Stats 112 Final Project

Group 2: Cherry Li, Ashley Munayco, Caitlin Ree, Shane Remo, Meiyi Ye

I. Abstract

In our study, we aim to analyze reflection papers from each group on the guest speaker Ms. Susan Philips. We are interested in identifying the underlying common themes among each group's different understanding of the chapters about individuals immigrating to the U.S. and Ms. Phillip's talk. To do this, we utilized text mining techniques such as frequency graphs, and word clouds. For our data model, we used text networks, a cluster dendrogram, and LDA topics modeling to find common and significant words in the reflection papers.

II. Statement of the Problem

The research question we are addressing is "What common themes unify the immigrant experience as seen through group reflections?". Using text mining of the reflection papers, we want to find if similar themes about the immigrant experience resonated among different groups.

III. Variables

The data we analyzed came in the form of reflection papers from Groups 1-8 along with a solo reflection paper. Each of the group reflection papers includes each member's response to the question "What chapter did you read, and what did you learn from this?" and the entire group's response to the question "Did reading the chapter and listening to Ms. Philips change your attitude or future behavior in any way?". The solo reflection paper is based on the individual's interview with a close friend named Rachel regarding her immigration experience. Our variables are thus in the form of character vectors of the text mined from each of the reflection papers.

IV. Exploratory Data Analysis

The reflection papers were given in raw PDF format, so we first converted the files into .txt format. As a standard method of cleaning text data, we read in the nine .txt files into R as character vectors and then proceeded to convert them into "corpus" objects. We then cleaned the corpora, getting rid of extraneous words and stemming words in order to combine different tenses of words (for example, "story" and "stories" would both become the stem "stori"). From this, we made frequency graphs and word clouds for each of the nine documents, using unstemmed versions of the previously stemmed words for clarity (now, "stori" would become "stories" again).
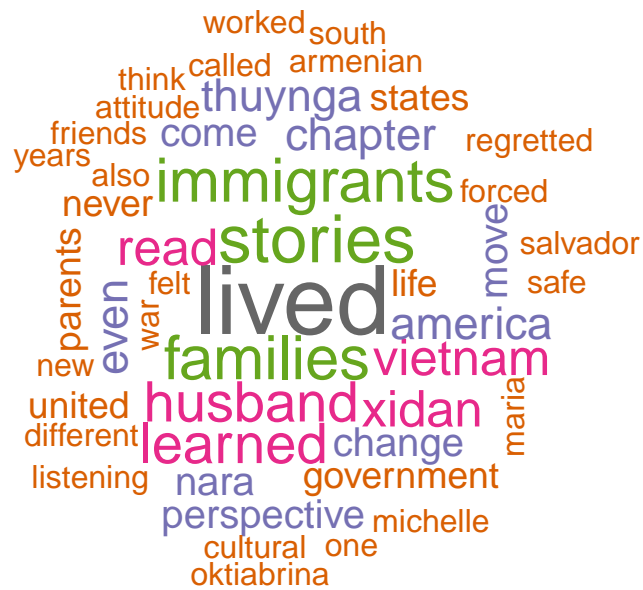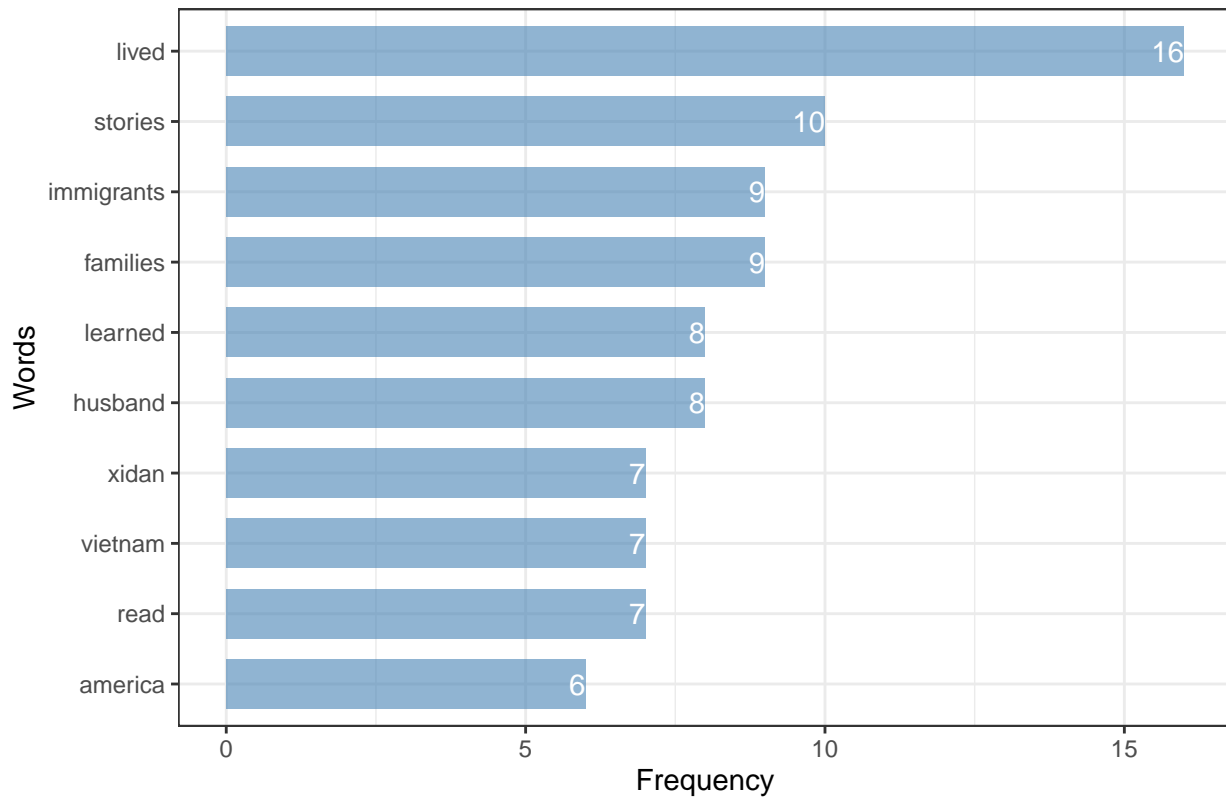
Group 1
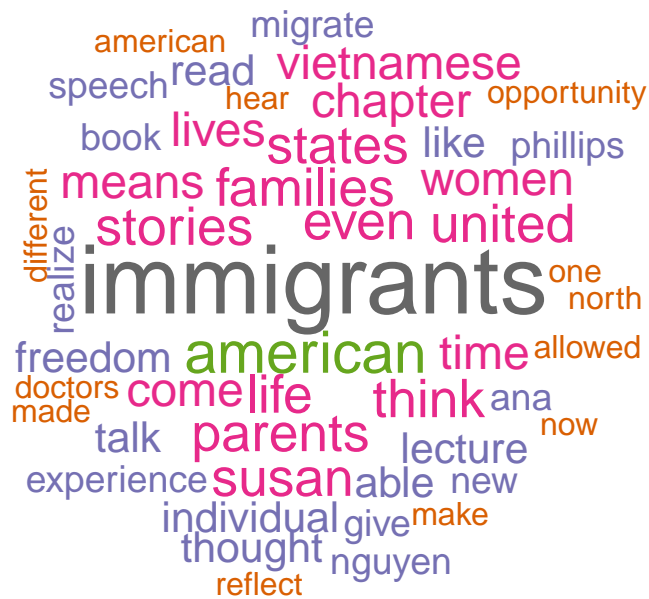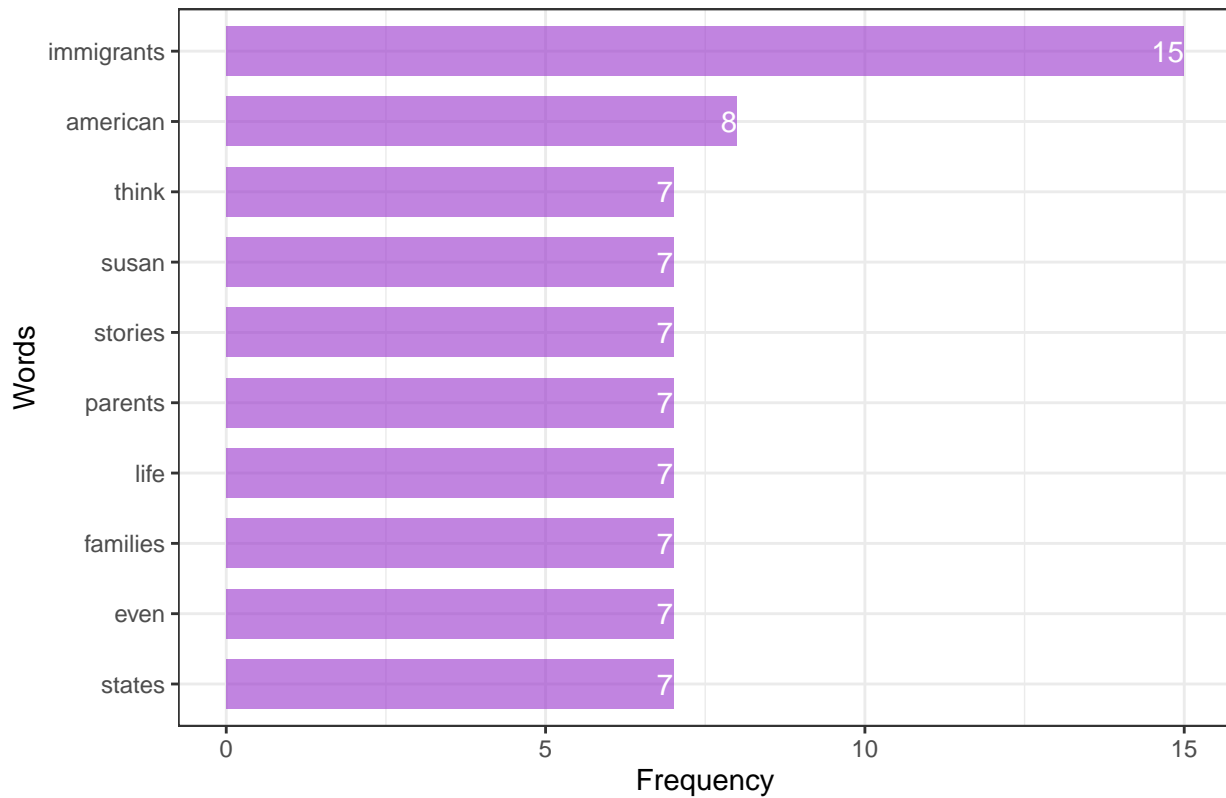
Top 10 Words in Group 1's Reflection
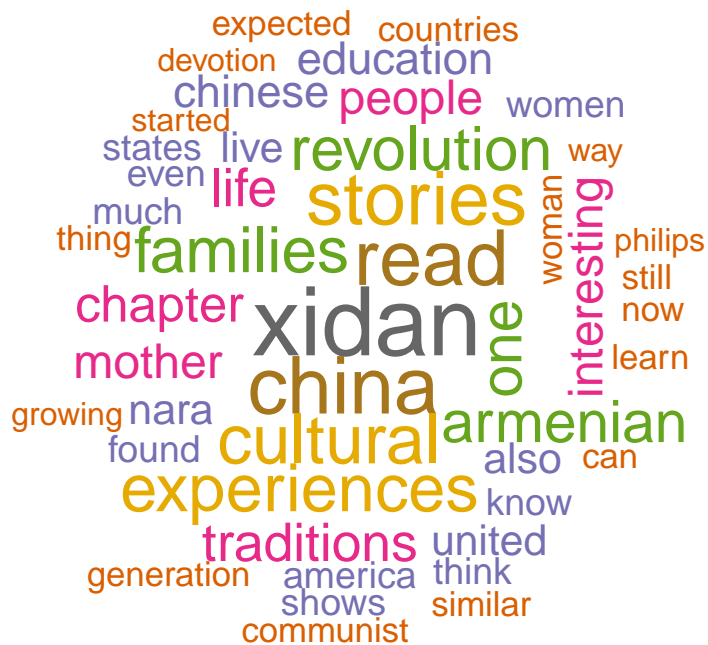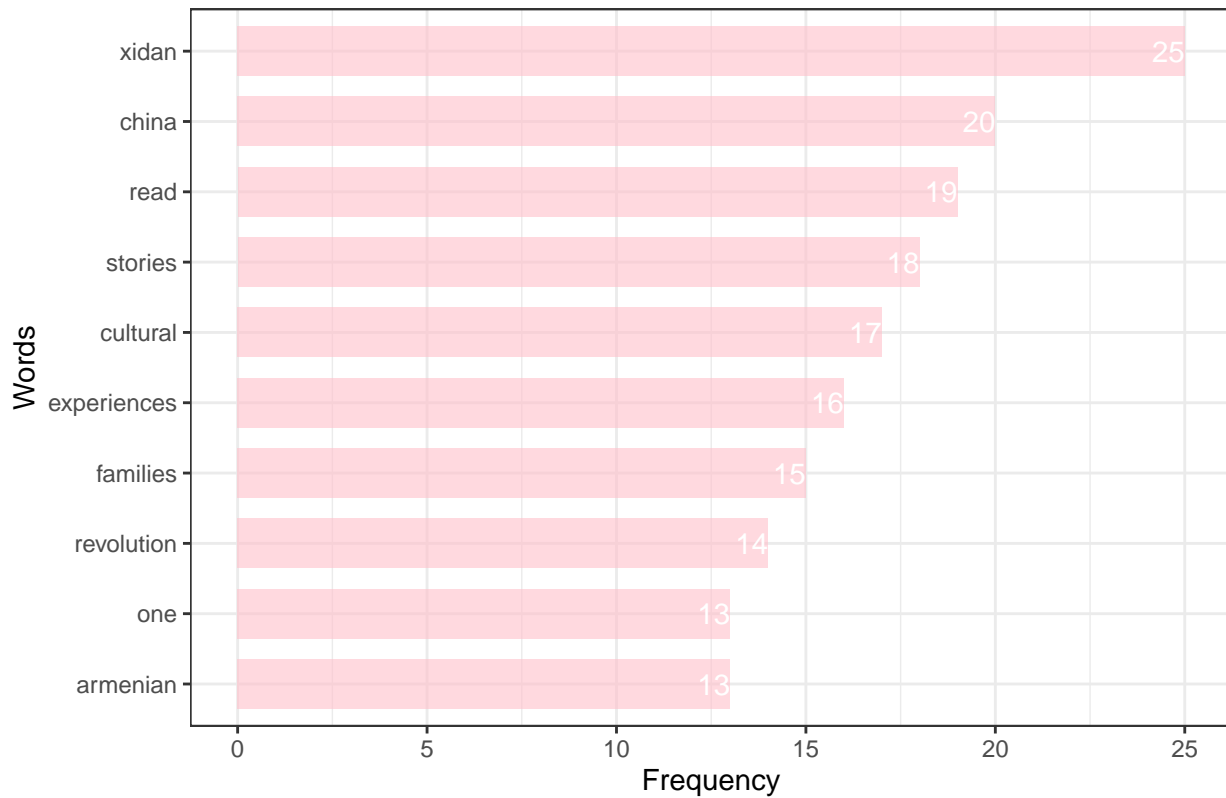
Group 2

