Cornell Sun Age Classification

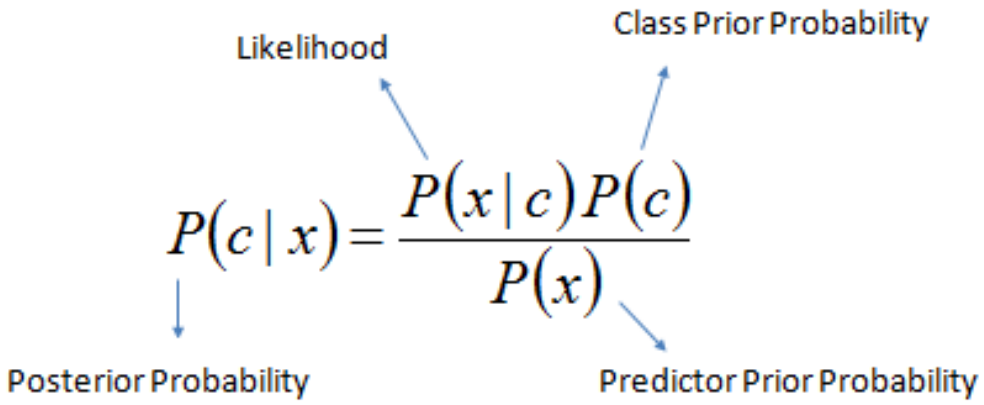
Mindy Lou and Chris Sciavolino

Project Overview

- Classify a Cornell Daily Sun article's likely readership
- Use data obtained by Google Analytics stored by the Cornell Daily Sun
- Naïve Bayes Classification model
- Create an iOS app that displays the classification information dynamically to the user

Model Overview

- Naïve Bayes Classification model
- Features: Words that appear in the content of an article
- Example:
    - { “hello”: 1, “world”: 2, “Cornell”: 3 }
    - [hello, world, world, Cornell, Cornell, Cornell]



