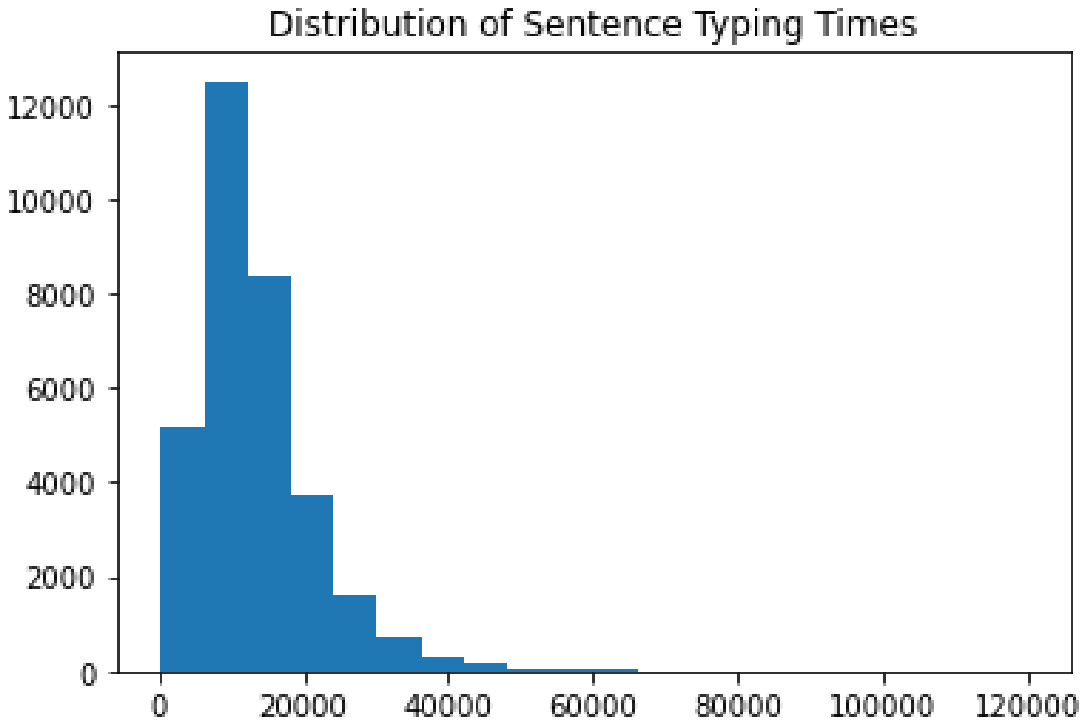
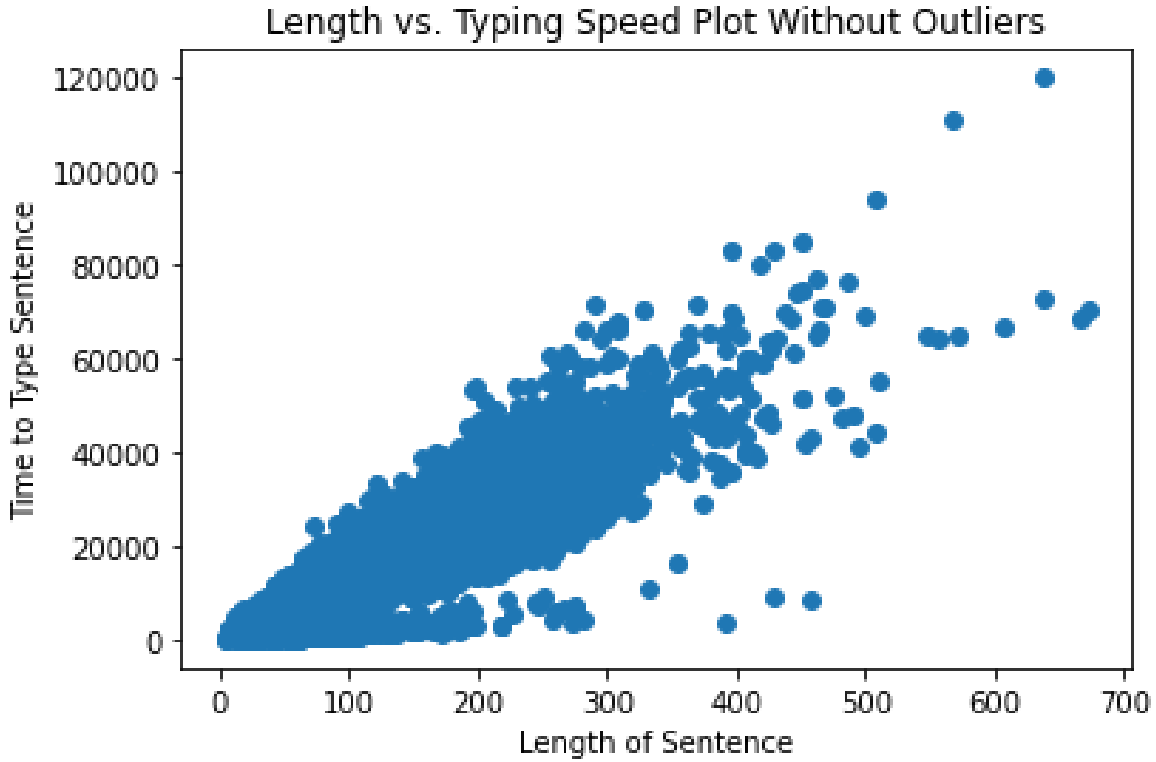
Capstone Project by Anand Ramakrishnan

