EDLD652 Final

Diana Dewald, Eliott Doyle

2/10/2022

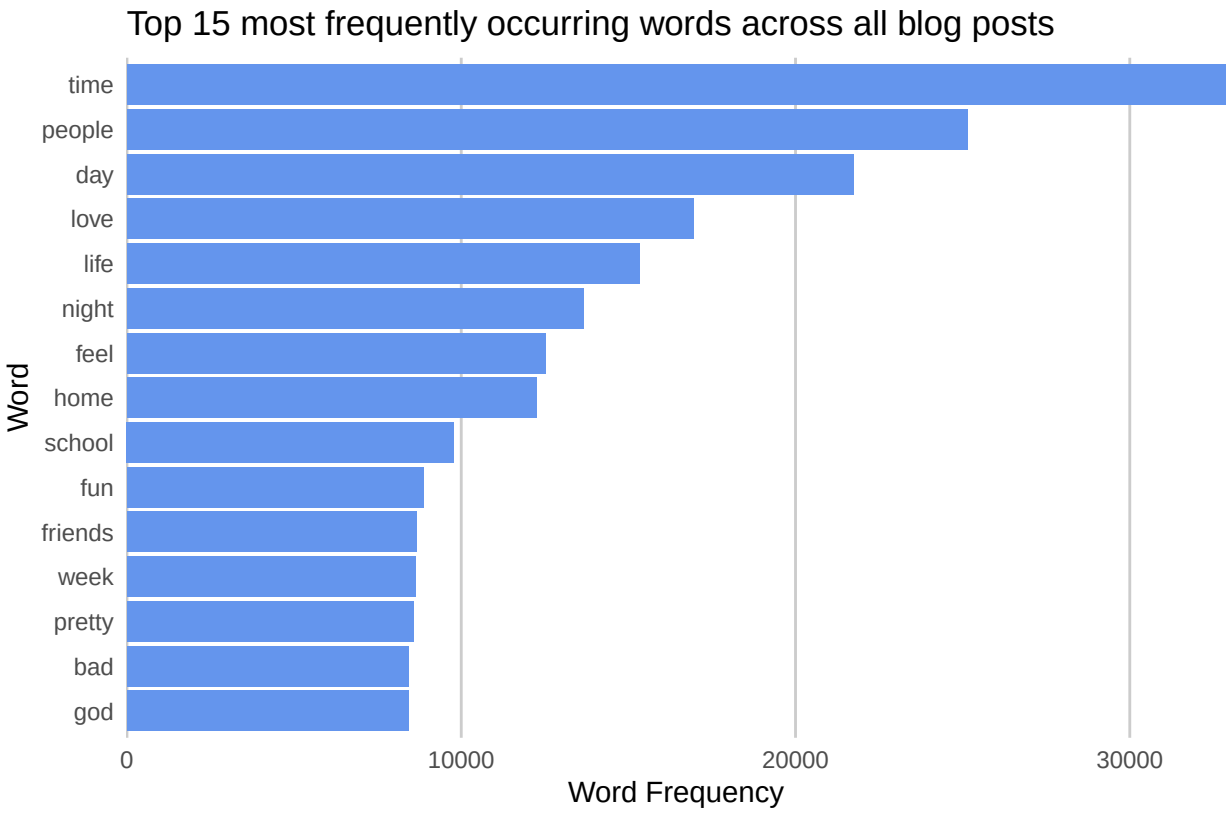
Description of the data

The data used for this project are from https://www.kaggle.com/rtatman/blog-authorship-corpus, a text dataset of approximately 681000 blogposts. For the purposes of this project, we will be working with one tenth of the full dataset. In addition to the text of each blog post, some information about each the post (topic, date posted) and blogger (gender, age, astrological sign) is included.

Research question 1:

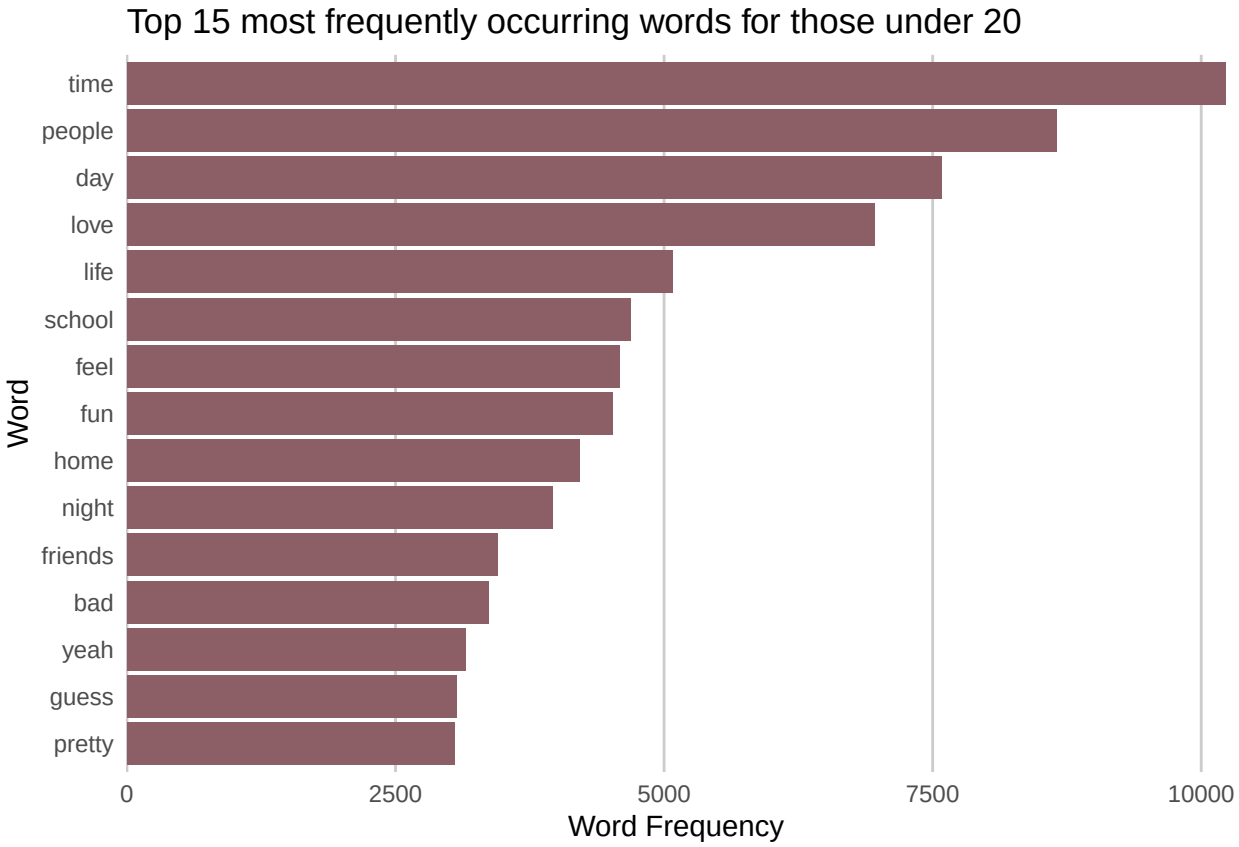
What are the most frequently used words in all writing samples?

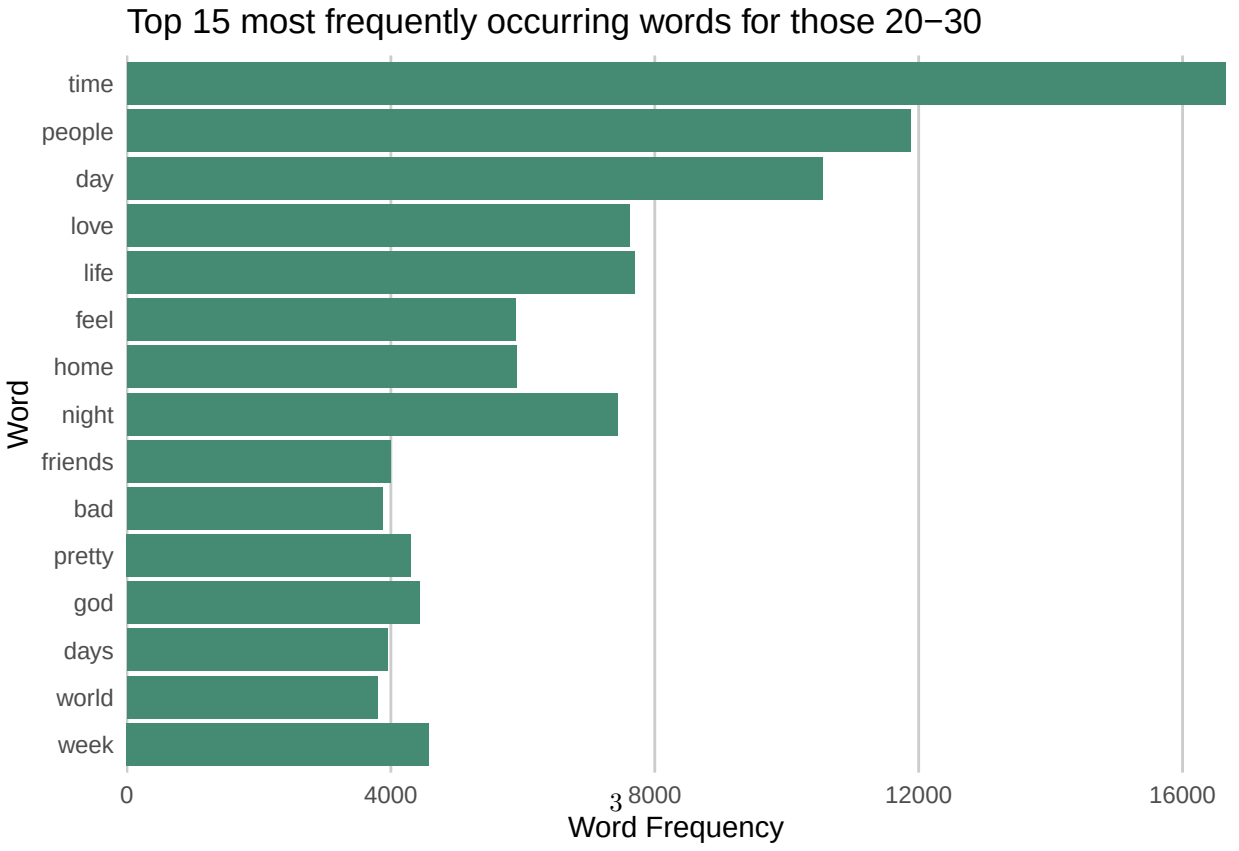
```
##   V1      id gender age   topic sign        date agegroup  word
## 1  1 2059027   male  15 Student  Leo 14,May,2004 Under 20  info
## 2  1 2059027   male  15 Student  Leo 14,May,2004 Under 20   has
## 3  1 2059027   male  15 Student  Leo 14,May,2004 Under 20  been
## 4  1 2059027   male  15 Student  Leo 14,May,2004 Under 20 found
## 5  1 2059027   male  15 Student  Leo 14,May,2004 Under 20 pages
## 6  1 2059027   male  15 Student  Leo 14,May,2004 Under 20   and
```