$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Obtaining Training Data

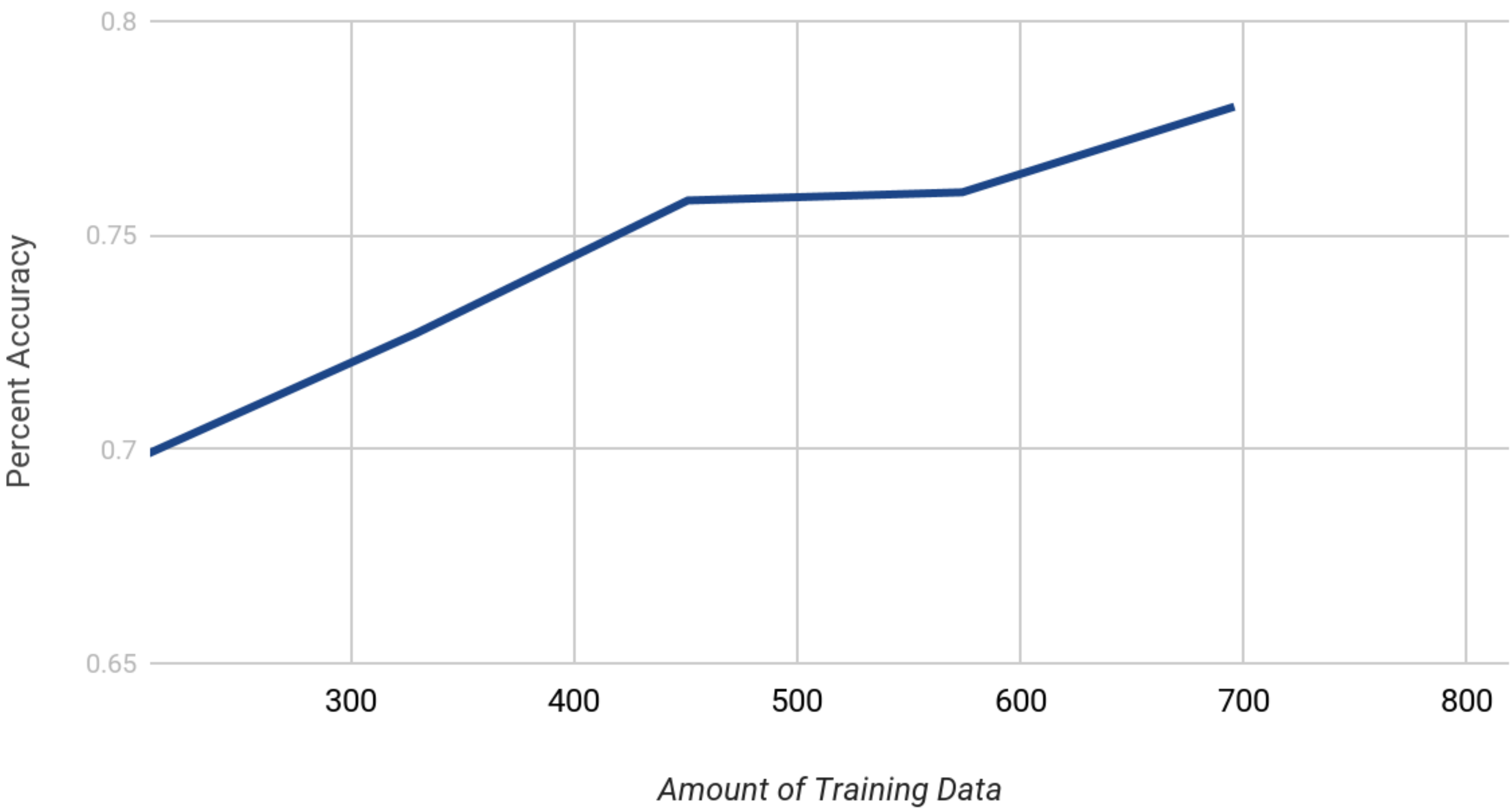
- Google Analytics only stores the title and sessions counts
- Strip out unnecessary unicode characters in the title
- Take the title and make an API request to the Cornell Sun’s WordPress account
  - Only search results to took the best result and checked for a rendered title match
- Counted the words and wrote the word counts to a text file
- Manually classify the Google Analytics articles by counting the sessions in each age group

Training the Model in Swift

- Use the .observe(…) method in the Bayes module
- Takes in the classification and the list of observed words
    - Ex. “18-24” and [hello, world, world, Cornell, Cornell]
- Read in both the classifications text file and the word counts text file into the app
- Observe each of the articles when training with all the words associated with said article

Judging Model Accuracy

- Total of 819 articles found and classified
- 70% of data used to train the model (574 articles)
- 30% of data used to test the model (245 articles)
- Accuracy of 76% on the testing data after training
- Given more training data, the model generally performed better

