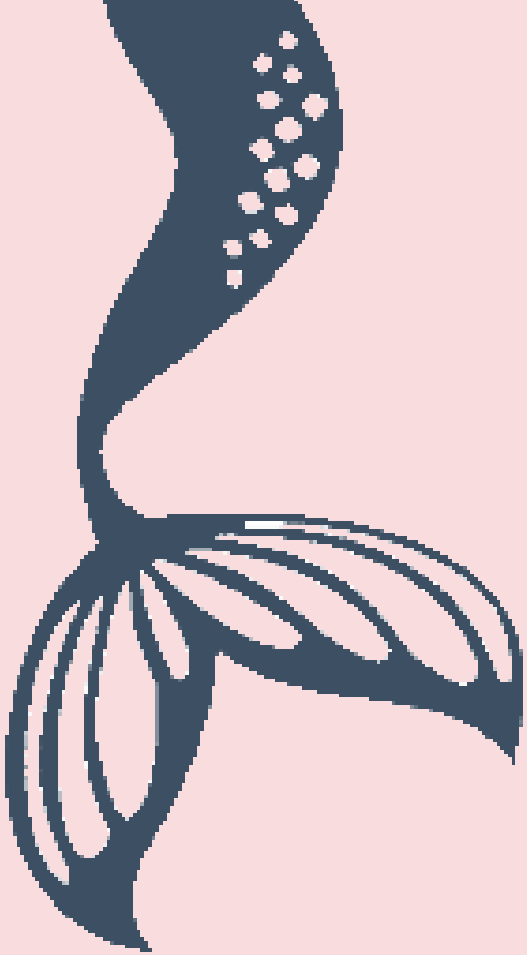
DSC680 FINAL PROJECT

spring 2020

Text Mining Visualizations of *The Little Mermaid*



Welcome

Business and Data Understanding

Analysis

Future Work and Conclusion

BUSINESS AND DATA UNDERSTANDING

WHAT IS...

TEXT MINING?

an overview of this technique and its applications

LITERARY ANALYSIS?

insight into the benefits of literary analysis

THE LITTLE MERMAID?

summary of the original story by Hans Christian Andersen

Text Mining

AND HOW IT IS DIFFERENT FROM TEXT ANALYTICS

Text Mining

- Portable Document Format
- unstructured data => structured data
- cleaning
- initial insights

Text Analytics

- machine learning
- predictive modeling applications

Literary Analysis

IN THE TRADITIONAL SENSE

- essays
- personal interpretation
- tone
- writing style
- syntax

The Little Mermaid

ORIGINAL SHORT STORY

- Hans Christian Andersen
- 1836
- Story
  - Little Mermaid falls in love with a prince
  - Little Mermaid makes sacrifice in attempt to obtain an immortal soul
  - Prince falls in love with someone else
  - Little Mermaid kills herself to spare the prince's life, obtains immortal soul

ANALYSIS

DATA EXTRACTION
the data source

METHODS
tools and packages used

RESULTS AND INSIGHTS
visualizations to help understand the story

Data Extraction and Methods

HOW AND WHAT

Data Extraction

- R - RStudio
- PDF

Methods

- Packages
  - *tm*
  - *stringr*
  - *qdapRegex*
  - *dplyr*
  - *tidytext*
  - *tidyverse*
- Data cleaning and manipulation
- Tableau

Results and Insights

WHAT INFORMATION DOES THE TEXT HIDE BETWEEN THE LINES?

