Capstone Project

Machine Learning Fundamentals

Samuel Bosshardt

12/15/2018

Questions

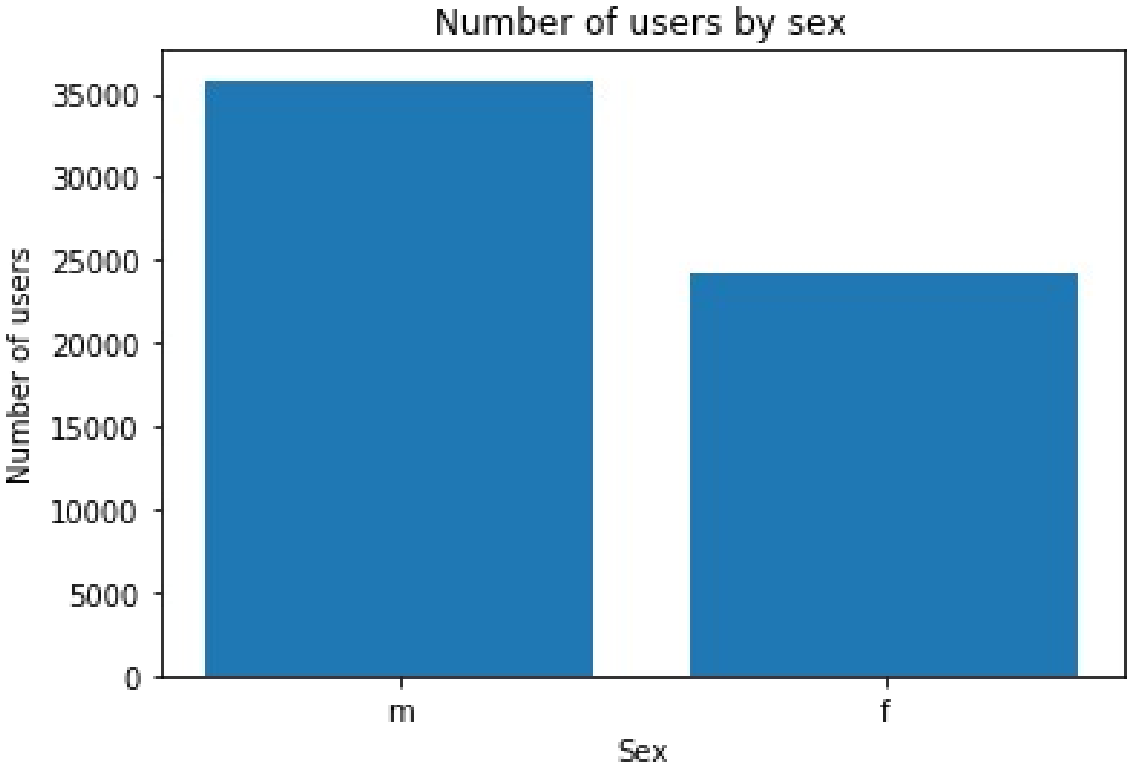
- Can the sex of the users be predicted by analyzing their essays?
- What machine learning models are most effective? (accurate and efficient)
- Although this is a classification task, how well can the regression models do?

Dataset Exploration

- Graph of number of males vs females.
- About 40% of the users are female.
- M: 35,829
- F: 24,117

```
from matplotlib import pyplot as plt
column = 'sex_int'
x_vals = df[column].unique().tolist()
plt.xticks(x_vals, df['sex'].unique().tolist())
plt.xlabel("Sex")
plt.ylabel("Number of users")
plt.title("Number of users by sex")
y_vals = []
for x_val in x_vals:
    vals_num = len(df[df[column] == x_val])
    y_vals.append(vals_num)

plt.bar(x_vals, y_vals, align="center")
```