Predicting the Typing Time of a Sentence

Background

- Professional typing tests, which assess words per minute, are a common part of job interviews.
- **Does any aspect of the sentence cause variation in typing speed?**
- The dataset was taken from a study by Nahuel Gonzalez of the University of Buenos Aires, in which the content of a sentence was predicted based on the timing of each keystroke. Since there were no time units listed in the data, I decided that **total typing time** would be the dependent variable.
- There are 32,716 sentences in total, split between those about gay rights, those about gun rights, and restaurant reviews.
- I cleaned the data to remove and account for function keys.


Distribution of Sentence Typing Times
12000
10000
8000
6000
4000
2000
0
0
20000
40000
60000
80000
100000
120000


Length vs. Typing Speed Plot Without Outliers
Time to Type Sentence
120000
100000
80000
60000
40000
20000
0
0
100
200
300
400
500
600
700
Length of Sentence

Plots

Word Cloud

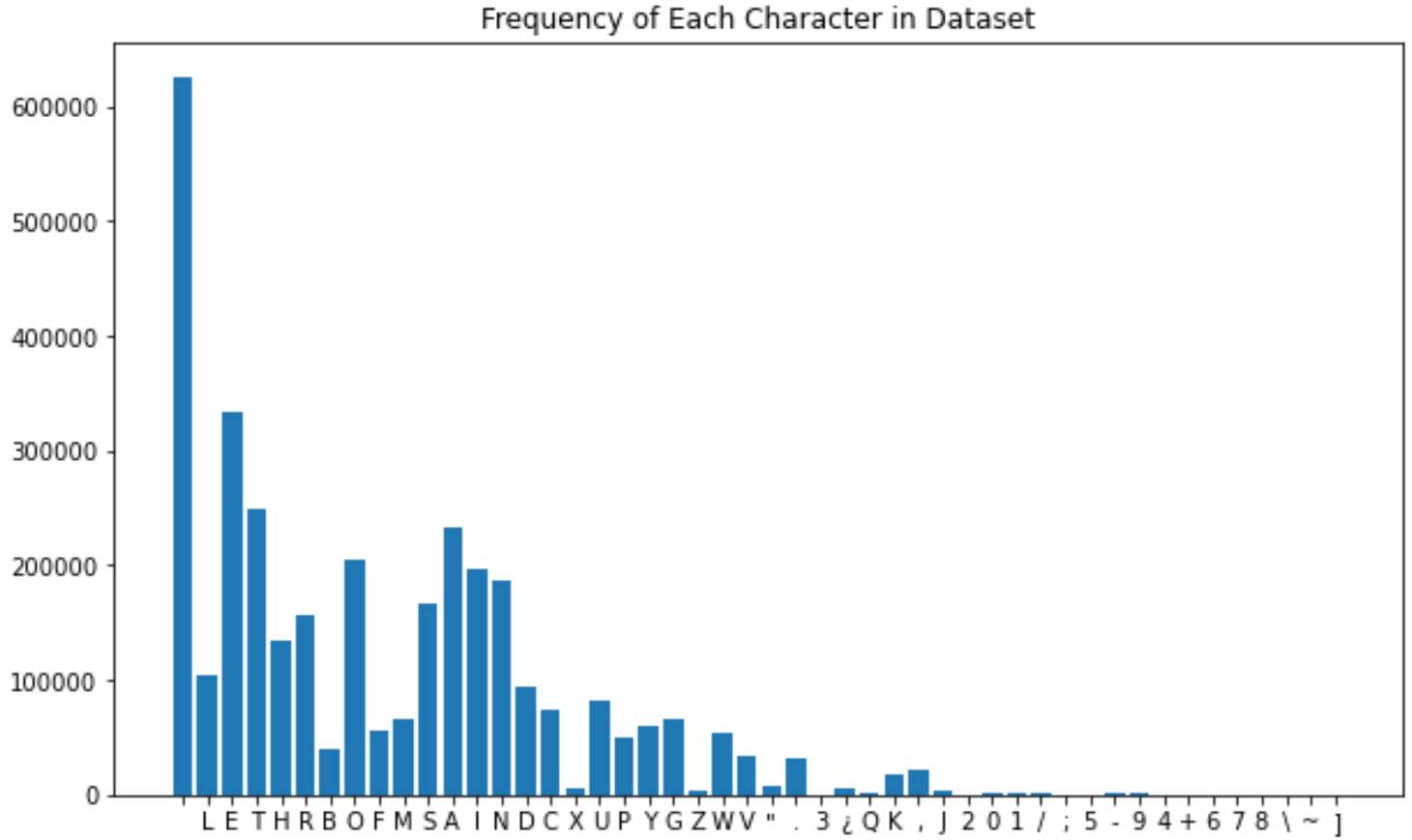

GUN CONTROL
LEGAL
CANOT
EVERYONE
MARIED
ABLE
THINK
WEL
FEL
AL
S
WITHOUT
WANT
GO
PROTECT
SOCIETY
EVERY
PERSON
GOVERNMENT
ALWAY
FAMILY
WIL
PLACE
FREDOM
WOMAN
USE
DON T
FRIEND
HAND