- Counts
- Sentiment Analysis
- Gender Analysis

Project Insights

Additional Applications

Counts

20%

OF THE WORDS IN THE NOVEL WERE QUOTES

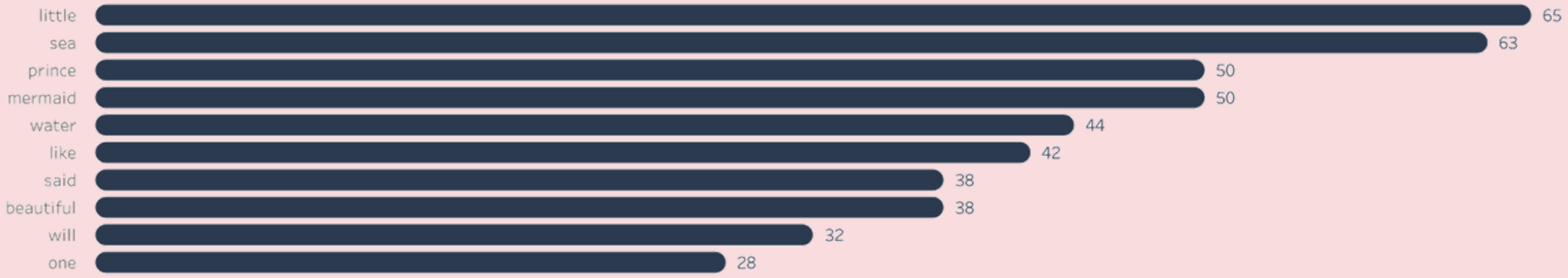
1849 spoken words

9251 total words

- action
- descriptive

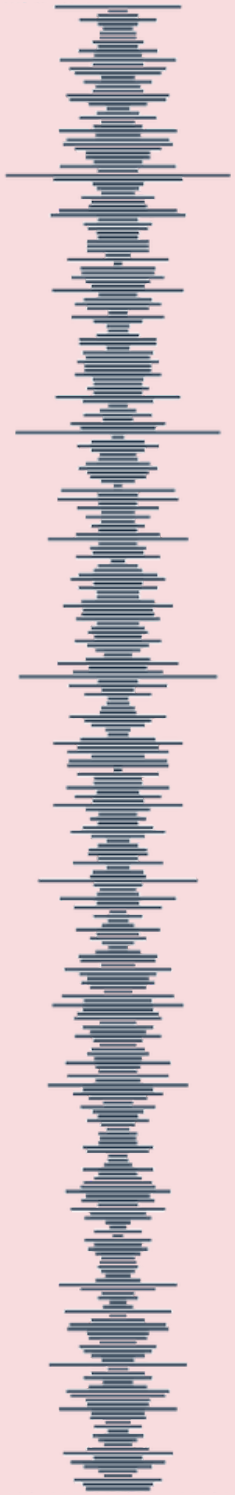
- screenplay adapatbility

TOP TEN WORDS



- motif
- theme
- characters

- document identifier
- market research

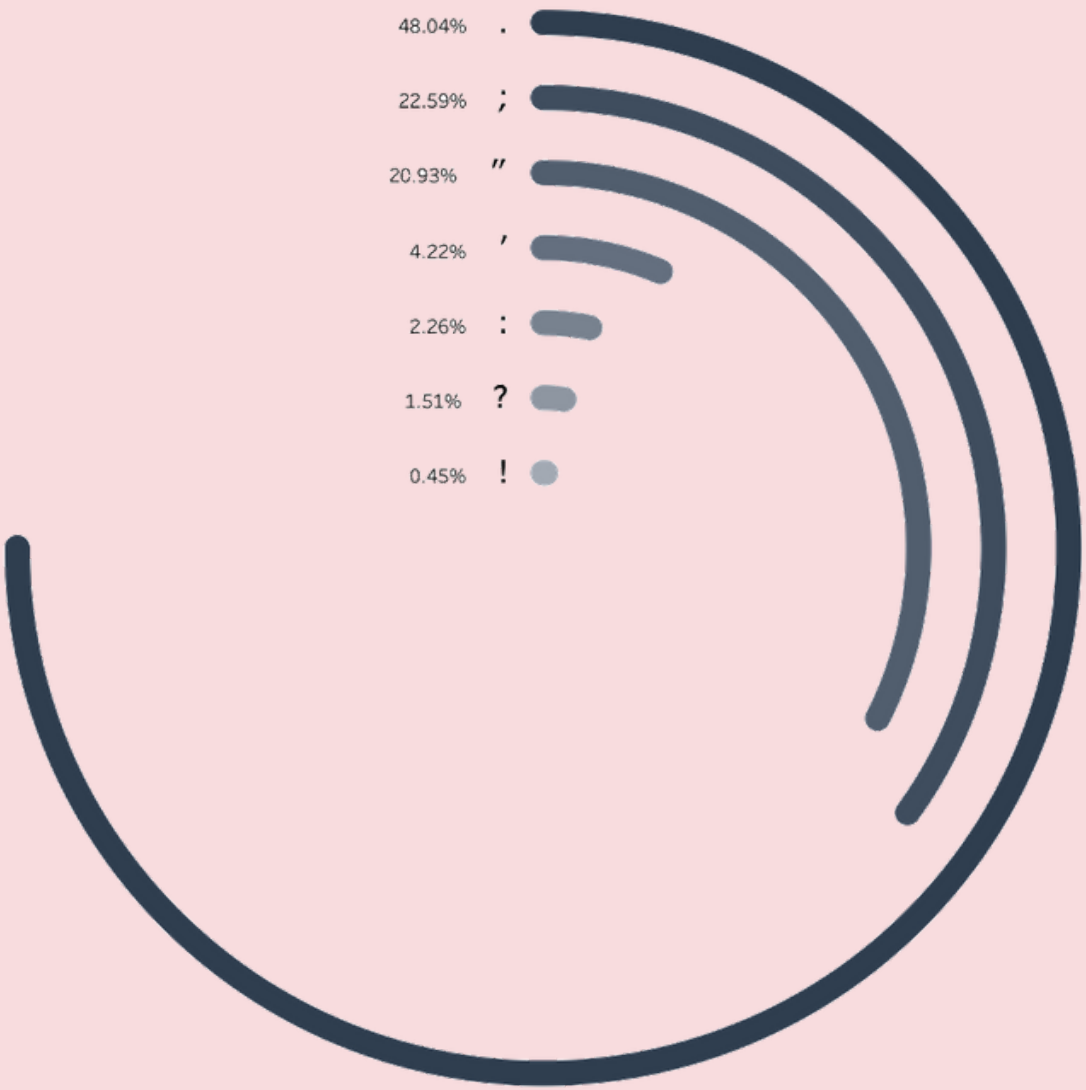


SENTENCE LENGTHS

- pattern identifier
- balanced

- document cadence
- essay proof-reading