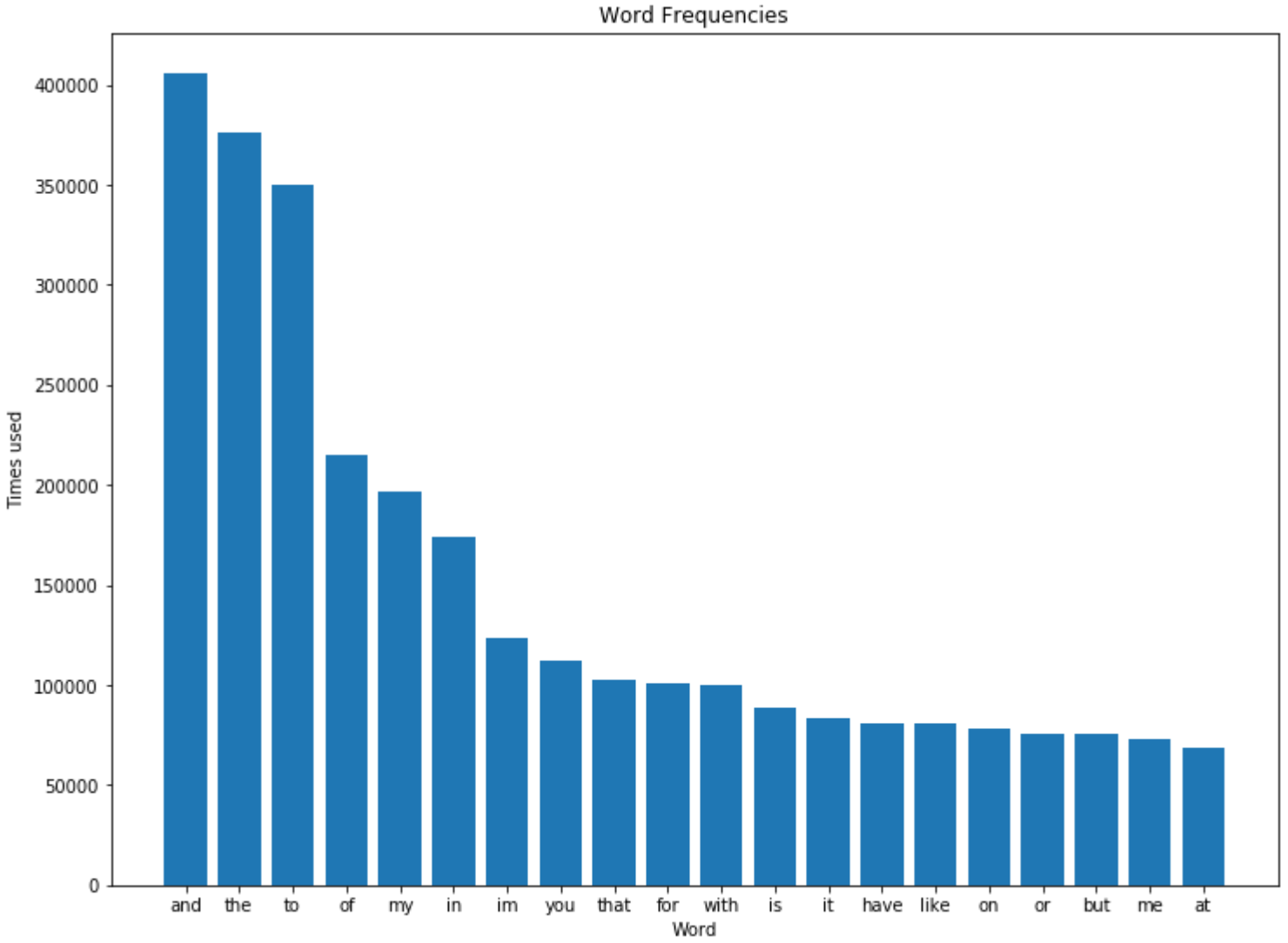


Dataset Exploration

- Exploration mostly consisted of looking at various essays and sentence structure.
- The essays themselves don't really lend themselves to graphing. There are many words that are used frequently.
- Still, we can look at the most frequently used words for all essays to see if there are differences.
- We can also look at the most frequently used words for “The 6 Things I could never do without” to see if there are differences.