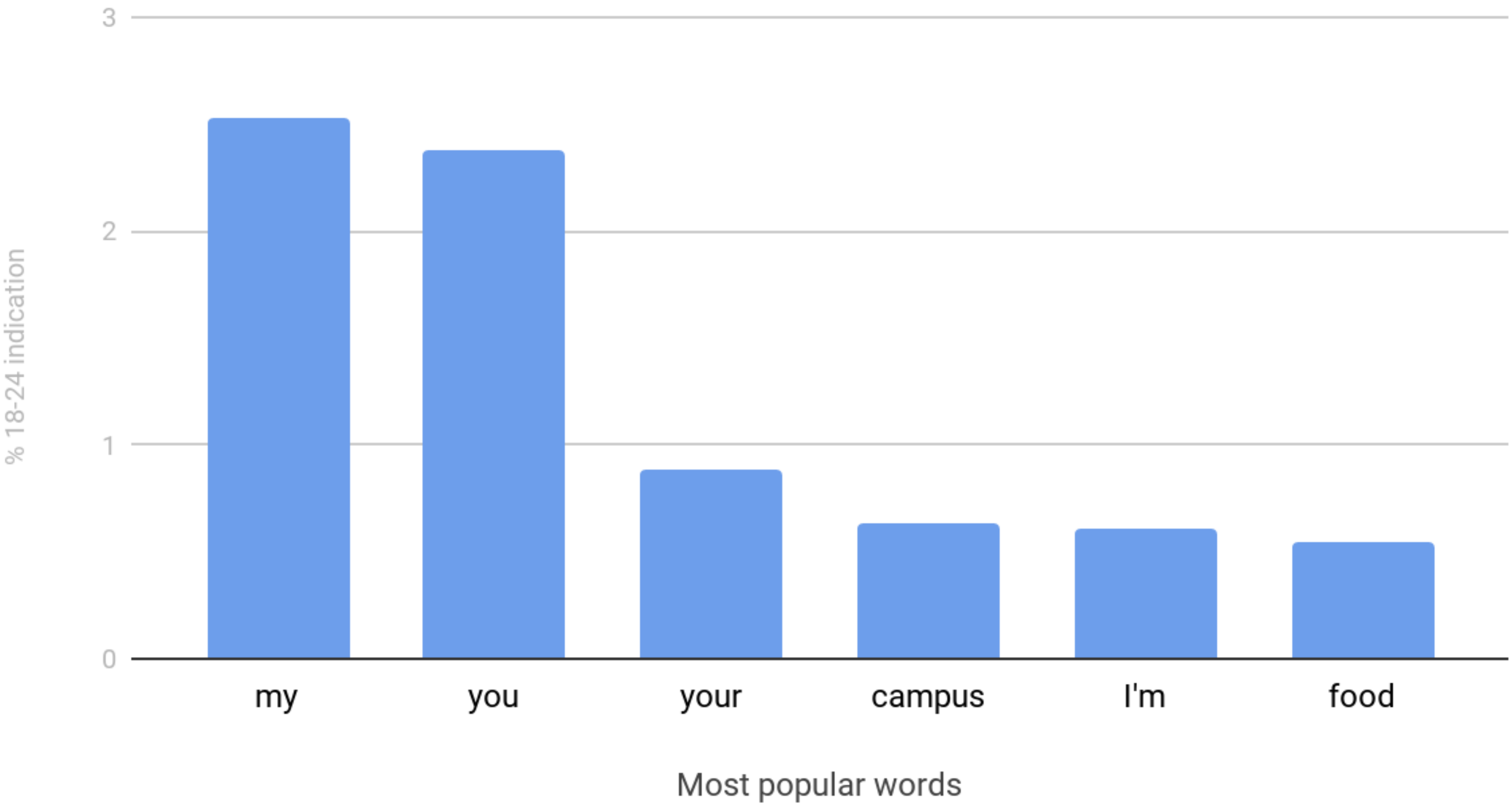


The Accuracy of our Naive Bayes Model

Analyzing Indicator Words

- Attempted to identify words that indicated the article’s classification
- Found the probability a word appears given a certain classification
- Subtracted off the equivalent probabilities for the other classifications
- If this value was greater than 0.05, deemed this word as an **indicator**
- Majority of articles containing this word classified accordingly

