Quora

DUPLICATE QUESTION PAIRS

Group # 16
Sai Krishna Kanneti (#31)
Tejas Dhrangadharia (#13)
Karan Nisar (#54)

University at Buffalo The State University of New York

INTRODUCTION

- What is Quora?
  - World's biggest forum with over 100 million users visits every month.
  - Best forum to share and gain knowledge
- Problem
  - People may ask similar questions
  - Important interest is to detect duplicated questions
  - Duplicate questions with same meaning:
    - **Should I tell my girlfriend about my past relationship?**
    - **How do I tell my past relationship to my girl?**
- Prediction Problem
  - From a question pair, predict whether question are the same or not.

OBJECTIVE

- Identify potential duplicate questions
- Tackle this natural language processing problem by applying advanced techniques to classify whether question pairs are duplicates or not.
- Build models of semantic equivalence, based on actual Quora data.

DATA OVERVIEW

id	qid1	qid2	question1	question2	is_duplicate
447	895	896	What are natural numbers?	What is a least natural number?	0
1518	3037	3038	Which pizzas are the most popularly ordered pizzas on Domino's menu?	How many calories does a Dominos pizza have?	0
3272	6542	6543	How do you start a bakery?	How can one start a bakery business?	1
3362	6722	6723	Should I learn python or Java first?	If I had to choose between learning Java and Python, what should I choose to learn first?	1

APPROACH

Text Mining

- Data cleaning using text mining
- Sentimental analysis

Models

- Logistic Regression
- Random Forest (RF)
- XG Boost

Comparison of Models

University at Buffalo The State University of New York

TEXT MINING

TEXT MINING

- Remove punctuations, white spaces, numbers and stop words
- Sentimental analysis
  - Positive words
  - Negative words
- Compare matching words in both questions and obtain match count

english Stop Words

```
 [1] "i"      "me"     "my"     "myself"
 [9] "you"    "your"   "yours"  "yourself"
[17] "himself" "she"    "her"    "hers"
[25] "they"   "them"   "their"  "theirs"
```

SMART Stop Words

```
 [1] "a"       "a's"      "able"     "about"
 [8] "across"  "actually" "after"   "afterwards"
[15] "all"     "allow"    "allows"   "almost"
[22] "also"    "although" "always"   "am"
```

Result after Text Mining:

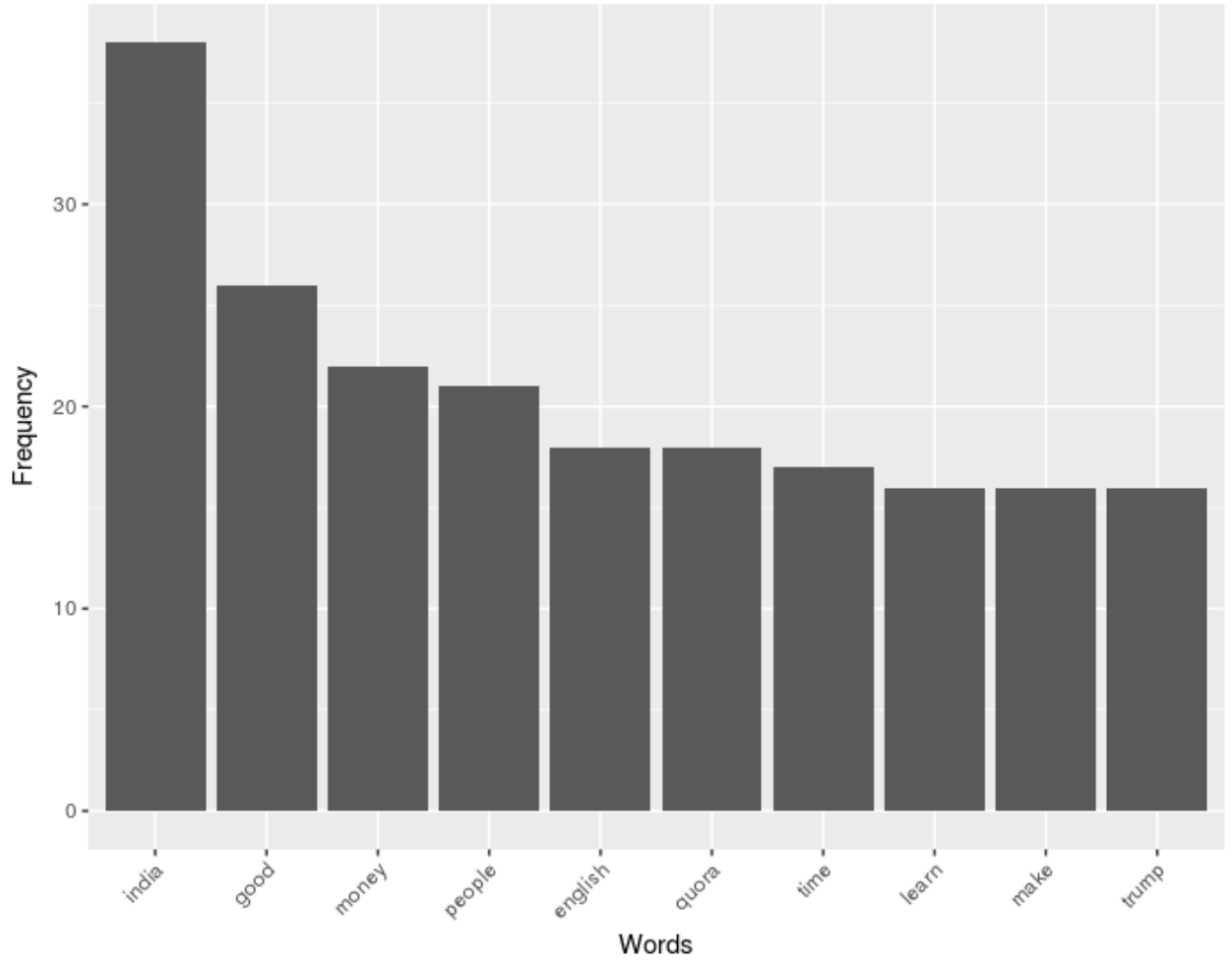
Question id	is_duplicate	match_count	p1	p2	n1	n2
7210	1	1	0	0	0	0
8757	1	0.4705882	1	1	3	1
7609	0	0	1	0	0	0
8859	1	0.6666667	1	2	0	0
4563	0	0	1	0	0	0
1663	1	0.75	1	1	0	0

Column Description

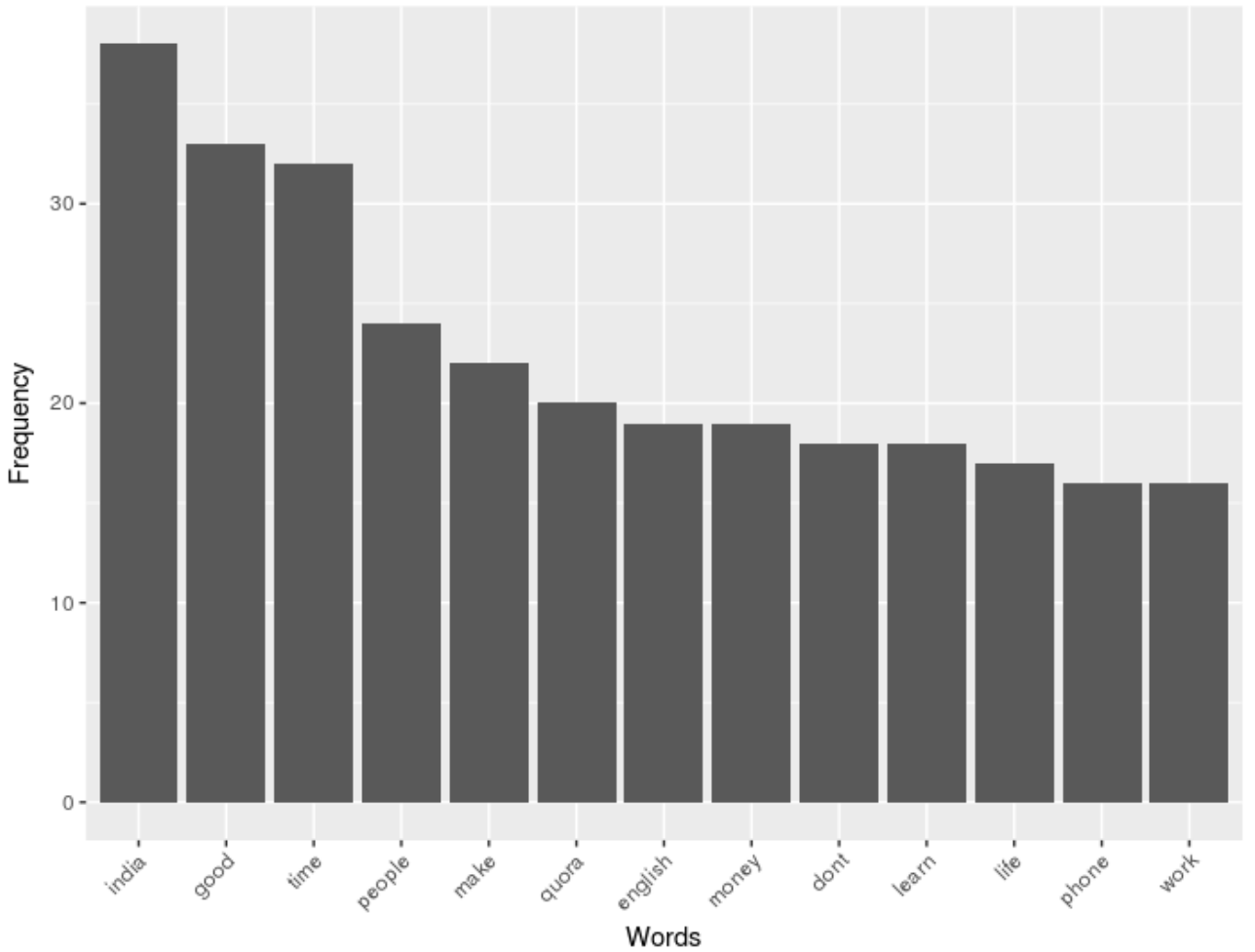
Columns	Description
Question id	Id of question
is_duplicate	Same as data set
match_count	Frequency of word matches in two questions
p1	Number of positive words in question 1
p2	Number of positive words in question 2
n1	Number of negative words in question 1
n2	Number of negative words in question 2