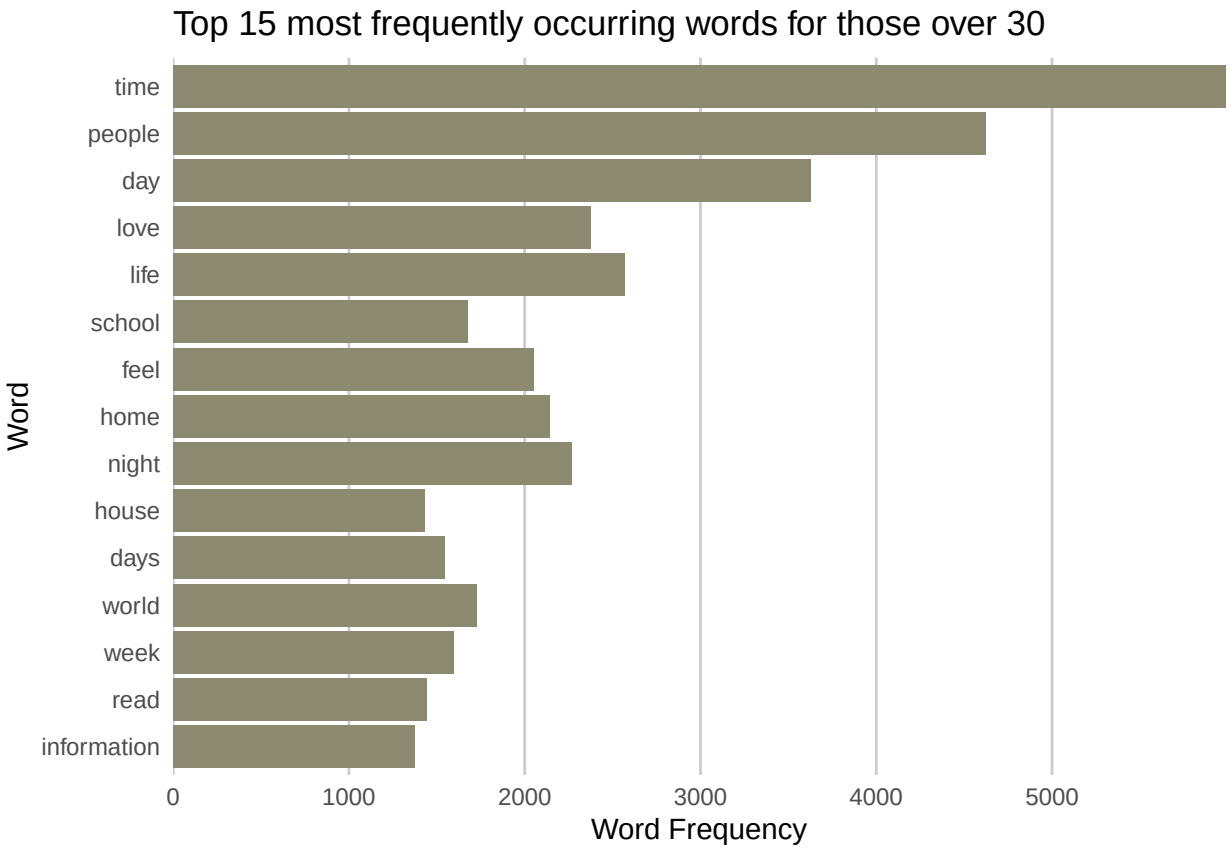


Data abbreviated from https://www.kaggle.com/rtatman/blog−authorship−corpus

Research questions 2a: What are the most frequently used words by age group? Which words are unique to certain age groups?






Top 15 most frequently occurring words for those over 30
Word
time
people
day
love
life
school
feel
home
night
house
days
world
week
read
information
0
1000
2000
3000
4000
5000
Word Frequency

Research question 2b: What is the distribution of topics discussed?

Research question 3:

Is there a relationship between blog post length and date posted?