MUST
GAY
MAKE
AMERICAN
NED
INDIVIDUAL
CHILDREN
SERVICE
FOD
MANY
ORDER
HOME
GUN
THING
MAN
WEAPON
COUPLE
EVEN
FIRST
NEVER
DELICIOU
WAY
T
LIVE
WORLD
MUCH
KNOW
SOMETHING
SOMEONE
MARIAGE
BELIEVE
STATE
STAF
REALY
COUNTRY
MARY
GREAT
BEAR ARM
PEOPLE
ANYONE
RIGHT
SAY
GOING
USED
TRY
ALOWED
FIREARM
CITIZEN
BEST
DAY
LAW
LOVE
EAT
LIFE
TIME
U
MEAL
ONE
BEN
FRESH
AMERICA
MEAN
MADE
COME
TAKE
REASON
PRICE
GOD
CRIMINAL
MENU
GAY
MARIAGE
RESTAURANT
PIZA
FRIENDLY

Word Analysis

- Four models were fit at each step of the analysis process: **linear regression**, **decision tree regression**, **bagged regression**, and **random forest regression**. The main statistic for success was **accuracy**.
- After splitting the dataset into training and test sets, I used a CountVectorizer to tally the counts of each word that appeared at least 33 times, or roughly 1% of the total number of sentences, in the training set.
- Using the word counts as predictors of typing time, the model that fit the best was linear regression, though it only had above-average accuracy for both the training and test sets. The other three models were very overfit.
- Choosing only words which are not real US English words, the models performed even worse.

