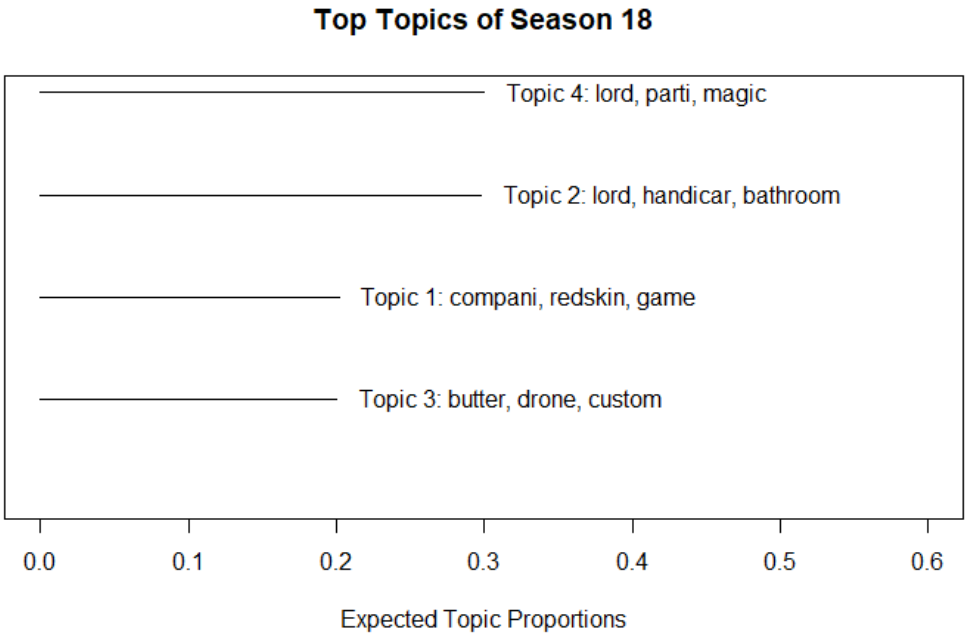
Series Produced by

Trey Parker	...	executive producer (278 episodes, 1997-2016)
Matt Stone	...	executive producer / producer (278 episodes, 1997-2016)
Anne Garefino	...	executive producer / producer / supervising producer (225 episodes, 1997-2016)
Frank C. Agnone II	...	supervising producer / executive producer / producer / post-production producer / line producer (212 episodes, 1997-2016)
Vernon Chatman	...	producer / consulting producer (109 episodes, 2006-2016)
Daryl Sancton	...	line producer (98 episodes, 2006-2016)
Pam Brady	...	creative producer / producer / consulting producer (95 episodes, 1997-2008)
Jennifer Howell	...	associate producer / supervising producer (93 episodes, 1997-2005)
Adrien Beard	...	producer (90 episodes, 2007-2016)
Bruce Howell	...	producer / executive producer (90 episodes, 2007-2016)
Eric Stough	...	producer (90 episodes, 2007-2016)
Erica Rivinoja	...	producer (61 episodes, 2007-2012)
Deborah Liebling	...	executive producer (58 episodes, 1997-2002)
Kurt Nickels	...	associate producer (53 episodes, 2007-2015)

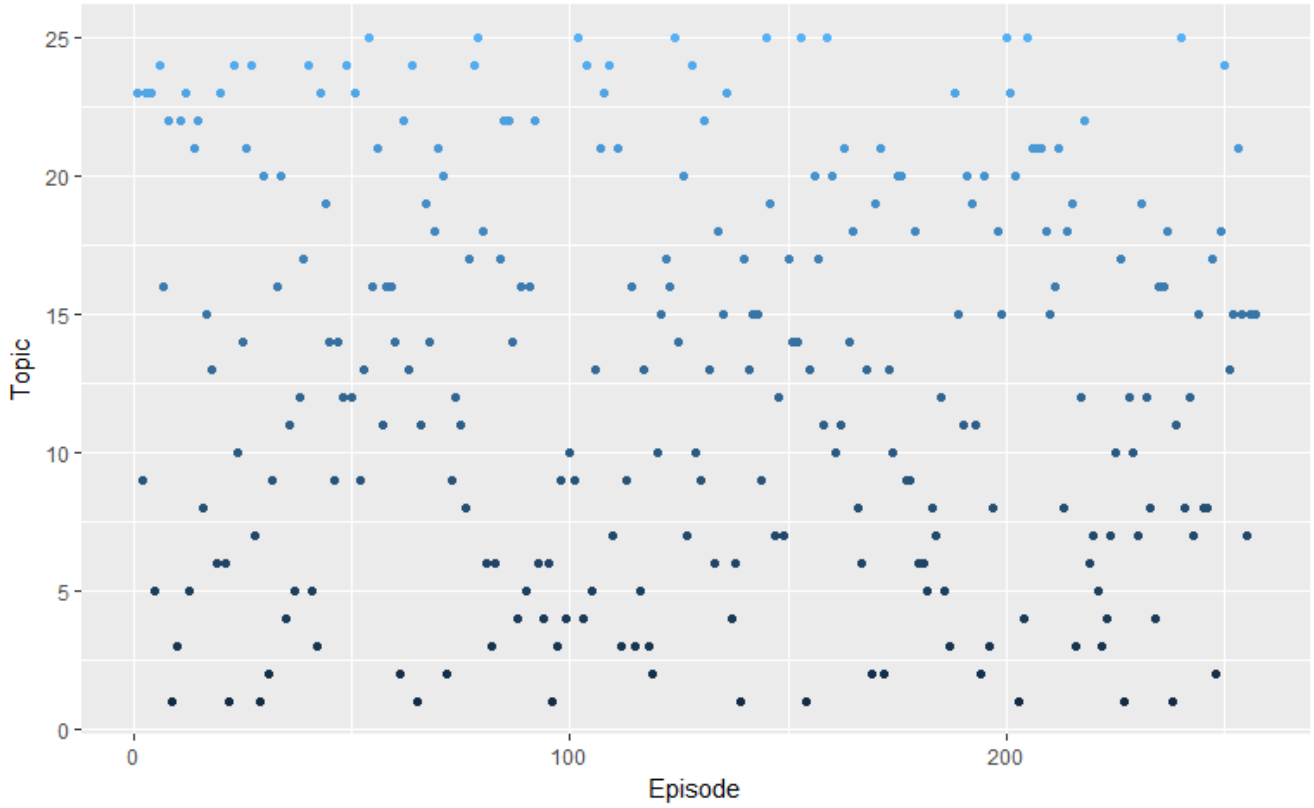
Overview

Structural Topic Model

- LDA with meta data



Structural Topic Model



Overview

Conclusion

- Extracted topics and corresponding episodes (what & when)
- Sentiment timeline (why & when)
- Future work
  - Change visualization of topic timeline
  - Incorporate persons

Thank you!

BACKUP