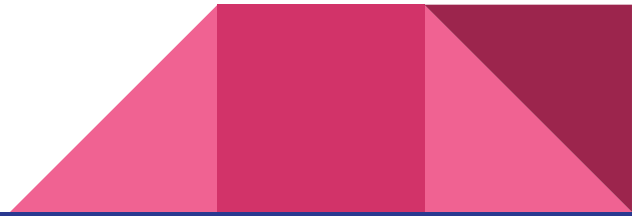
# Graduate Group 3: n-Gram Distributions

Annika Sparrell, Kasey La, Ge Gao

COSI 116A — Prof. Dylan Cashman, Brandeis University

# Motivation and Background

- Explore a topic of interest to the group and our degrees
- Create a visualization for a domain that lends itself to interesting statistics
- N-grams are sequences of text that appear sequentially– unigrams are each single word, bigrams are each pair of words, etc.
- We decided to explore unigram distributions over different texts as we felt this would lend itself to interesting visualizations.



# Data Analysis

- We used three freely available texts as the data from which we drew n-grams, *Alice in Wonderland*, *Pride and Prejudice*, and *The Great Gatsby*.
- We preprocessed the data in python to make it usable by D3
- Ultimately we obtained a CSV of unigrams across each text along with their counts and probabilities

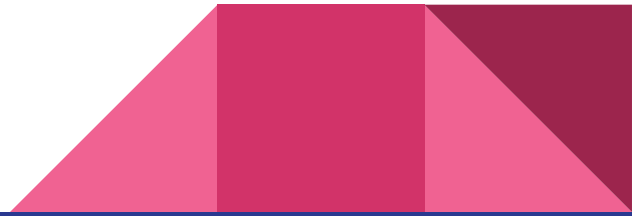
# Task Analysis

- Our most important tasks were chosen to be “Understand n-grams through visualization” and “Determine how different texts produce different distributions”.

- Our most important mid-level task was “explore” and our most important high-level tasks were “discover” and “enjoy”

# Design Process

- We stuck fairly close to one of our ideas from the paper sketches as it lent itself well to the medium
- Although we wanted to explore additional aspects of n-grams, especially text generation, these avenues were too ambitious to include in the scope of this project
- We did succeed in implementing most of our ideas from our interactive sketch in our final visualization



# Visualization Explanation

- We used areas and text as the marks in our visualization, as our data originates in text format and area is an effective way of encoding proportion

- Color was our main channel, identifying data as belonging to a given text. Size was also used within the bar charts.

- Our main interactions were hovering over a bar to gain insight into the numerical details, as well as brushing over words in the text to see the distributions of those words in the charts.

# Conclusions

- Ultimately our visualization is a fun view on the distributions of words across three specific texts, but it does not get at any deep aspects of NLP.
- Given more time we would certainly implement the rest of the features on the nice-to-have checklist.
- We would also like to explore different visualizations that can showcase other aspects of this domain if we had more time.

# Visualization Demo