Dataset Exploration

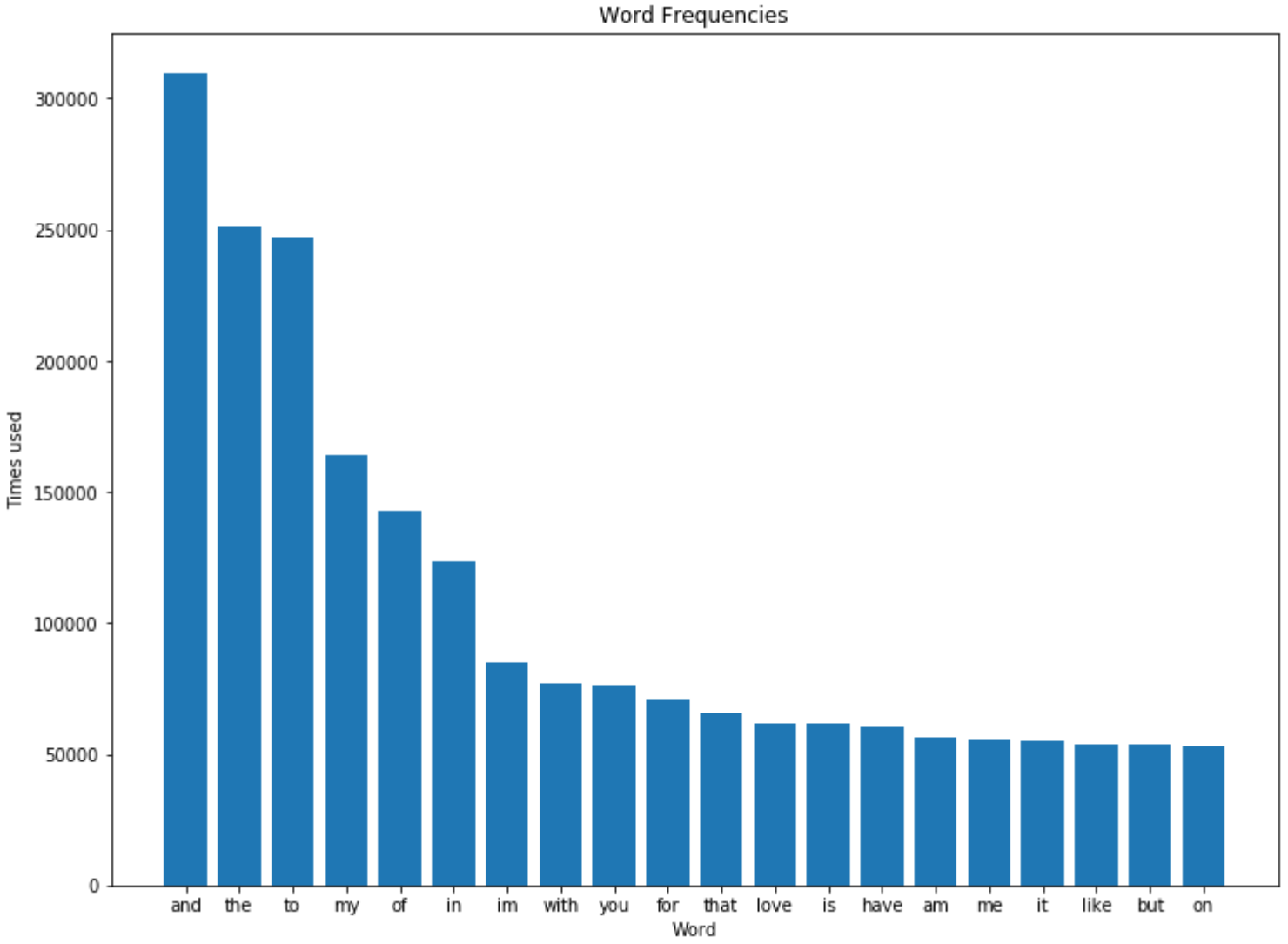
Male Essays - Top 20 Words




```
and:  405802
the:  376457
to:  349619
of:  214668
my:  196876
in:  173726
im:  123128
you:  111904
that:  102256
for:  100777
with:  99767
is:  88254
it:  83044
have:  81248
like:  80891
on:  77953
or:  75931
but:  75450
me:  72790
at:  68850
```