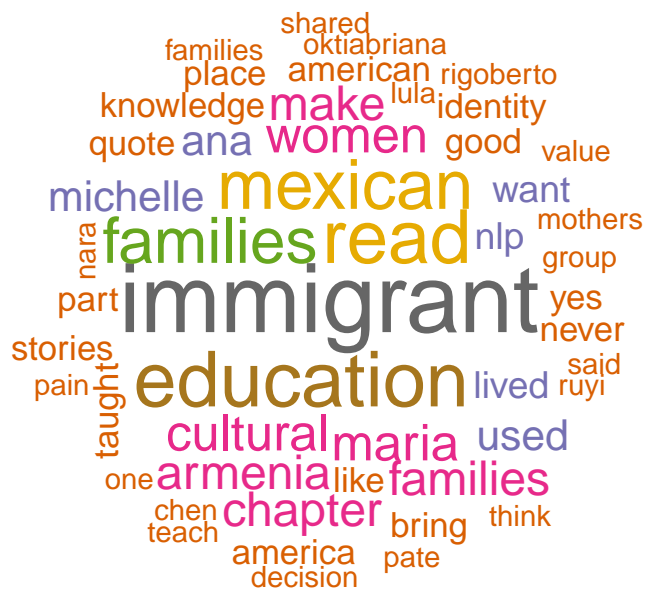## Top 10 Words in Group 2's Reflection

Group 3

Top 10 Words in Group 3's Reflection

Group 4

Top 10 Words in Group 4's Reflection

Group 5

Top 10 Words in Group 5's Reflection

Group 6

Top 10 Words in Group 6's Reflection

Group 7

Top 10 Words in Group 7's Reflection
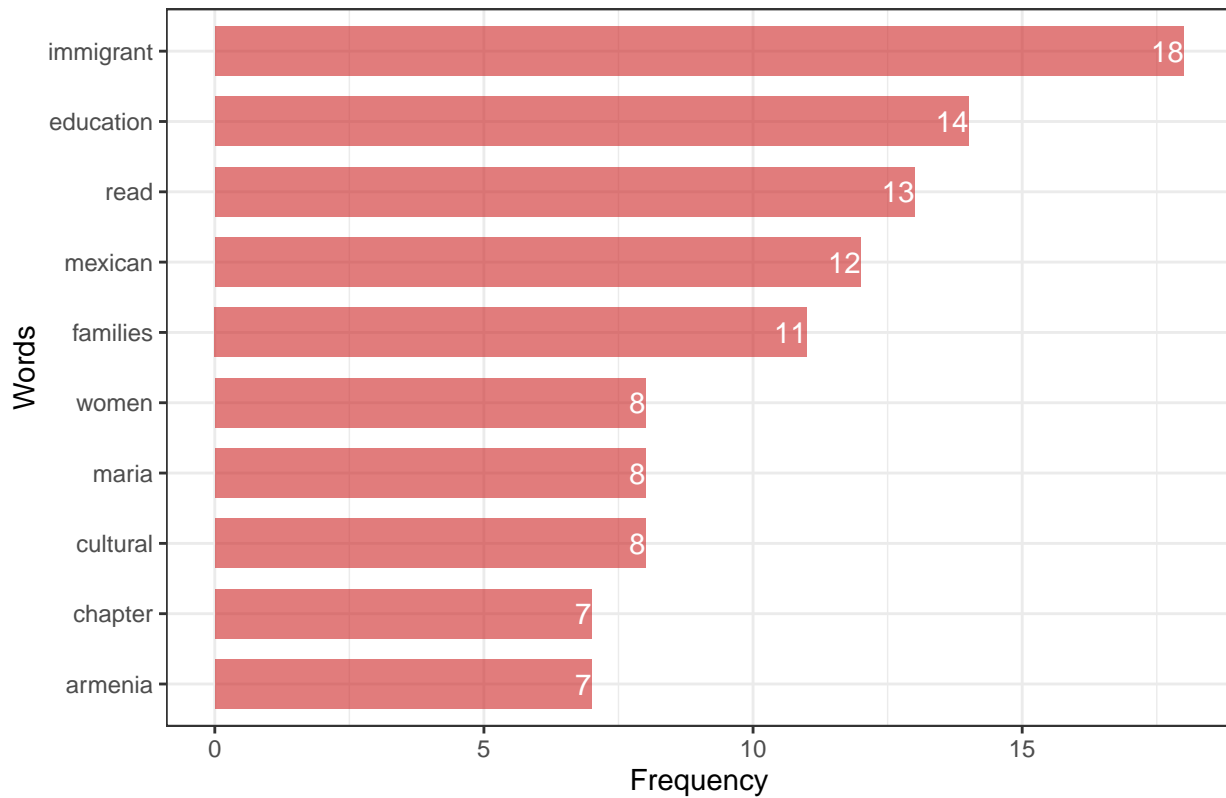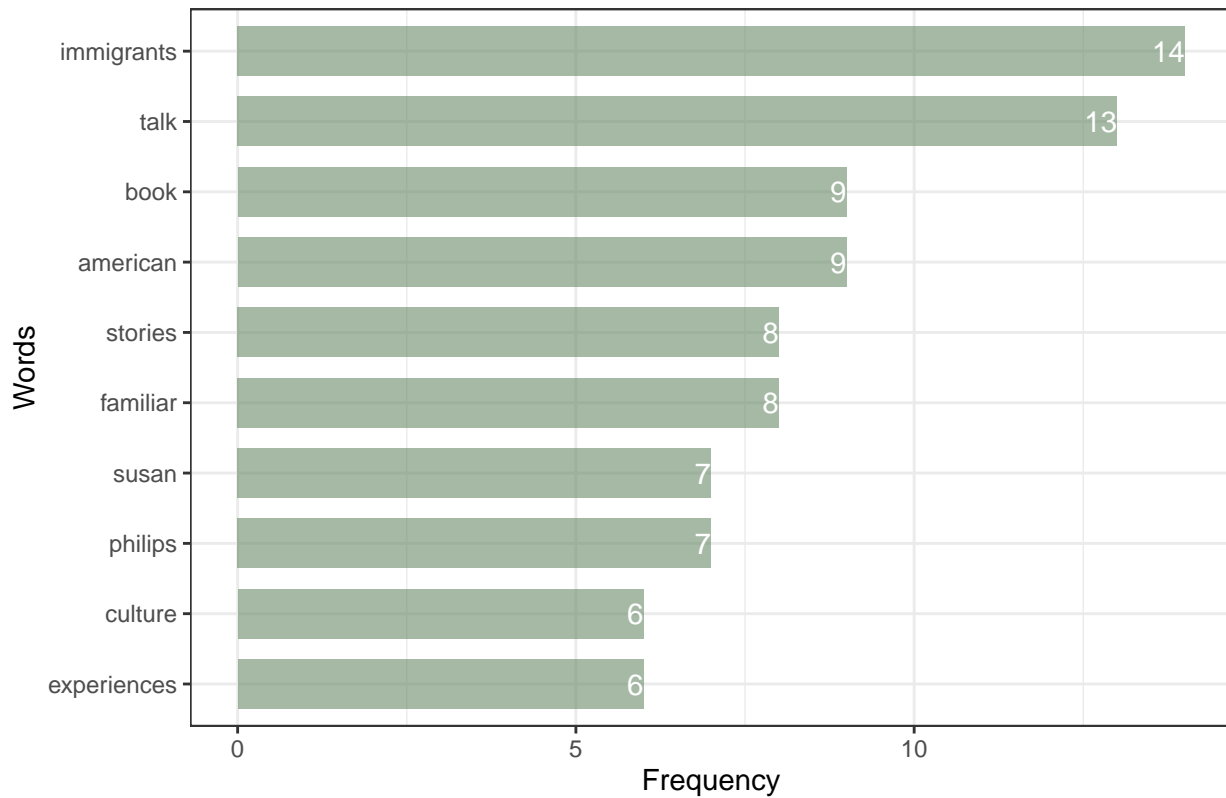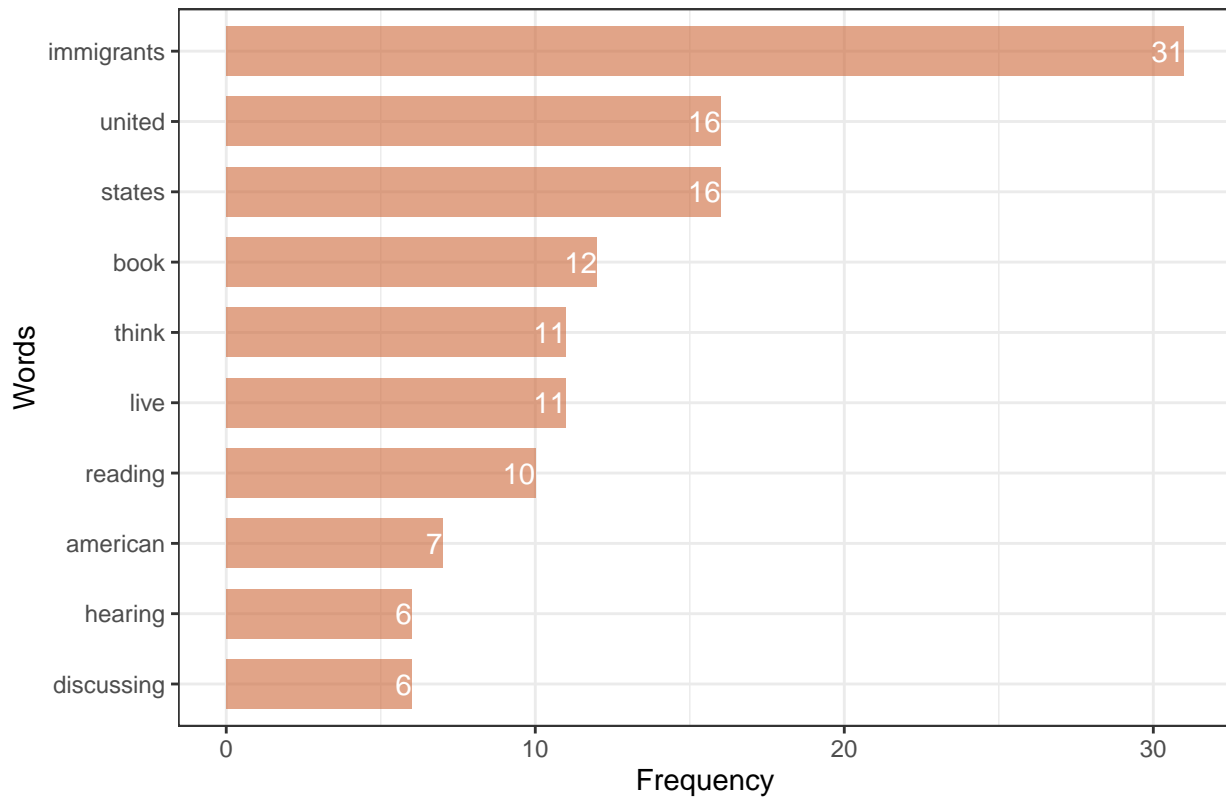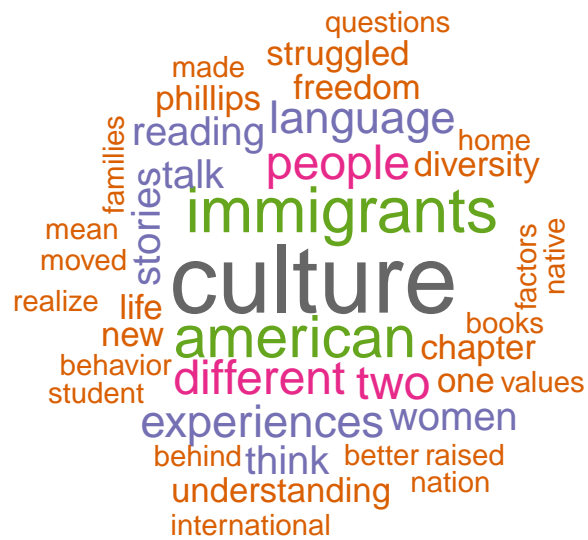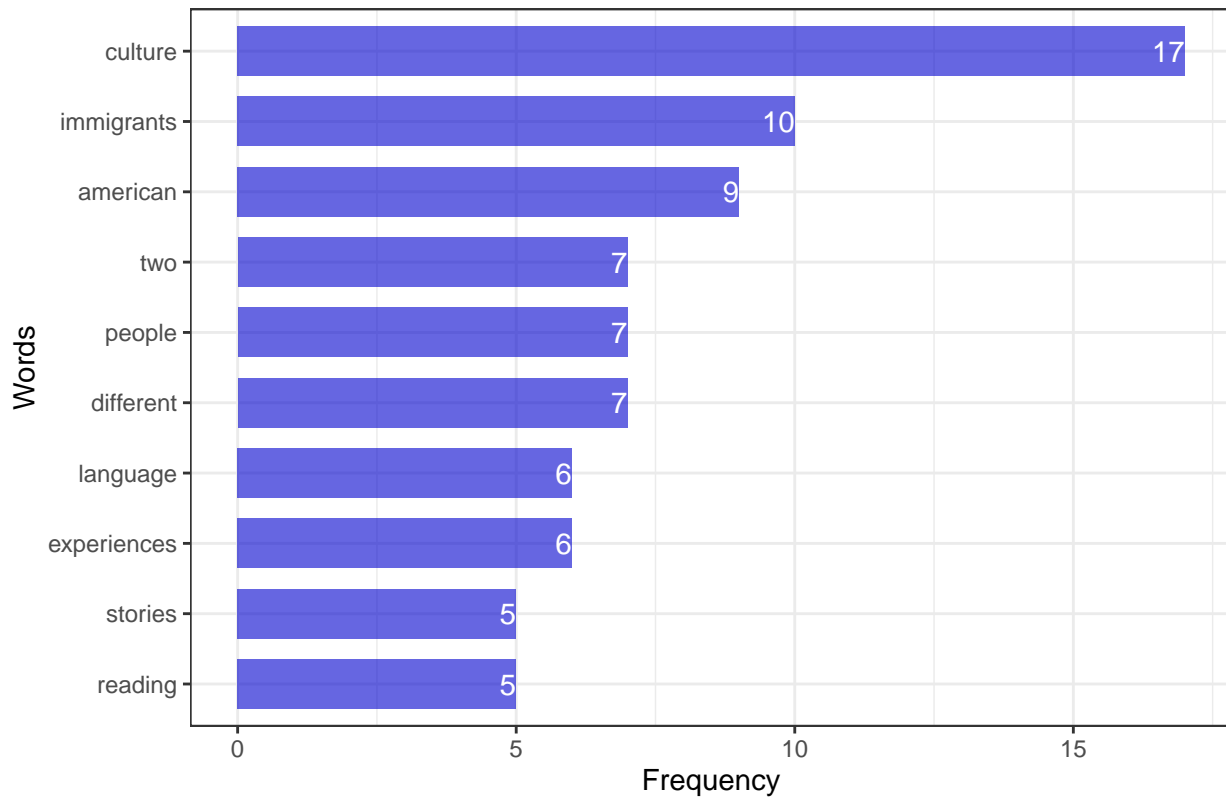
Group 8