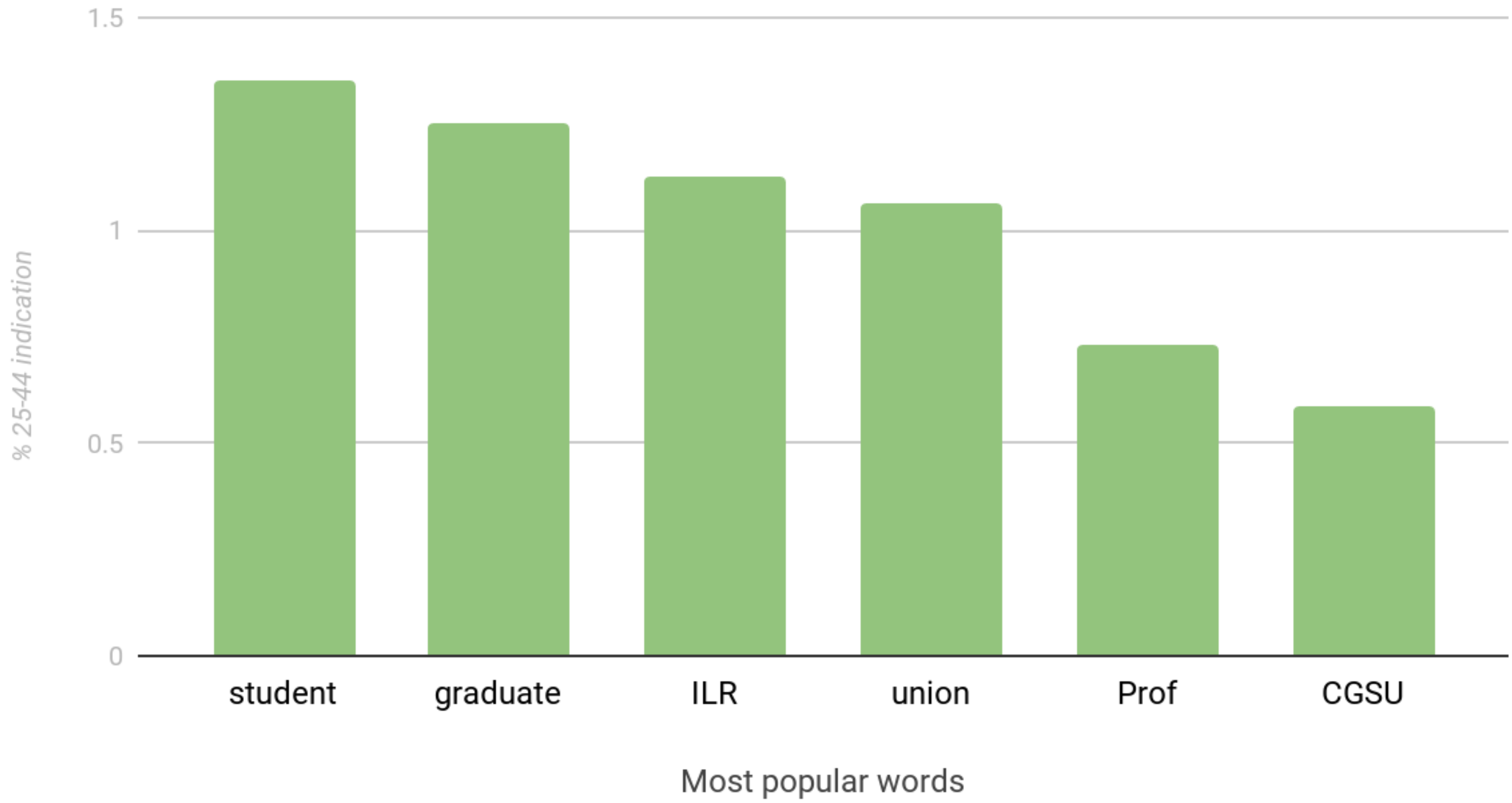
Most Popular Words Read By Ages 18-24



18-24 Classification Conclusions

- Articles containing casual language more likely to be consumed by younger audiences
  - “You”
  - “Your”
  - “My”
  - “Food”
- More close to campus and social-oriented