Attribute Analysis

- For my next group of models, I used these attributes as predictors: the length of the sentence, the topic of the sentence, the number of function keys (such as shift or backspace) used, and the sentiment of the sentence.
- The sentiment was calculated using the VADER tool, on a scale of -1 (most negative) to +1 (most positive).
- The random forest regressor performed the best, though its parameters needed to be fine-tuned afterwards in order to limit overfitting.
- The length was easily the most important attribute in the model, but both sentiment and the number of function keys also had a significant effect.

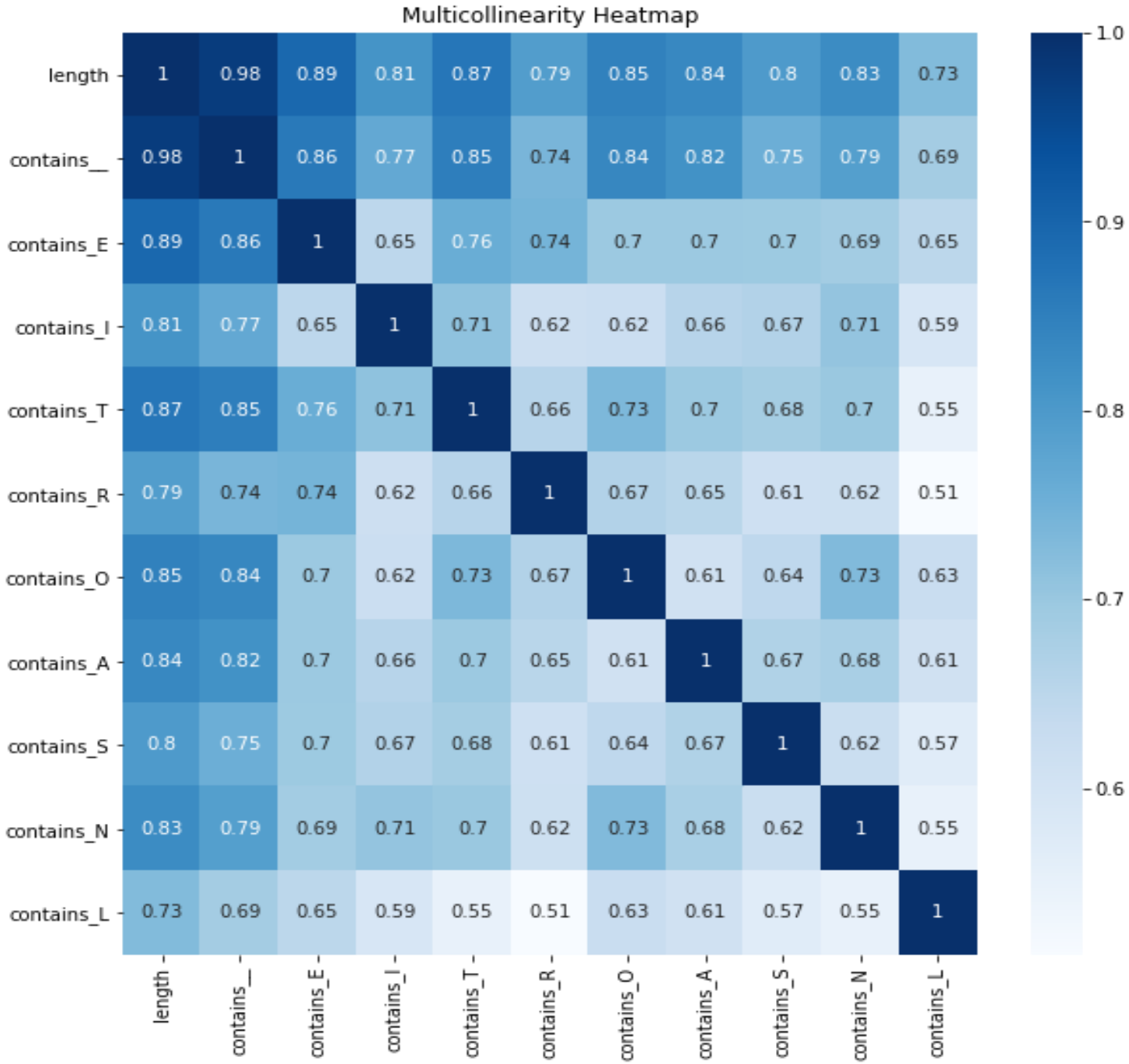


Character Analysis

- Using the Counter function in Python's Collections library, I found out the characters contained in the sentences as well as how much each character appeared in the sentences.
- The random forest regressor performed the best again.
- The features with the highest coefficients were space and nine of the ten most common letters.