Top 10 Words in Group 8's Reflection
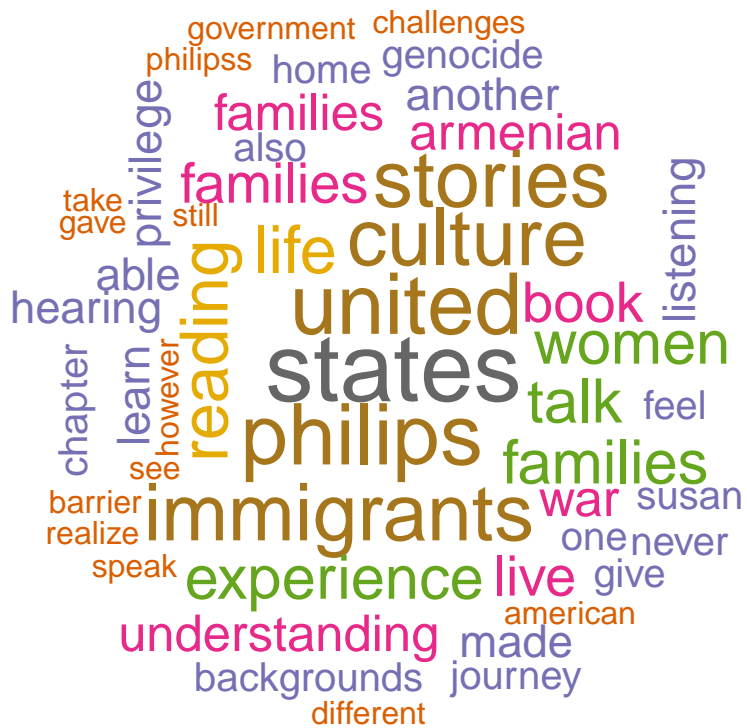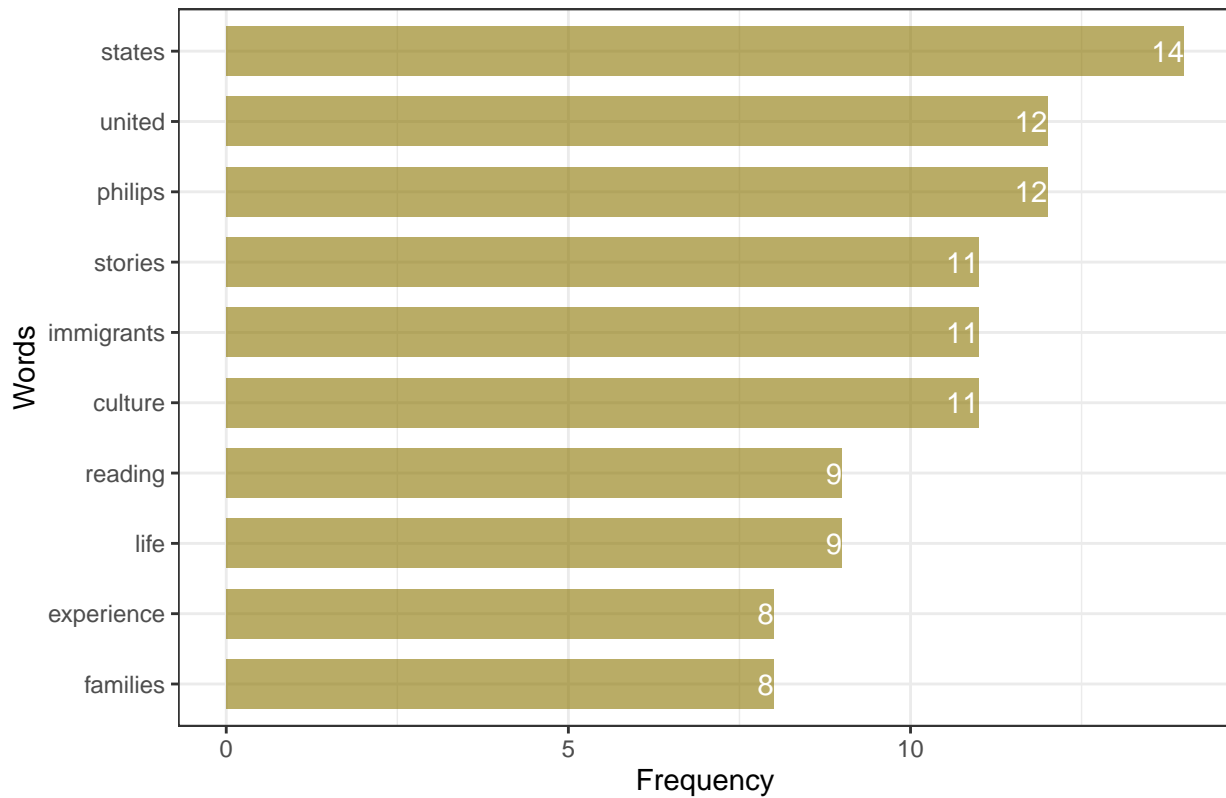
Solo

Top 10 Words in Solo's Reflection
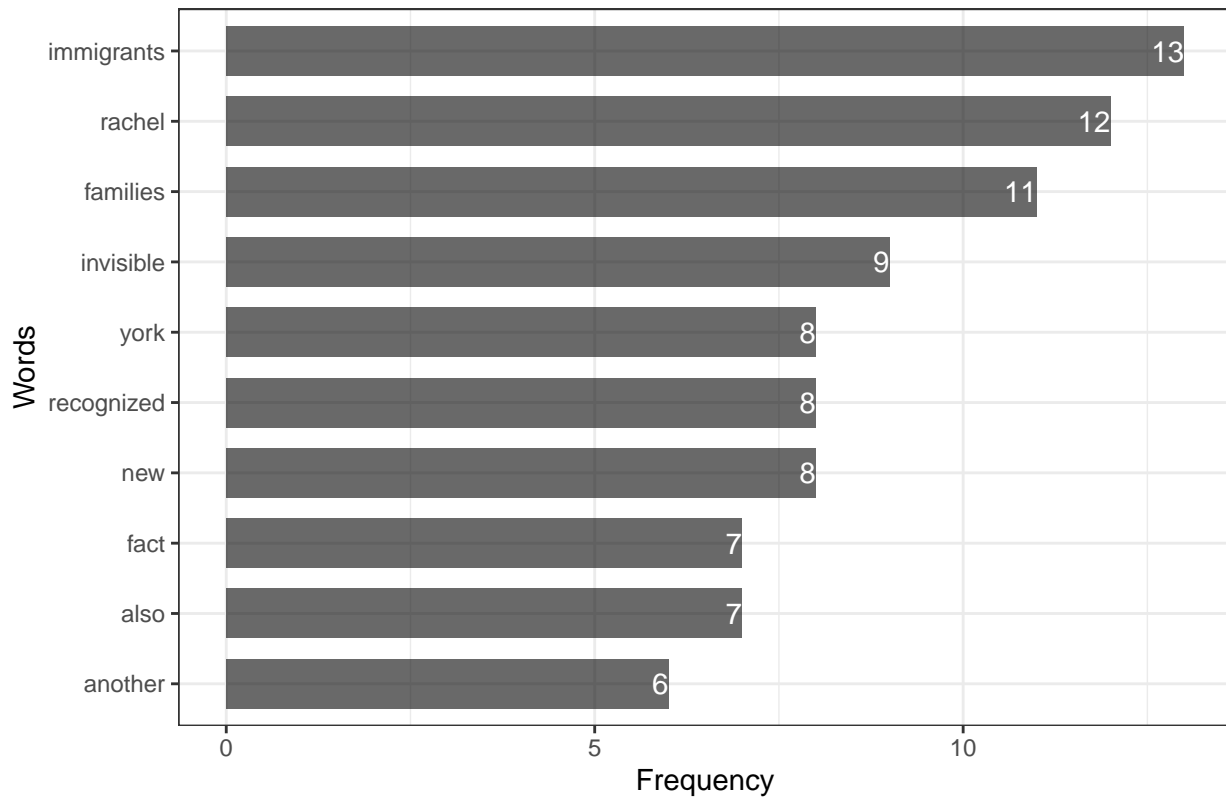
As can be seen through these frequency charts and word clouds, there are some common words that frequently appear across all of the reflection papers. For instance, "immigrants", "families", and "culture" are among the top ten most frequent words. For a closer look at each group's reflection paper, some common words unique to Group 1, for instance, are "Xidan", "China", and "revolution." Group 4's reflection commonly

used the words "education" and "women", and the solo reflection commonly used the words "invisible" and "recognized." All of these words speak to the various significant factors of the immigrant experience, such as the separation or unification of families, the mixture between host culture and guest culture, educational opportunities or lack of opportunities, and much more.

The following frequency chart and commonality word cloud highlights the words that appeared most frequently for all ten reflection papers combined. As mentioned previously, these words include "immigrants", "families", "reading", "stories", "culture", "live", "United States", "life", and "experiences." Some other interesting words that are used less frequently and shown in the word cloud are "parents" and "mother", which contributes to the common theme of "family" across all reflection papers. Lastly, the words "understanding" and "different" are also important to point out as they speak to the notion of how despite these commonalities, each immigrant experience is still different in its own way.

Commonality

Top 10 Words for Reflection Papers Combined
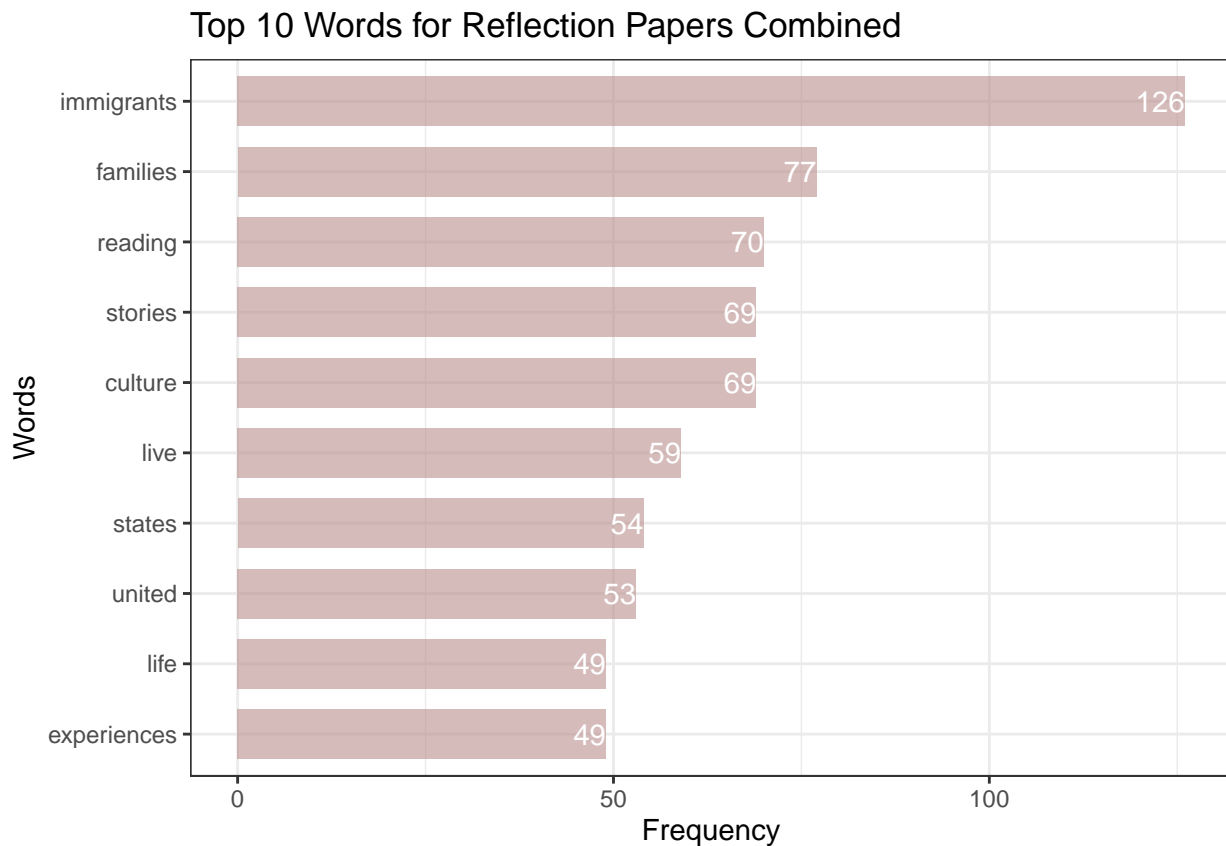
V. Statistical Methods and Summary of Results

Text Networks

Text Networks use linguistic analysis to show the relationships between words, with the location of nodes revealing the physical closeness of the words in the text and thickness of lines revealing how often the words appear together in the text. It's worth noting that the words are represented by their stems, so some context is necessary to determine what the unstemmed word should be.