Dataset Exploration

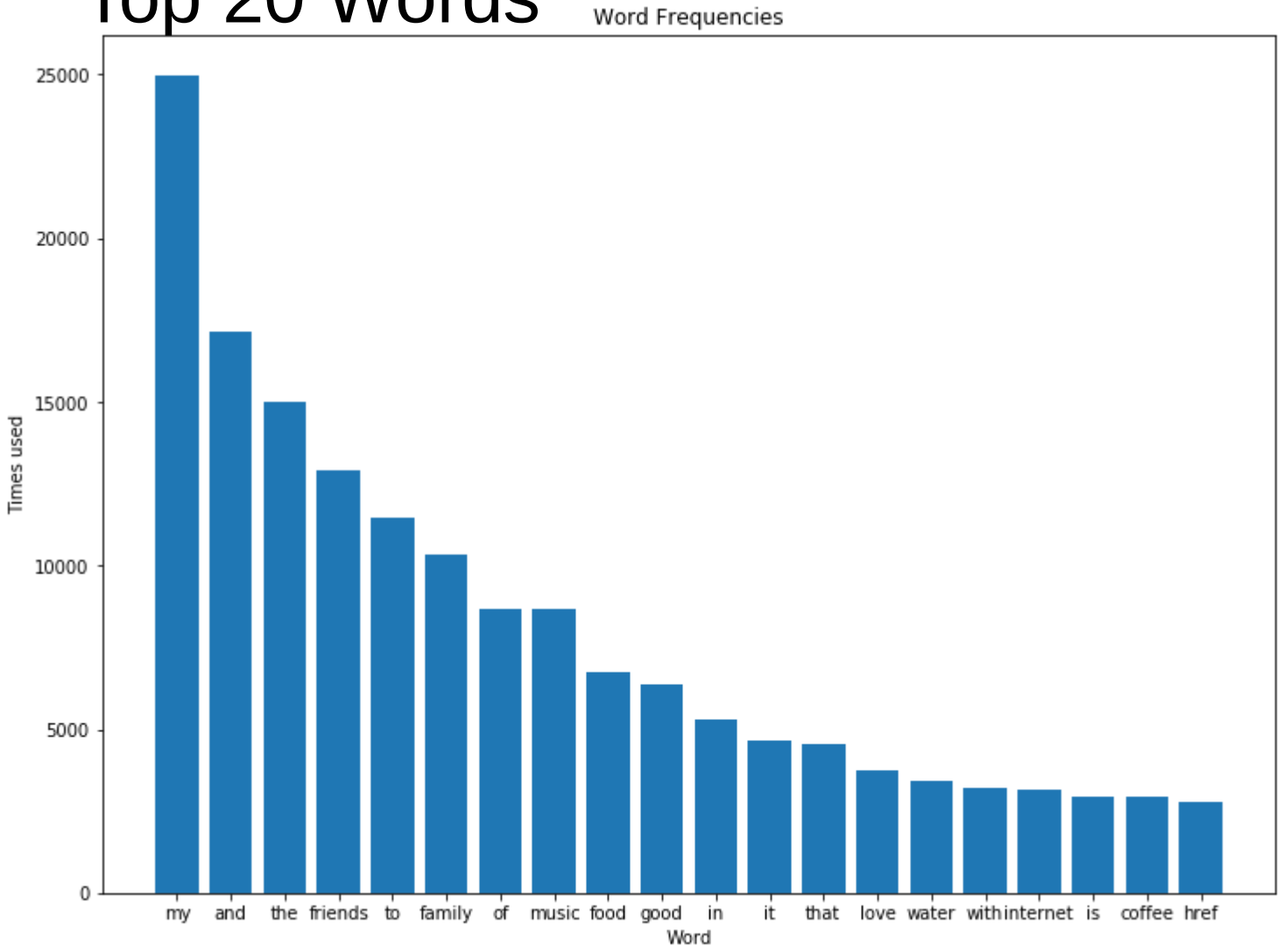
Female Essays - Top 20 Words




```
and:  309731
the:  250885
to:  246979
my:  164225
of:  142478
in:  123662
im:  84644
with:  77272
you:  76550
for:  70929
that:  65892
love:  61920
is:  61637
have:  60416
am:  56310
me:  55464
it:  55274
like:  54019
but:  53973
on:  52845
```