Multicollinearity

- The frequencies of the 10 most important characters had high correlation with the length of the sentence. In addition, the frequencies of the 9 letters had high correlation with the number of occurrences of the space. However, the frequencies of the letters did not have high correlation with each other.
- Since multicollinearity (high correlation between predictors) can lead to a wider confidence interval, it is best that we remove highly-correlated predictors. Thus, the length and number of spaces will no longer be included.


Multicollinearity Heatmap
length 1 0.98 0.89 0.81 0.87 0.79 0.85 0.84 0.8 0.83 0.73
contains__ 0.98 1 0.86 0.77 0.85 0.74 0.84 0.82 0.75 0.79 0.69
contains_E 0.89 0.86 1 0.65 0.76 0.74 0.7 0.7 0.7 0.69 0.65
contains_I 0.81 0.77 0.65 1 0.71 0.62 0.62 0.66 0.67 0.71 0.59
contains_T 0.87 0.85 0.76 0.71 1 0.66 0.73 0.7 0.68 0.7 0.55
contains_R 0.79 0.74 0.74 0.62 0.66 1 0.67 0.65 0.61 0.62 0.51
contains_O 0.85 0.84 0.7 0.62 0.73 0.67 1 0.61 0.64 0.73 0.63
contains_A 0.84 0.82 0.7 0.66 0.7 0.65 0.61 1 0.67 0.68 0.61
contains_S 0.8 0.75 0.7 0.67 0.68 0.61 0.64 0.67 1 0.62 0.57
contains_N 0.83 0.79 0.69 0.71 0.7 0.62 0.73 0.68 0.62 1 0.55
contains_L 0.73 0.69 0.65 0.59 0.55 0.51 0.63 0.61 0.57 0.55 1
length
contains__
contains_E
contains_I
contains_T
contains_R
contains_O
contains_A
contains_S
contains_N
contains_L
1.0
0.9
0.8
0.7
0.6

Final Model

- The final models had 21 predictors:
  - The 10 words with the highest coefficients for linear regression: 'thethe', 'tha', 'health', 'consenting', 'sandy', 'wouldn', 'hte', 'fof', 'considering', and 'walked'
  - The 9 letters with the highest feature importances in the character random forest regression model: 'E', 'I', 'T', 'R', 'O', 'A', 'S', 'N', and 'L'
  - The number of function keys used while typing each sentence
  - The sentiment of each sentence
- The eventual model chosen was a **random forest regressor**.
- The model had higher variance if the 10 words were not included, indicating that the words helped to prevent overfitting.

Table of Accuracy Scores

Model (set)	Words as predictors	Attributes as predictors	Characters as predictors	Final choice of predictors
Linear (training)	0.719	0.854	0.82	0.85
Linear (test)	0.672	0.858	0.827	0.854
Decision tree (training)	0.998	0.987	0.999	0.999
Decision tree (test)	0.384	0.722	0.689	0.714
Bagged (training)	0.925	0.959	0.965	0.97
Bagged (test)	0.597	0.827	0.822	0.849
Random forest (training)	0.946	0.967/0.883	0.976	0.979
Random forest (test)	0.624	0.836/0.851	0.838	0.862

Conclusions & Limitations

- The letter E had a much higher impurity-based feature importance than any other predictor in my final model, with a value of 0.569. The only other predictor with an importance higher than 0.1 was the number of function keys, which had a value of 0.123.
- Possible reasons for the dominance of the letter E are that E is such a common letter in the English language and that it is not in the center row of the QWERTY keyboard, so typists must do a little reaching to press the key.
- There are a few limitations to this project:

1. Since the data never had consecutive [BACK] instances, it is impossible to tell how many letters were deleted upon using backspaces, so my cleaned data might not be accurate.
2. The sentiment analyzer could not assign a score to misspelled words.