Group 1
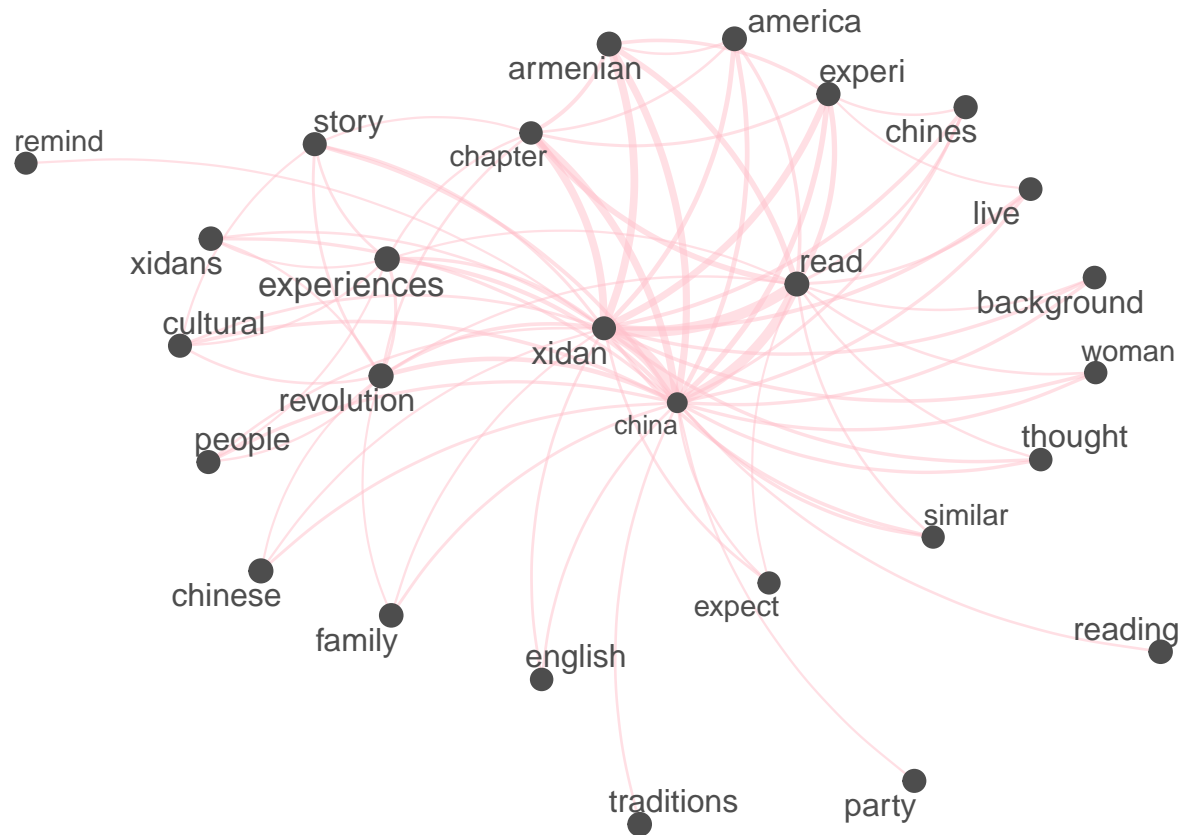
Xidan and China are, expectedly, quite close in the network, along with the word "read", since that was the story that Group 1 focused on the most. Other words that appear in the network that reveal themes are "experiences", "similar", and "family".

Group 2
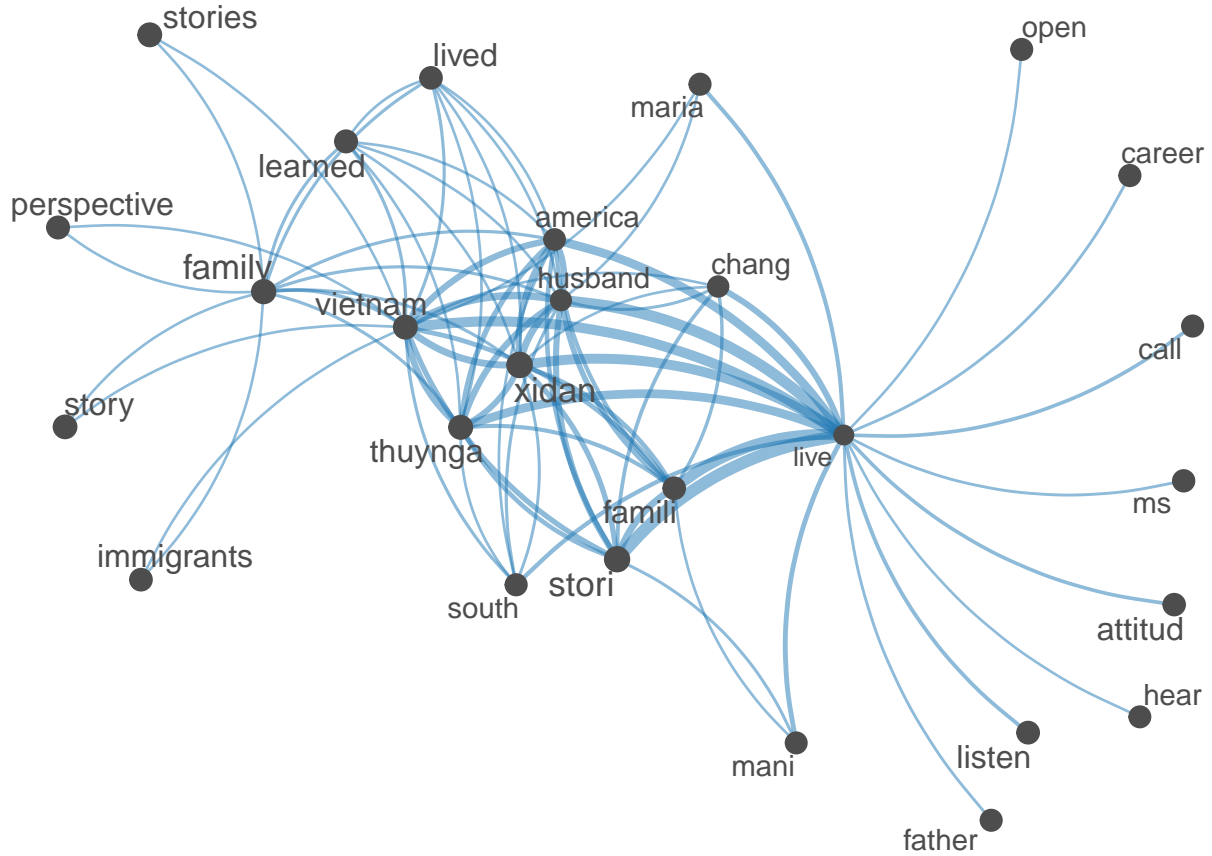
There are very strong lines in the center of Group 2's network, including among "family", "story", "lived", and "husband", as well as the locations and names of people in the book.

Group 3
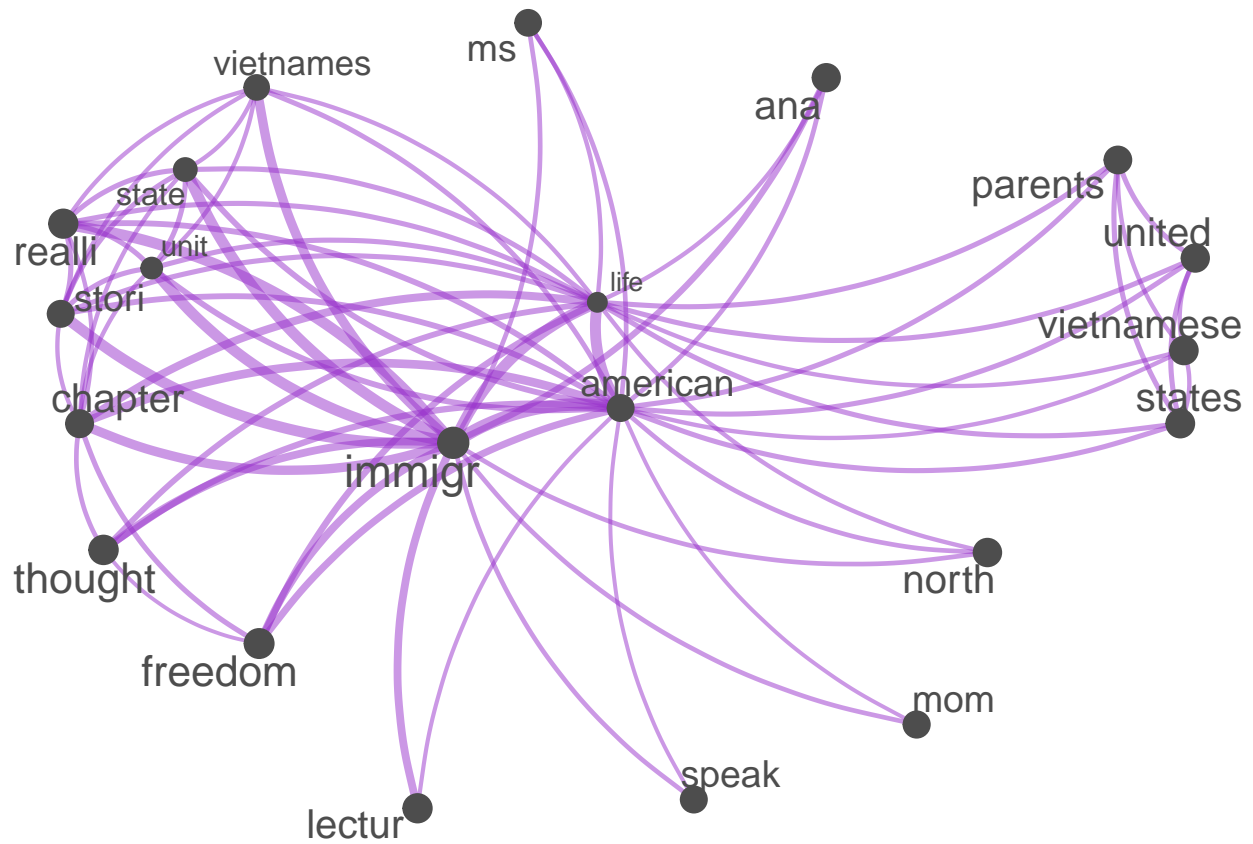
"Immigrants", "American", "freedom", and "life" appear predominantly in Group 3's network, signaling the emphasis they placed on the experiences of people after their immigration to America.