Dataset Exploration

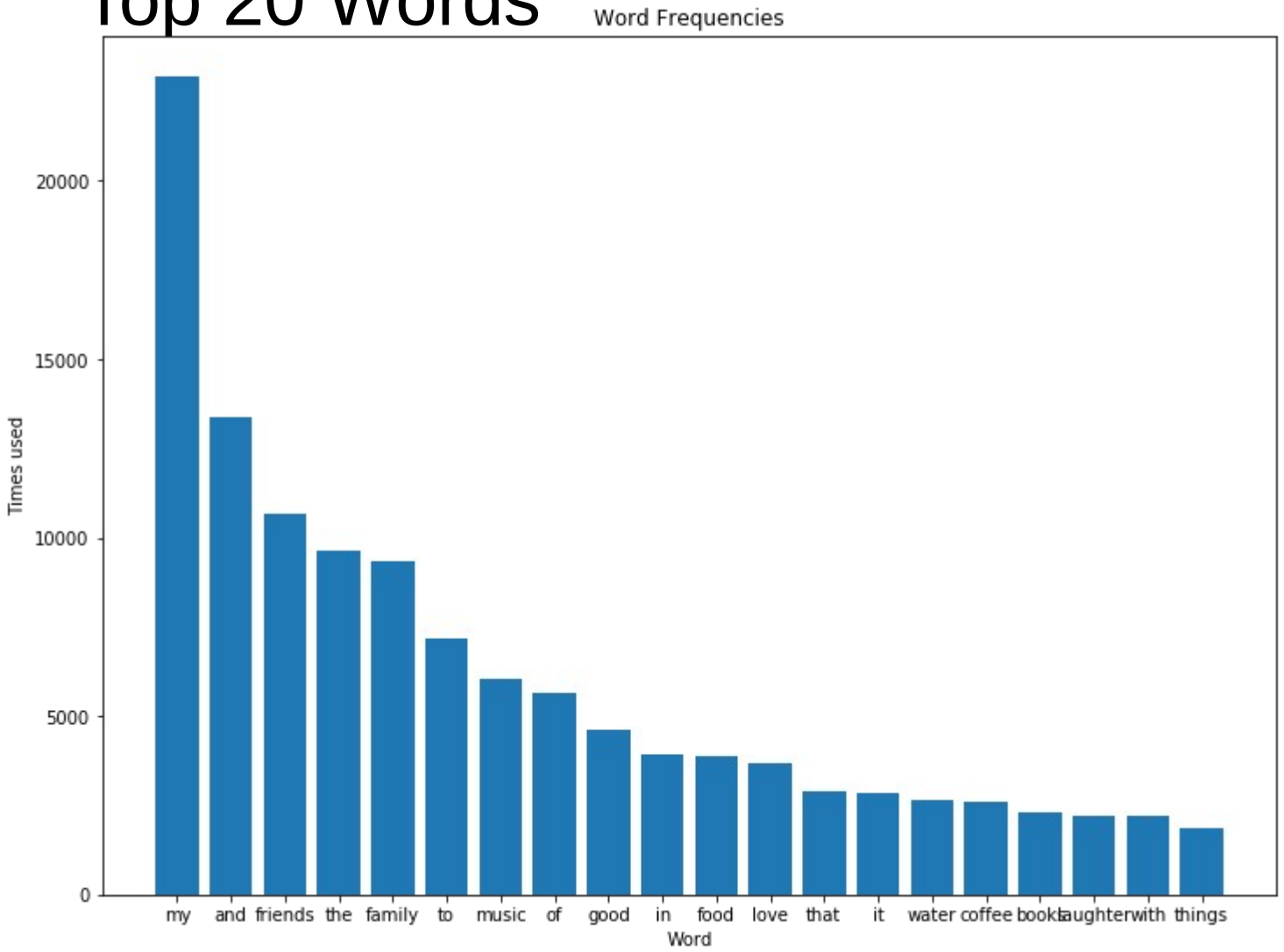
Male “6 things I could never do without”
Top 20 Words




```
my:  24979
and:  17170
the:  15020
friends:  12887
to:  11446
family:  10326
of:  8668
music:  8660
food:  6732
good:  6344
in:  5310
it:  4662
that:  4520
love:  3724
water:  3406
with:  3190
internet:  3160
is:  2959
coffee:  2950
href:  2779
```