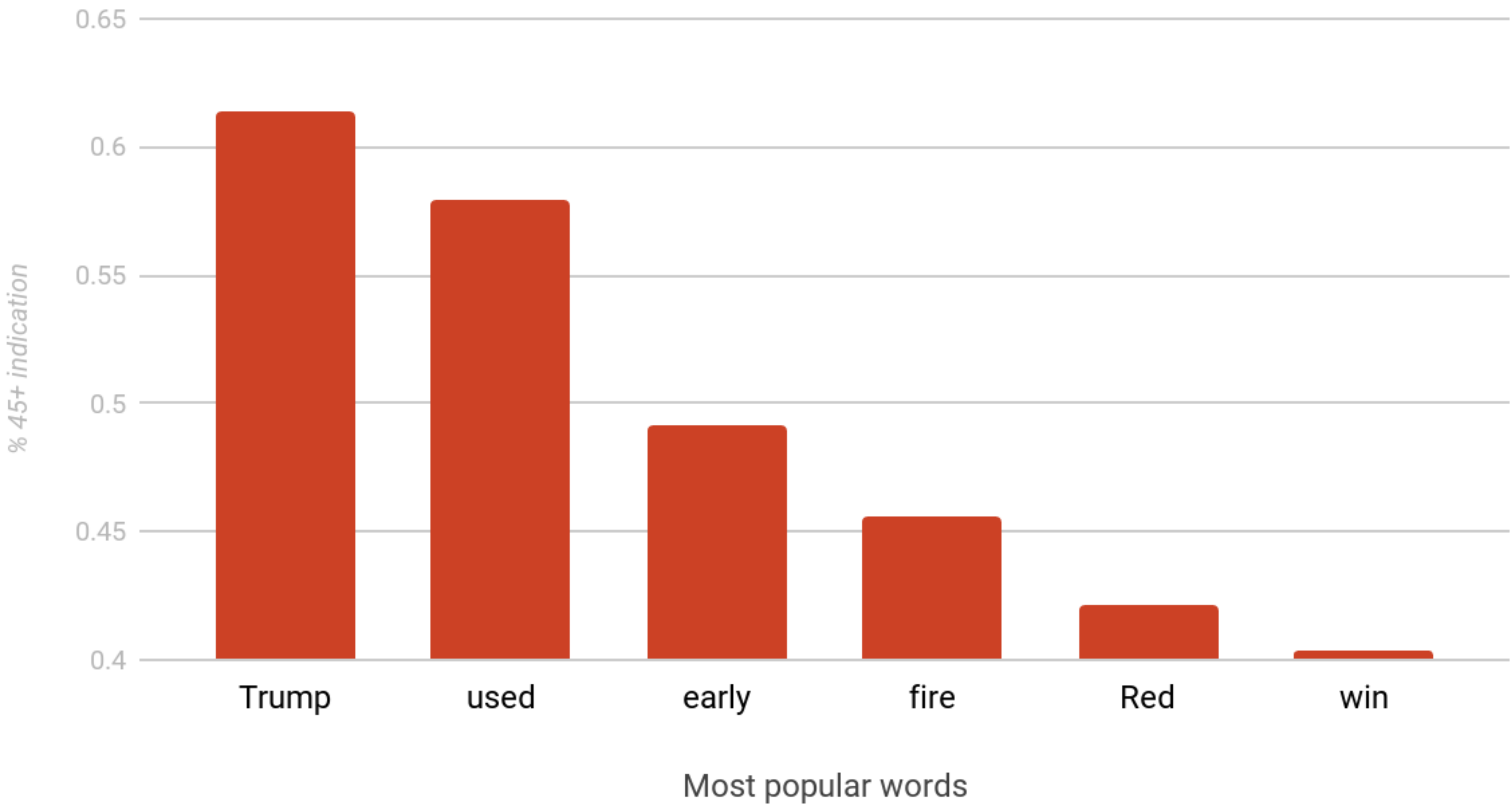
Most Popular Words Read by Ages 25-44



25-44 Classification Conclusions

- Articles containing language pertaining to economics, unionization, or academics
  - “Student”
  - “Graduate”
  - “Union”
  - “ILR”
- More academic or organization related