Group 4
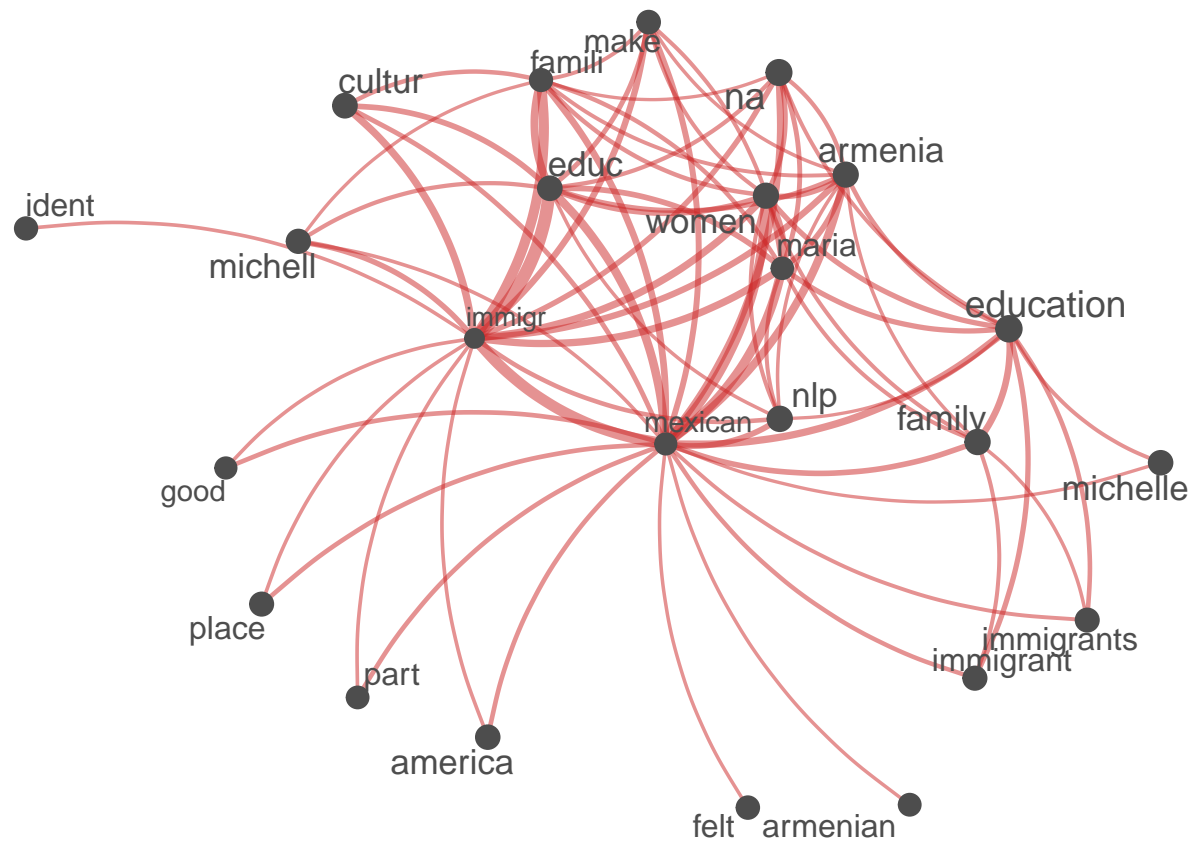
Besides country and person names, words like "education", "culture", and "identity" show some of the losses immigrants experience when they move to a new country.

Group 5
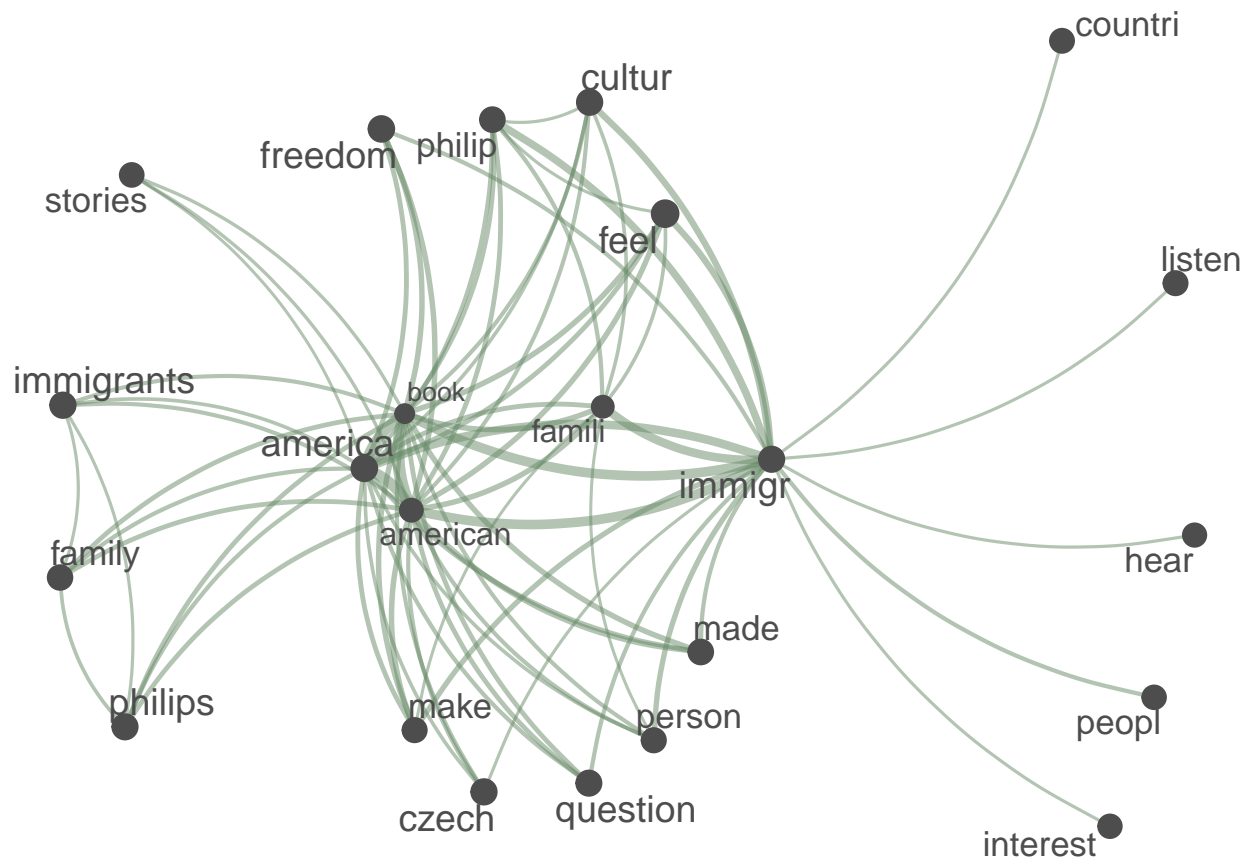
"Freedom", "culture", "America", and "immigrants" appear strongly in this network as well, similarly to other groups.

Group 6
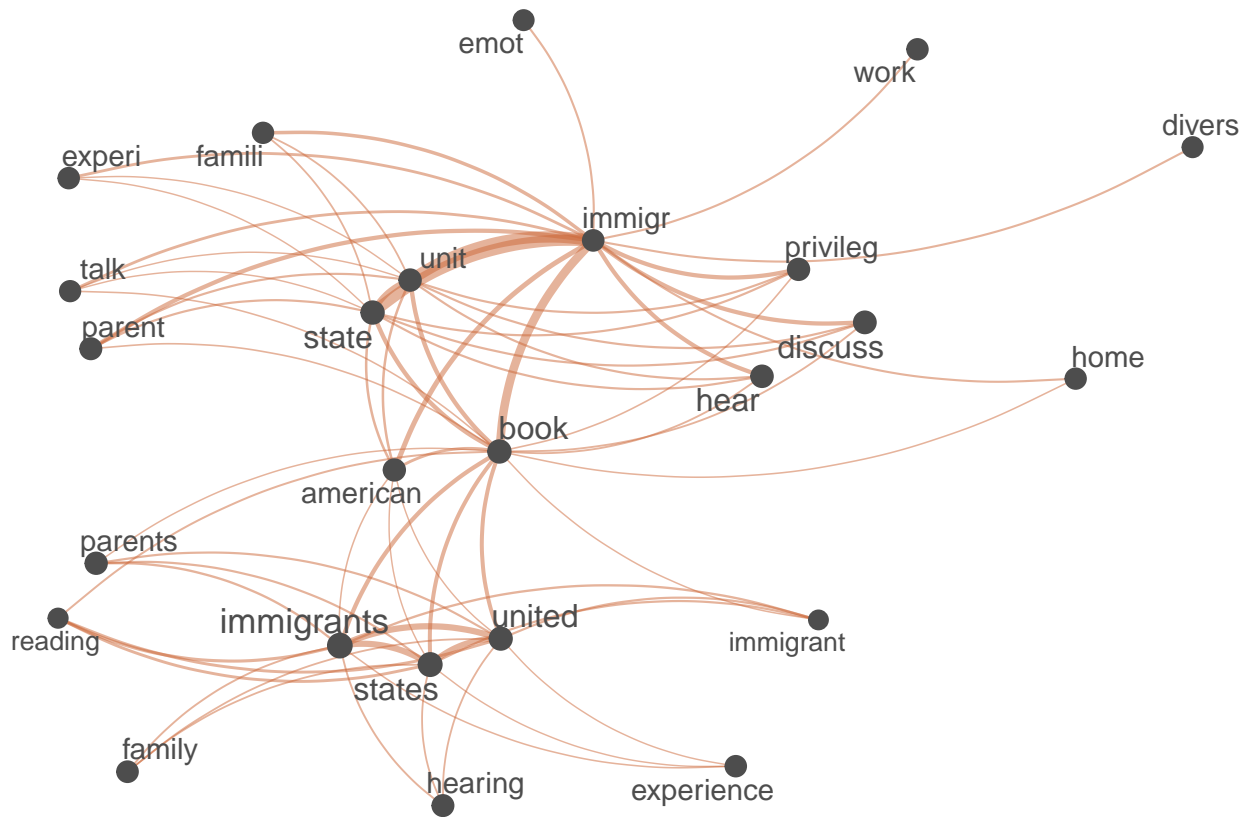
Group 6's network looks slightly different, with the only strong connection being between "United", "States", and "immigrate". Words that have not appeared in other networks appear farther out, such as "diverse" and "emotion".