Dataset Exploration

Female “6 things I could never do without”
Top 20 Words




```
my:  22927
and:  13374
friends:  10657
the:  9618
family:  9359
to:  7184
music:  6058
of:  5668
good:  4621
in:  3946
food:  3901
love:  3662
that:  2872
it:  2859
water:  2631
coffee:  2603
books:  2312
laughter:  2217
with:  2199
things:  1869
```

Additional Columns

- Concatenated/Filtered essay text
  - There was junk data in the essay text (html like
, "). Apostrophes, commas, and parenthesis seemed like they could also be problematic. I wrote functions to get rid of the junk, normalize the essays into lower case words with all special characters filtered out. Regular expressions were used for this.
- Sex as 0 or 1 instead of “m” or “f”
  - This is useful for using regression models.
- Count vectors based on the concatenated and filtered essay text.
  - from sklearn.feature_extraction.text import CountVectorizer

Classification Approaches

- Naive Bayes
  - Best performance and accuracy of all models tested.
- K-Nearest Neighbors
  - Not too slow, somewhat accurate. But it often classifies females as males.
- Support Vector Machines
  - Very slow, least accurate. Rarely classifies essays as female.

Naive Bayes Classifier

- Run time (for fitting & predicting the model) is under 1 second on my machine.
- Accuracy: 75.9%
- Precision: 75.7%
- Recall: 83.1%
- F1: 79.2%