PUNCTUATION COUNT



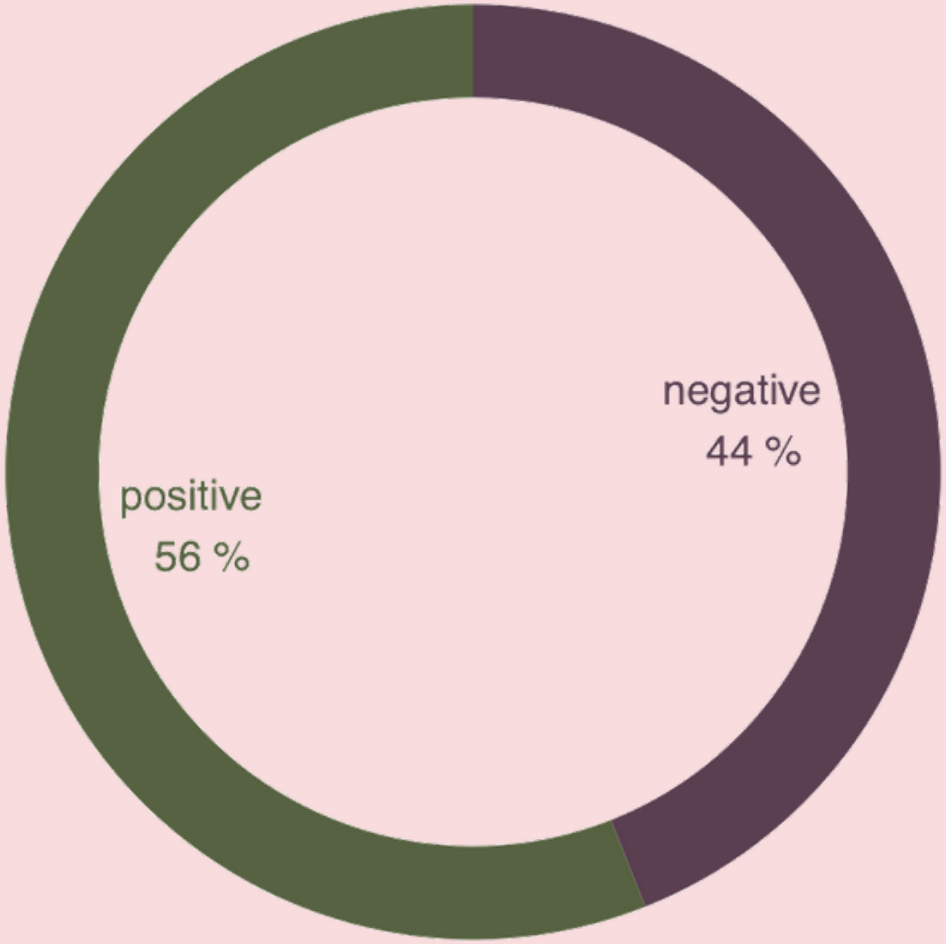
- period most used
- sentence choice

- tone
- document categorization

Sentiment Analysis

SENTIMENT SKEW

Sentiment Breakdown



- generally positive
- tone

- market research
- product research

ASSIGNED POLARITY

Sentiment Contribution

by word counts greater than five in *The Little Mermaid*

- sentiment polarity
- light imagery
- mostly positive

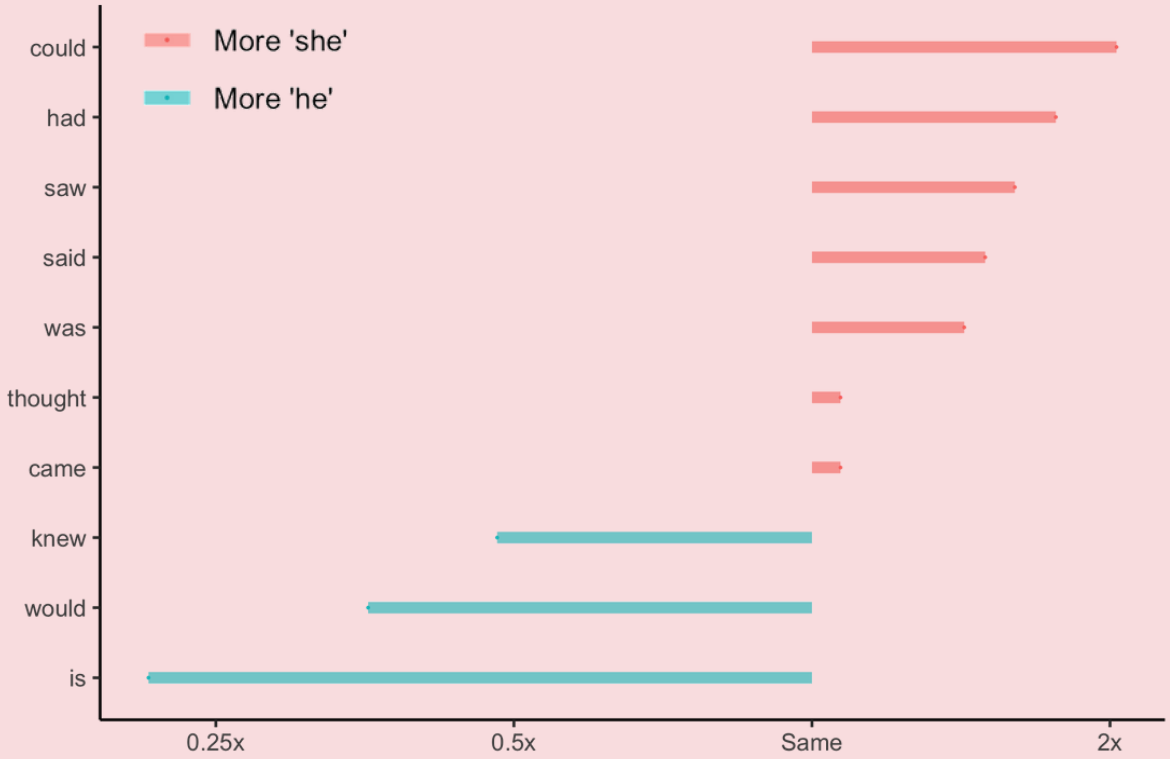
- individual word significance
- writer's block

Gender Analysis

TRAILING WORDS

Bigram Frequency

trailing words paired with 'he' and 'she' in *The Little Mermaid*



- 'she' more possessive
- women spoke more
- predominantly female narrative

- gender bias through generations
- trailing and leading word sentiment

FUTURE WORK AND CONCLUSION

FUTURE EXPLORATION

- comparison between Andersen's other works
- literary gender analysis

RESULTS AND INSIGHTS

- pattern detection
- unexpected insights
- vast sentiment analysis applications

Thank You

Resources and references found in written report