Group 7
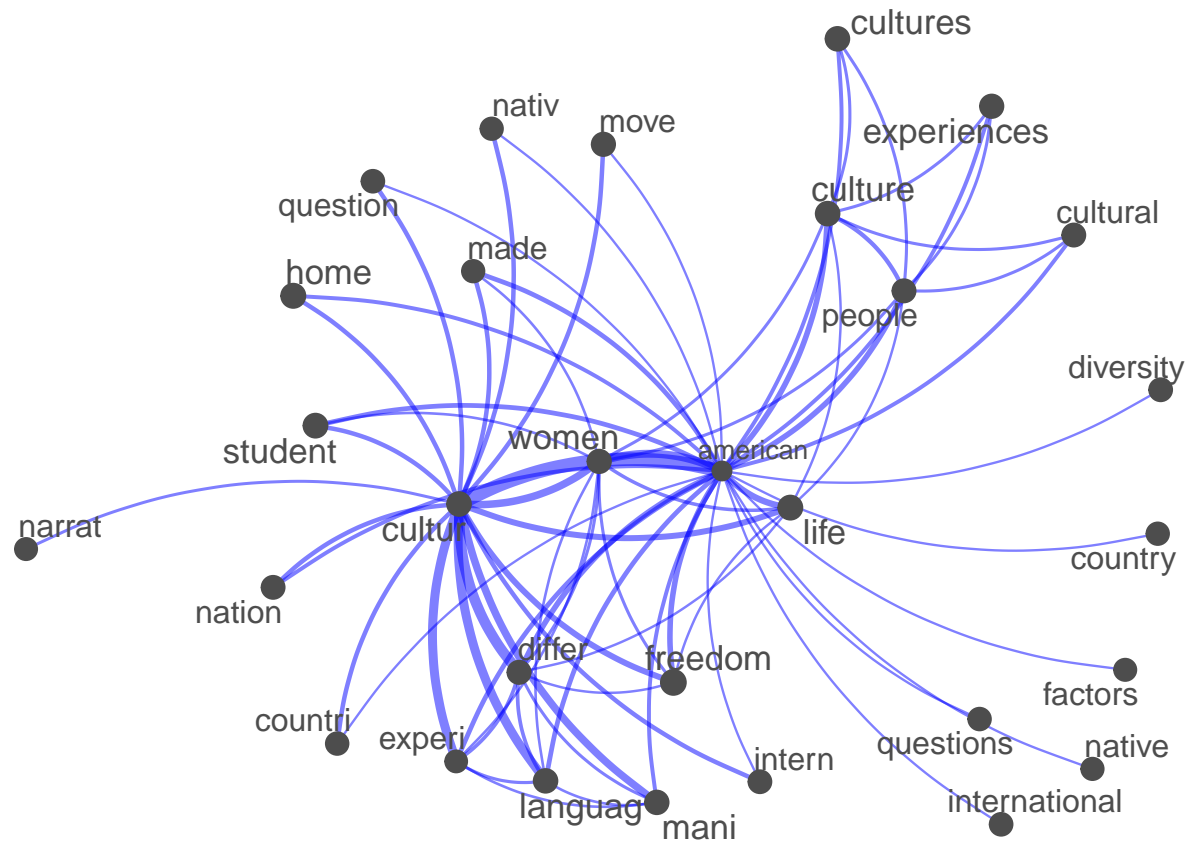
There are many strong connections in Group 7's network, including among "women", "culture", "language", "experience", and "American". This reveals some of the commonalities of the women's feelings after immigrating.

Group 8
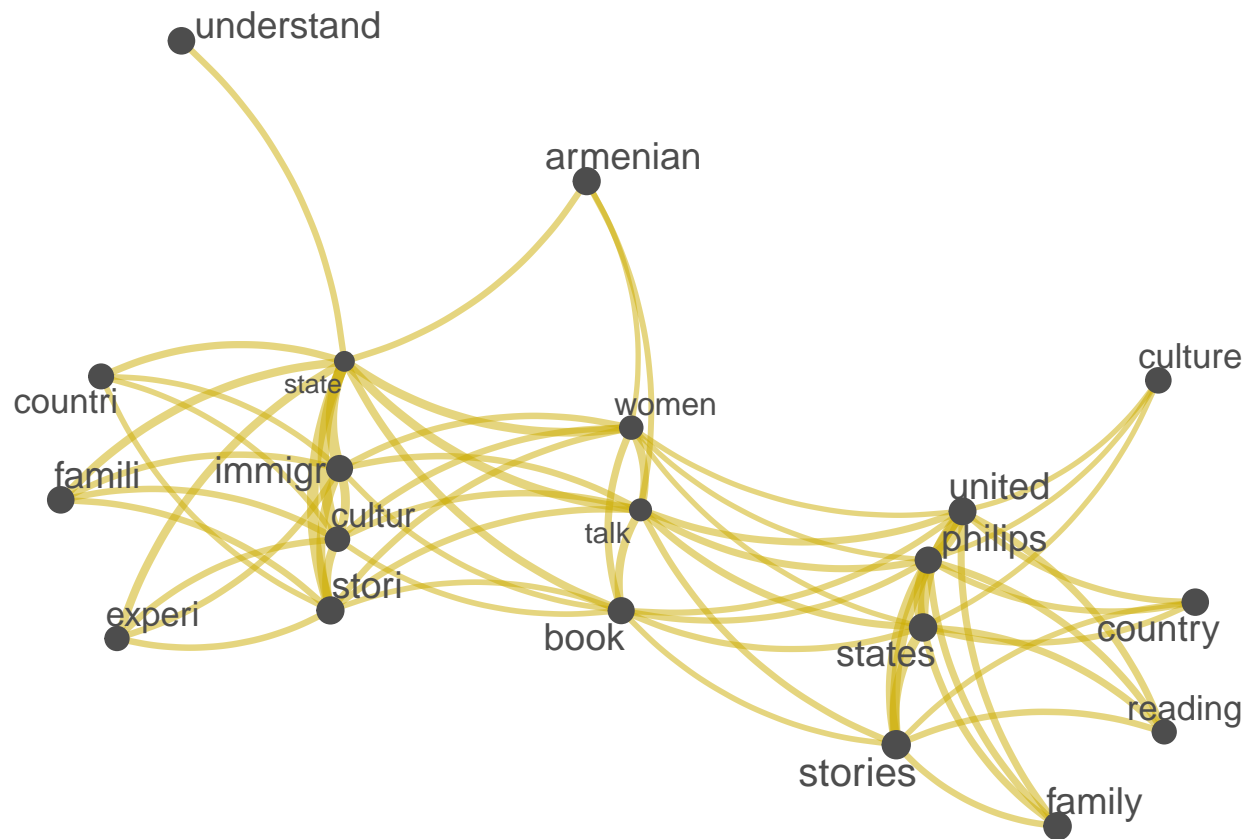
Words like "Philips", "United", and "States" are expected to be close and strongly connected because of the purpose of the book, along with "immigrate" and "culture".

Solo

restaurant

family
people
immigrants
recognized
back
door
invisible
care
career
korea
york
invis
famili
rachel
common
heart
recogn
desir
understand
anoth
hold
mention
citizen
small
wear
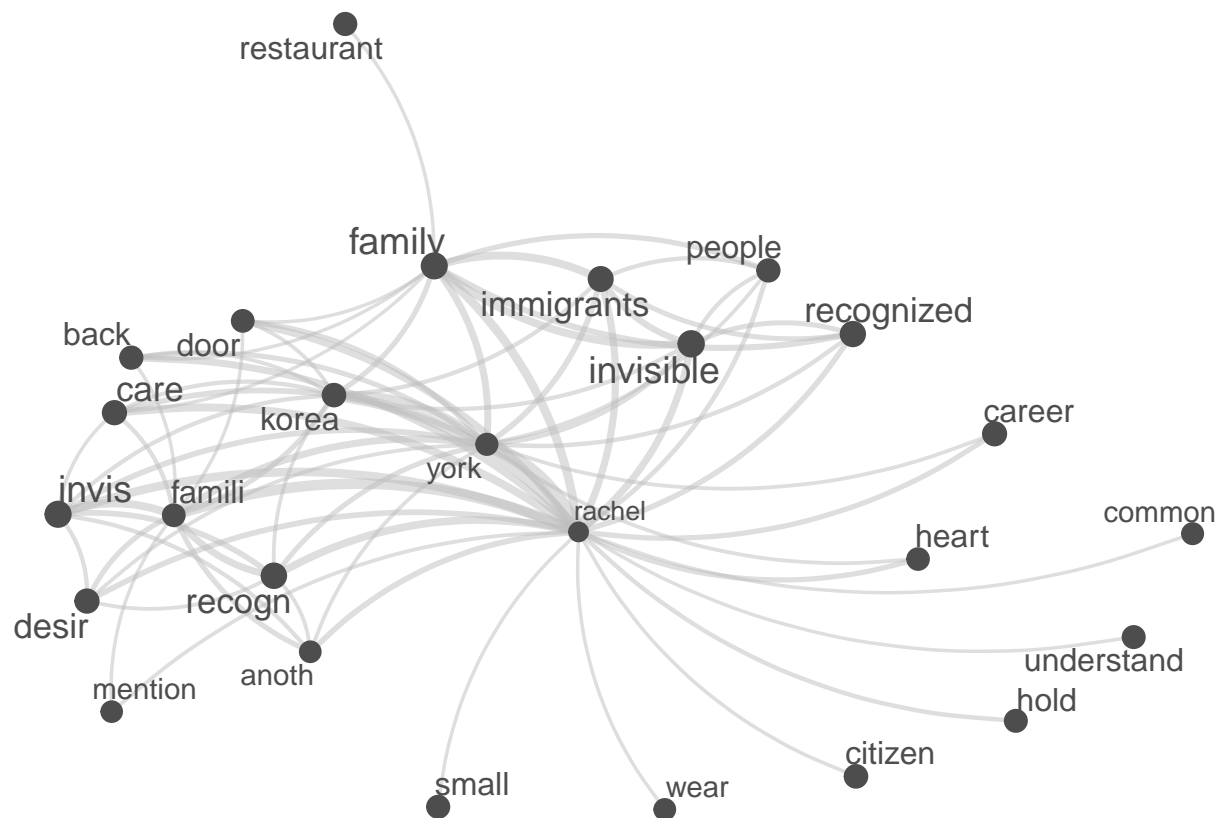
Because the subject of the interview was Rachel, her name shows up predominantly in the center of the network, with lines connected to "invisible", "recognize", and "care".

Combination
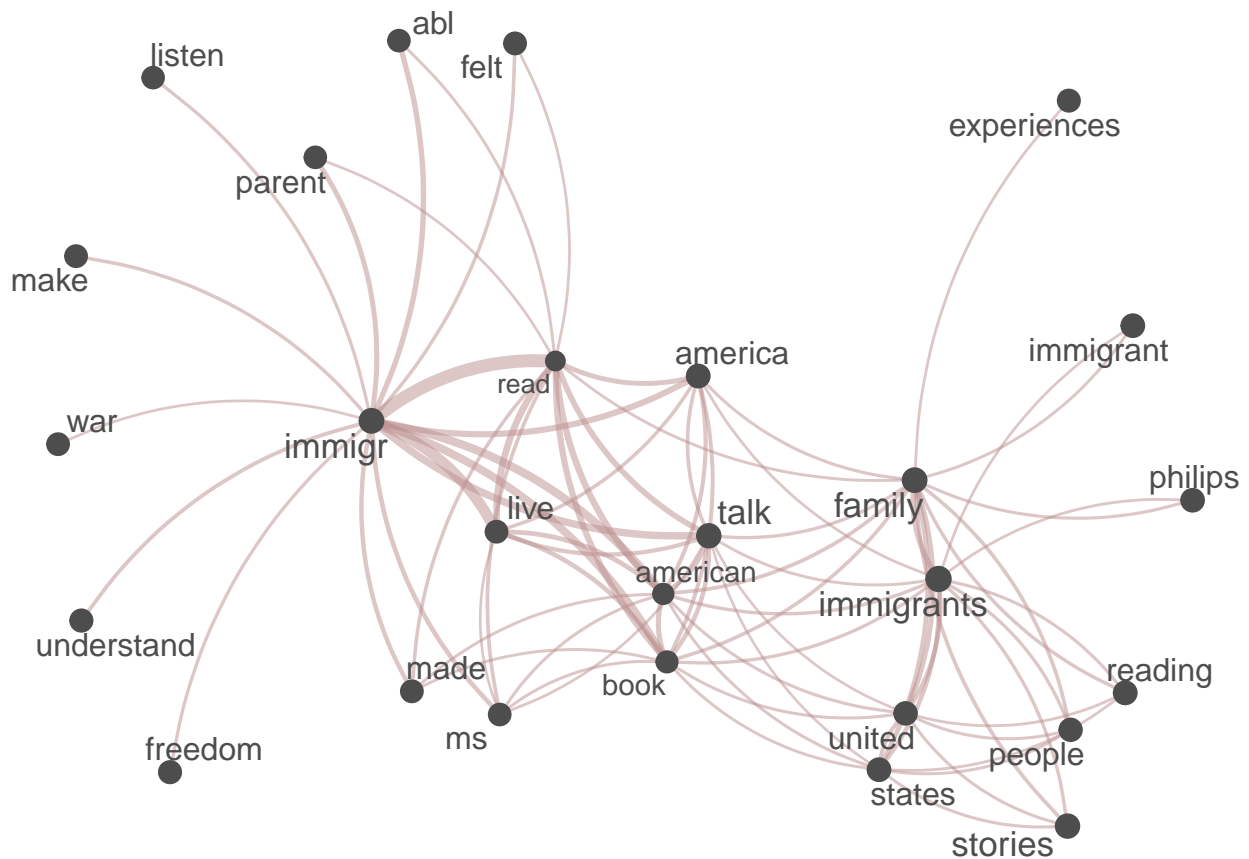
Some of the connected words found in all of the documents were "family", "live", "immigrants", and "America". Words on the outside of the network like "felt", "war", and "freedom" reveal further themes in all of the stories.