Confusion Matrix

	Guessed M	Guessed F
User is M	5505	1765
User is F	1122	3598

K-Nearest Neighbors Classifier

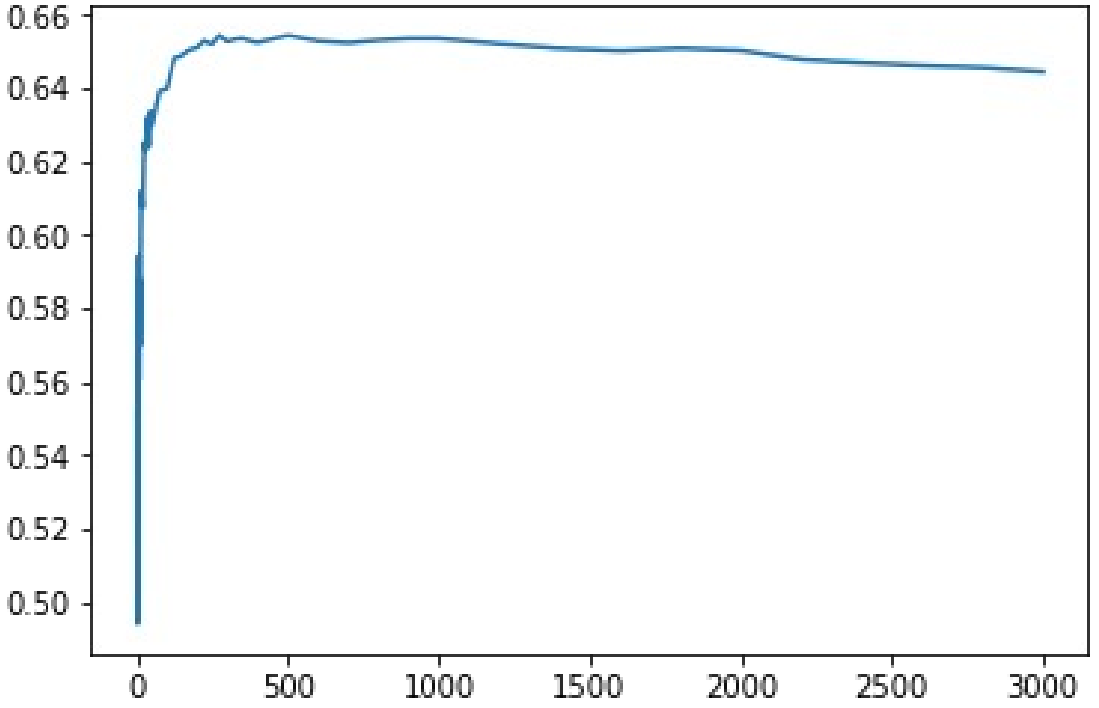
- Accuracy: 65.5%
- Precision: 82.7%
- Recall: 67.6%
- F1: 74.4%
- Guesses male 74% of the time even though males represent 60% of the users.

Confusion Matrix

	Guessed M	Guessed F
User is M	6012	1258
User is F	2880	1840

K-Nearest Neighbors Classifier

- Optimal num_neighbors is around 500.
- Run time for fitting the model and making predictions is 2 minutes 20 seconds on my machine.



Support Vector Machines Classifier

- Run time (for fitting & predicting the model) is 1 hour and 15 minutes on my machine.
- Accuracy: 62.0%
- Precision: 99.1%
- Recall: 61.6%
- F1: 75.9%

Confusion Matrix

	Guessed M	Guessed F
User is M	7205	65
User is F	4497	223

- Although females make up 40% of the users, only 2.4% of users were classified as female!

Learned Bias

- Support Vector Machines (and to a lesser extent K-Nearest Neighbors) seems to have learned males are more likely to be in the dataset than females.
- Of the female users in the test set, SVM guessed they were male 95.3% of the time!
- This could potentially be worked around by creating a training set with an equal number of males and females.

Regression Approaches

- K-Nearest Neighbors Regressor
  - Running time and accuracy similar to the K-Nearest Neighbors Classifier.
- Multiple Linear Regression
  - Very slow, Least accurate.

