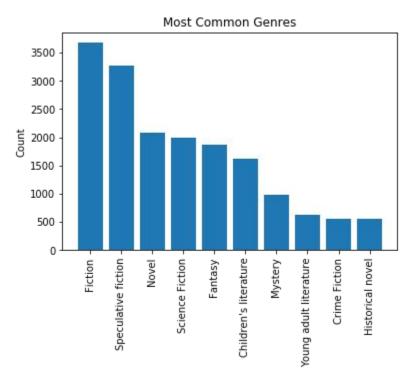
Genre Discovery: Using Naive Bayes Modeling to Predict
Genre

Samantha Gunnerson

#### Introduction

- Genre of form vs. Genre of content
- Unique problem:
  - Poorly Defined
  - Subgenres via combining genres
- Originally from the work of David Bamman and Noah Smith at Carnegie Melon
- 16,559 Wikipedia book summaries with extra information from Freebase

## Initial EDA



Most Common by Genre 3500 3000 2500 Count 2000 1500 1000 500 Children's literature Horror Speculative fiction Fantasy Historical fiction Young adult literature Thriller Suspense

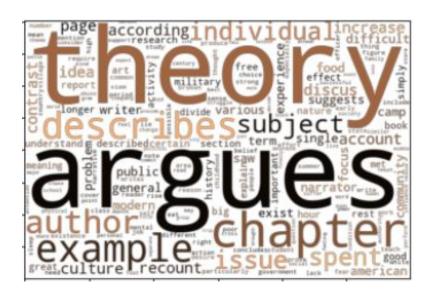
\*Including genres of form

\*Within only fiction genres

#### Models and Predictive Words

- Modelled using Naive Bayes
- Model genres of form (fiction vs nonfiction)
- Model genres of content (various fiction genres)

Fiction vs. Nonfiction

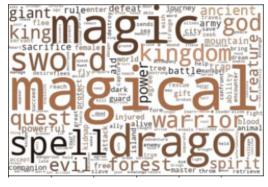


#### Predictive Words P2



Speculative Fiction





Fantasy



Children's Literature



Science Fiction Young Adult Literature

#### Predictive Words P2



Mystery



agent Case shortly per those accident was commit hire criminal have crimes accident was commit and crimes accident was committed accident

Suspense



Horror



Historical Fiction Thriller

# **Application Online**

#### Twitter:

Cleans tweets based on a username and checks fiction genre

#### Summary:

Runs hierarchical model described in rest of presentation

### Conclusion

- Consistency is surprising in the genres
  - Suggests a common language of genre in Western culture
- Further Analysis
  - Actual text from literature
  - Other languages/cultures