Cluster Dendrogram

Cluster dendrograms present a visual flowchart to represent the relationships between words, with words at the top branching down to words that they are often found with. Words stemming from the same branch are often found directly paired with each other.

```
## <<TermDocumentMatrix (terms: 1473, documents: 1)>>
## Non-/sparse entries: 1473/0
## Sparsity           : 0%
## Maximal term length: 17
## Weighting          : term frequency (tf)
## Sample             :
##        Docs
## Terms   Combination.txt
##   cultur          69
##   experi          49
##   famili          77
##   immigr         126
##   life            49
##   live            59
##   read            70
##   state           54
```

```
##   stori            69
##   unit             53
```

Cluster Dendrogram

d
hclust (*, "ward.D")
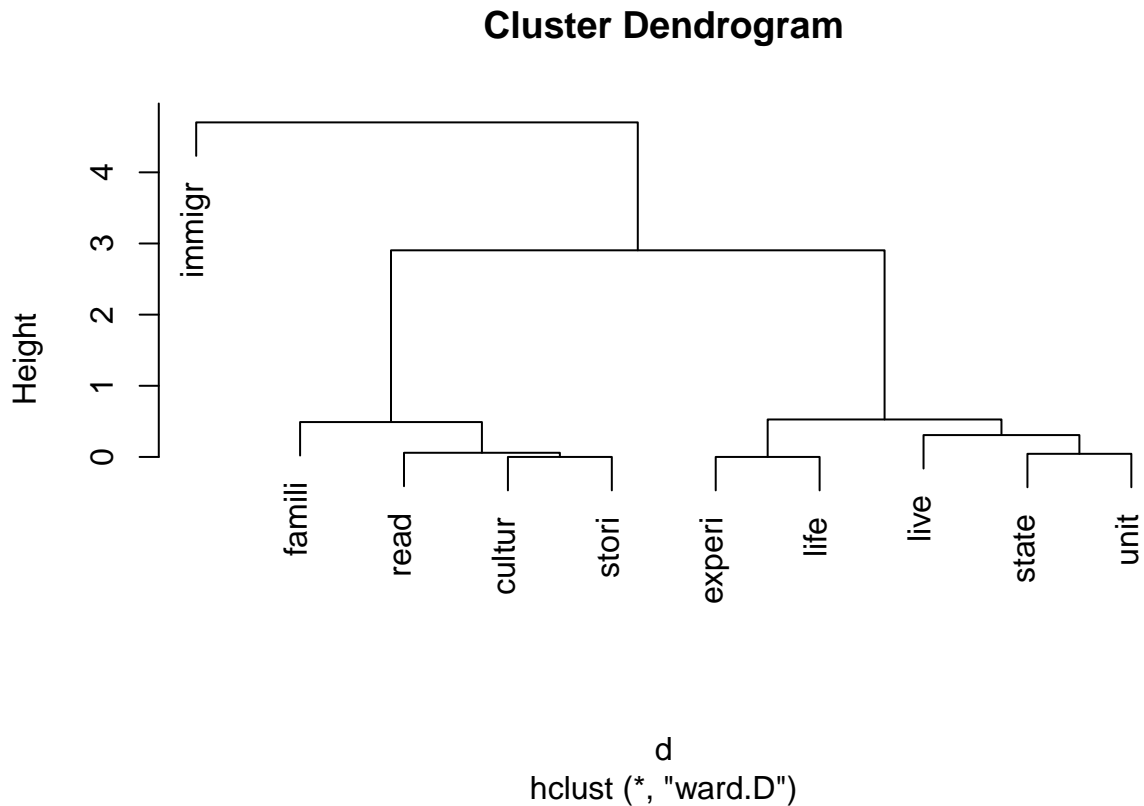
From this, we can see that the stem "immigr" is very commonly found. The words "United" and "States" are paired together, along with "life"/"experience" and "culture"/"story". Grouped with the first two pairs is "live", and grouped with the other pair is "read" and "family".

Topics Model

Topics modeling is a form of unsupervised learning that classifies words without having any references for potential grouping. Topics models are often fitted using latent Dirichlet allocation (LDA), which treats text as a mixture of topics, topics as a mixture of words, and mathematically determines which words make up a topic and which topics make up a text.
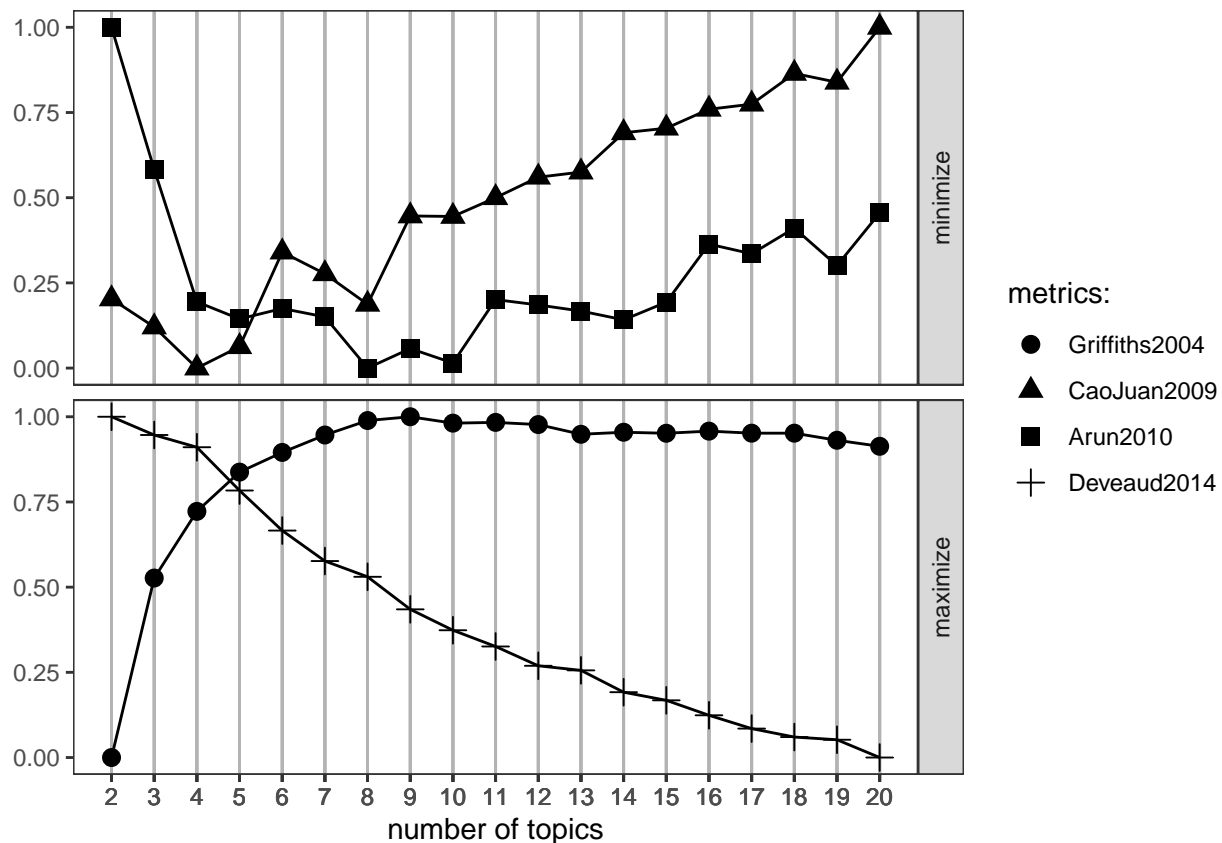
By finding the intersection of the "maximize" graph above, we see that the optimal number of topics is 5.

```
##         Topic 1       Topic 2         Topic 3     Topic 4          Topic 5
## [1,]  "family"      "back"          "american"  "another"        "think"
## [2,]  "immigrants"  "maria"         "mother"    "know"           "immigrant"
## [3,]  "united"      "immigration"   "talk"      "understanding"  "experiences"
## [4,]  "life"        "america"       "xidan"     "story"          "way"
## [5,]  "states"      "freedom"       "like"      "fact"           "live"
```

The model organizes important words into these 5 themes. Topic 1 gives a broad overview, topics 2 and 3 focus on Maria and Xidan specifically, topic 4 highlights more personal words like "understanding" and "story", and topic 5 includes "experiences", "way", and "live".

VI. Conclusion

Through this project, we analyzed the common themes of what different groups got from Susan Philips's book and talk. Based on her name coming up often, Xidan's immigration story seemed to resonate among groups even though her chapter of the book was not given. Each group took away something slightly different, but common themes included women, education, freedom, and culture.

VII. Shortcomings and Suggestions

One shortcoming is the small sample size we were working with. An analysis of all papers that have ever been written in this class for this book may provide a more thorough insight into what students learn from this lesson. Additionally, because of the open-endedness of the questions asked for the reflection paper, groups took very different approaches in their answers. For example, some groups focused more on Susan Philips's presentation, while others only wrote about things they read in her book. Finally, textual data is not as

straightforward to analyze as numerical data, so the models and conclusions in our analysis are somewhat abstract. A suggestion we have moving forward would be for the reflection paper questions to ask each individual not just about the chapters they read but also about the talk itself. Many groups chose to write about the talk in response to the group question about what they learned, so they may have had more to say individually about the talk and discussion.