K-Nearest Neighbors Regressor

- Accuracy: 65.5%
- Precision: 82.3%
- Recall: 67.7%
- F1: 74.3%

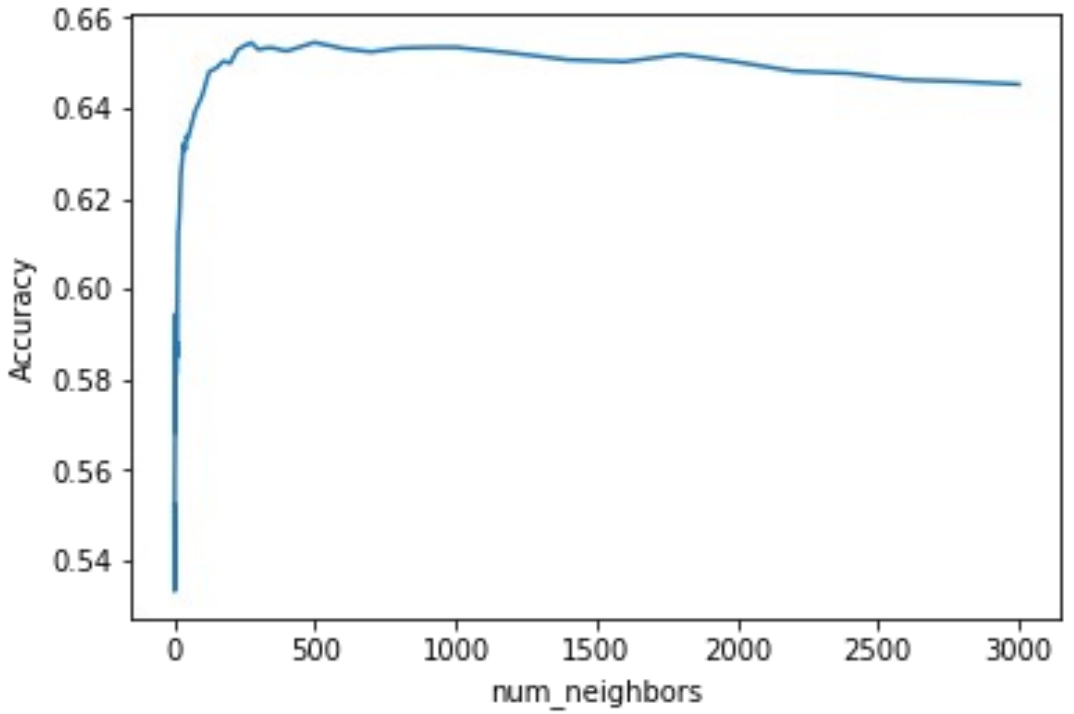
Confusion Matrix

	Guessed M	Guessed F
User is M	5981	1289
User is F	2853	1867

- Guesses male 74% of the time even though males represent 60% of the users.
- Similar results to K-Nearest Neighbors Classifier.

K-Nearest Neighbors Regressor

- Optimal num_neighbors is around 500. Again, similar to K-Nearest Neighbors Classifier.
- Run time for fitting the model and making predictions is just under 2 minutes on my machine.



Multiple Linear Regression

- Run time (for fitting & predicting the model) is 2 hours 20 minutes on my machine.
- The model is multi-threaded. Unlike all other models tested, this utilized my 8 CPU cores.
- Accuracy: 52.7%
- Precision: 53.8%
- Recall: 62.9%
- F1: 58.0%

Confusion Matrix

	Guessed M	Guessed F
User is M	3912	3358
User is F	2312	2408

Conclusions to Questions

- Q: Can the sex of the users be predicted by analyzing their essays?
- A: Yes. The Naive Bayes classifier did the best overall (F1 score of 79.2%).
- All five machine learning models that were tested were more accurate than using a coin toss to predict.

Conclusions to Questions

- Q: What machine learning models are most effective? (accurate for the task of predicting sex from essay text)
- A: Of the models tried, the Naive Bayes classifier worked the best (accurate around . The K-Nearest Neighbors models were accurate about 65% of the time.

Conclusions to Questions

- Q: Although this is a classification task, how well can the regression models do?
- A: K-Nearest Neighbors Regressor did as well as the K-Nearest Neighbors Classifier. Multiple Linear Regression did poorly (barely above 50% accuracy, very slow to fit).

Next Steps

- There are many possible ways to further improve the recognition.
  - Identifying phrases rather than relying only on word counts.
  - Discarding the most frequently used words (e.g. “and”, “the”, “to”).
  - Performing analysis on each essay column individually, rather than the concatenation.
  - Using multiple models in conjunction with each other.