Most Popular Words Read by Ages 45+

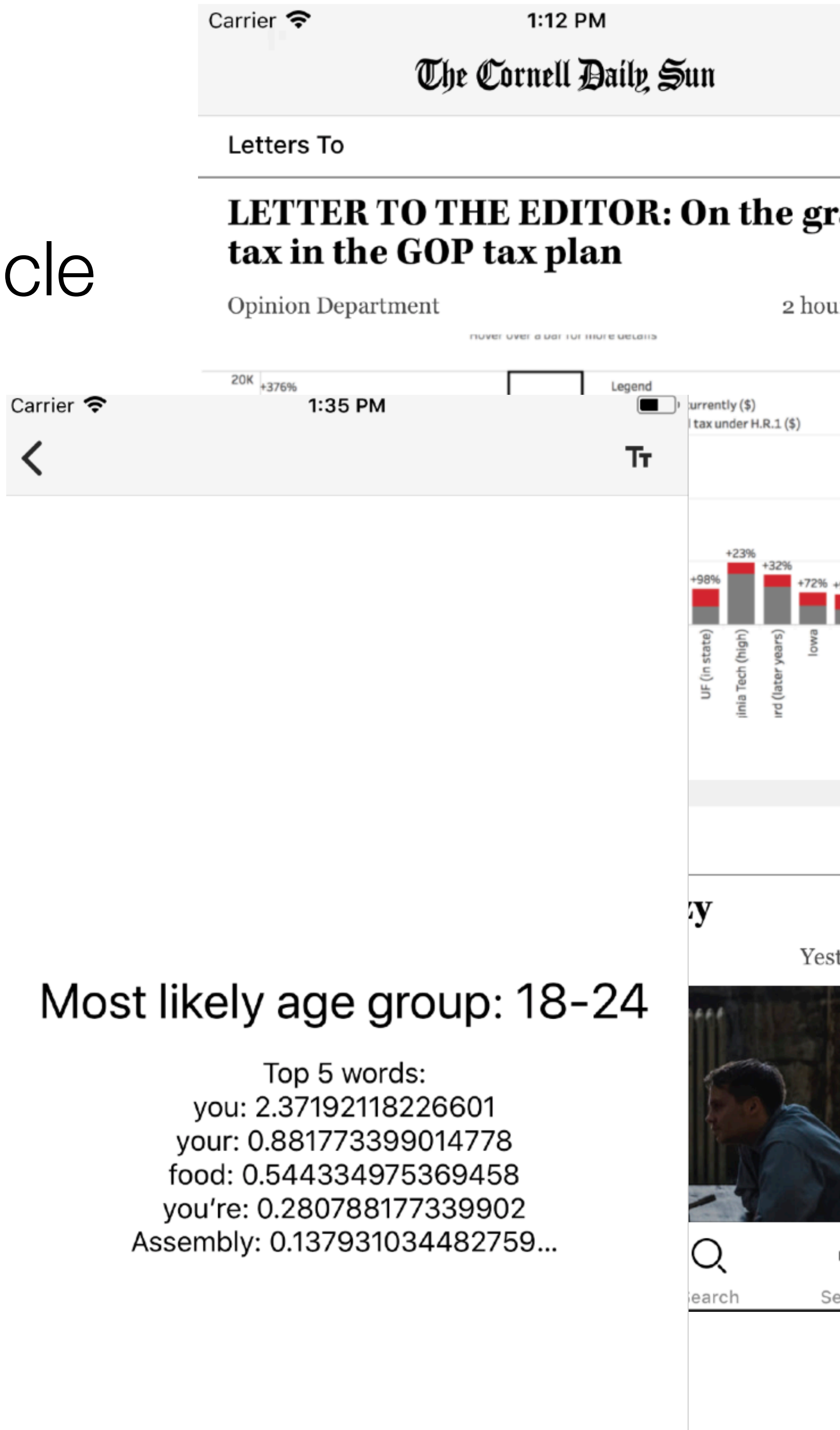


45+ Classification Conclusions

- Articles containing language pertaining to politics or local news events
  - “Trump”
  - “Fire”
  - “Used”
- More oriented around politics or news stories

iOS Application and Demo

- Tap on article
- Shows statistics and classification for article
  - Most likely classification
  - Top 5 words
  - Corresponding weight for each word



LETTER TO THE EDITOR: On the gr
tax in the GOP tax plan

Most likely age group: 18-24

Top 5 words:
you: 2.37192118226601
your: 0.881773399014778
food: 0.544334975369458
you're: 0.280788177339902
Assembly: 0.137931034482759...