FREQUENCY PLOTS OF WORD COUNT

Question 1

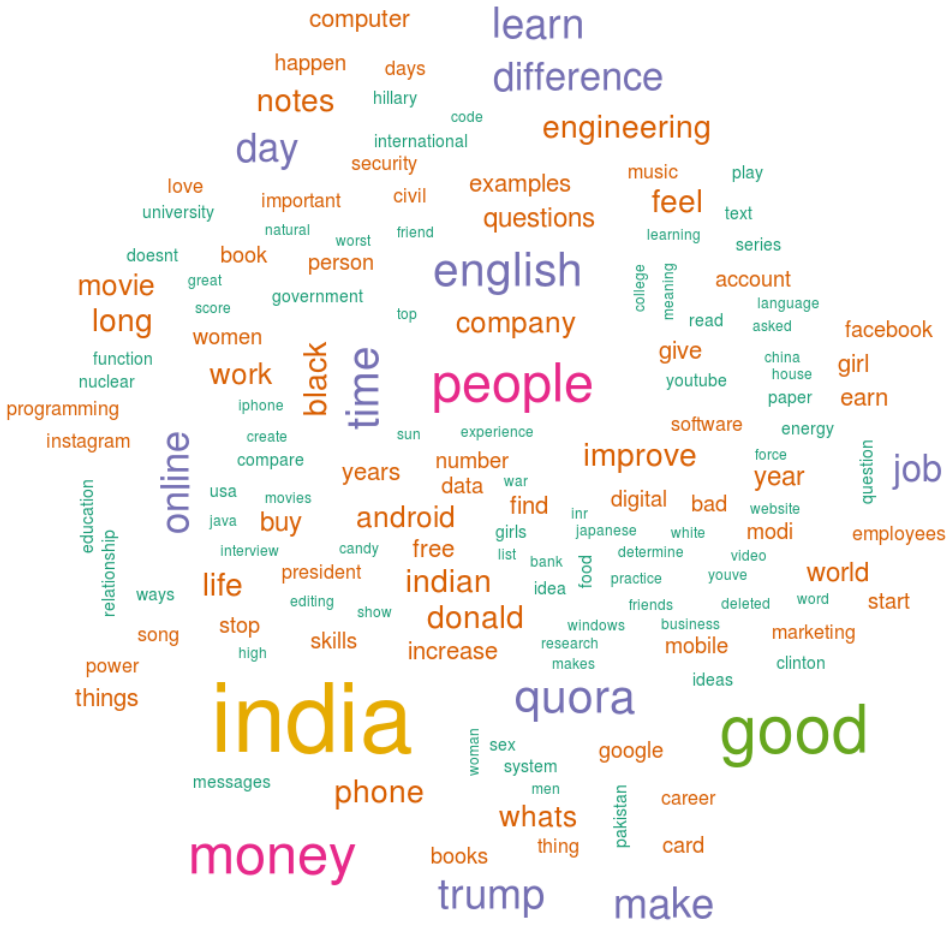


Question 2

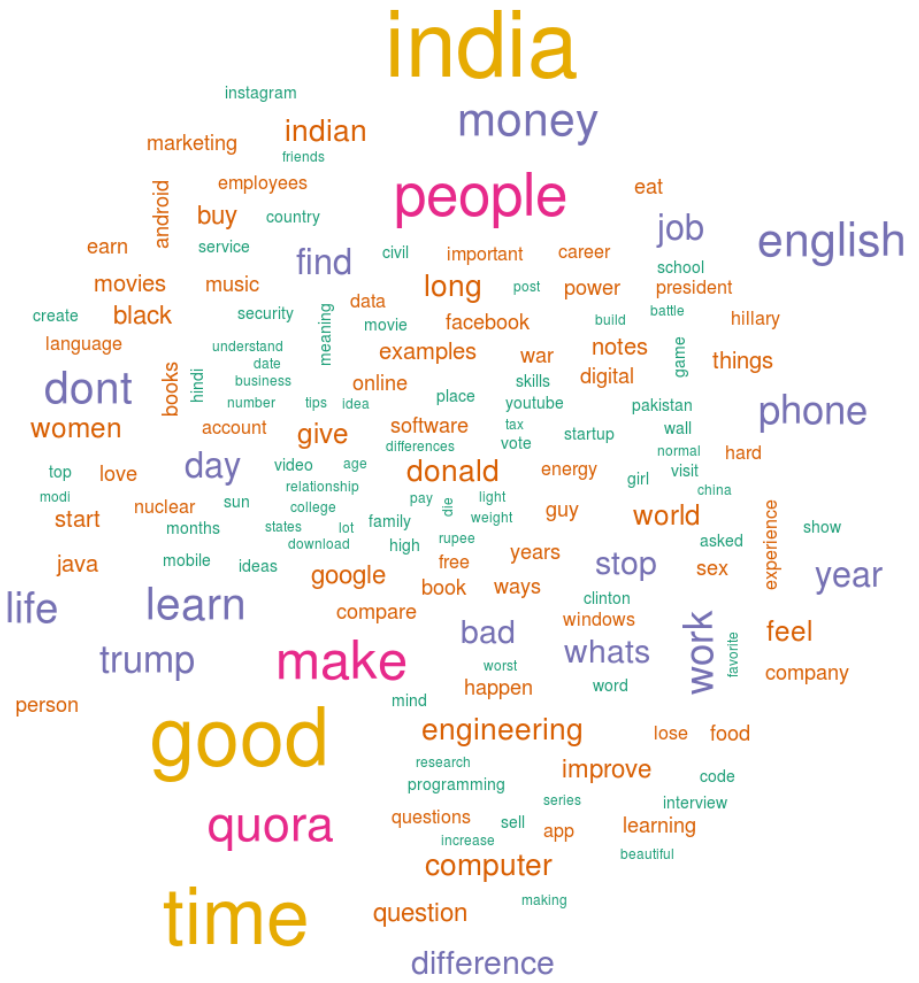


WORD CLOUD

Question 1



Question 2



APPROACH 1: LOGISTIC REGRESSION

LOGISTIC REGRESSION

```
> summary(predict_log)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.9629  1.2482  1.3966  1.3759  1.5334  1.7411
> ?logLoss
> Metrics::logLoss(as.numeric(actual),predict_log)
[1] Inf
Warning message:
In log(1 - predicted) : NaNs produced


> table(test$is_duplicate)

   0    1
1876 1124
> table(predict_log)
predict_log
   0    1
   1 2999
>
```

APPROACH 2: RANDOM FOREST

RANDOM FOREST

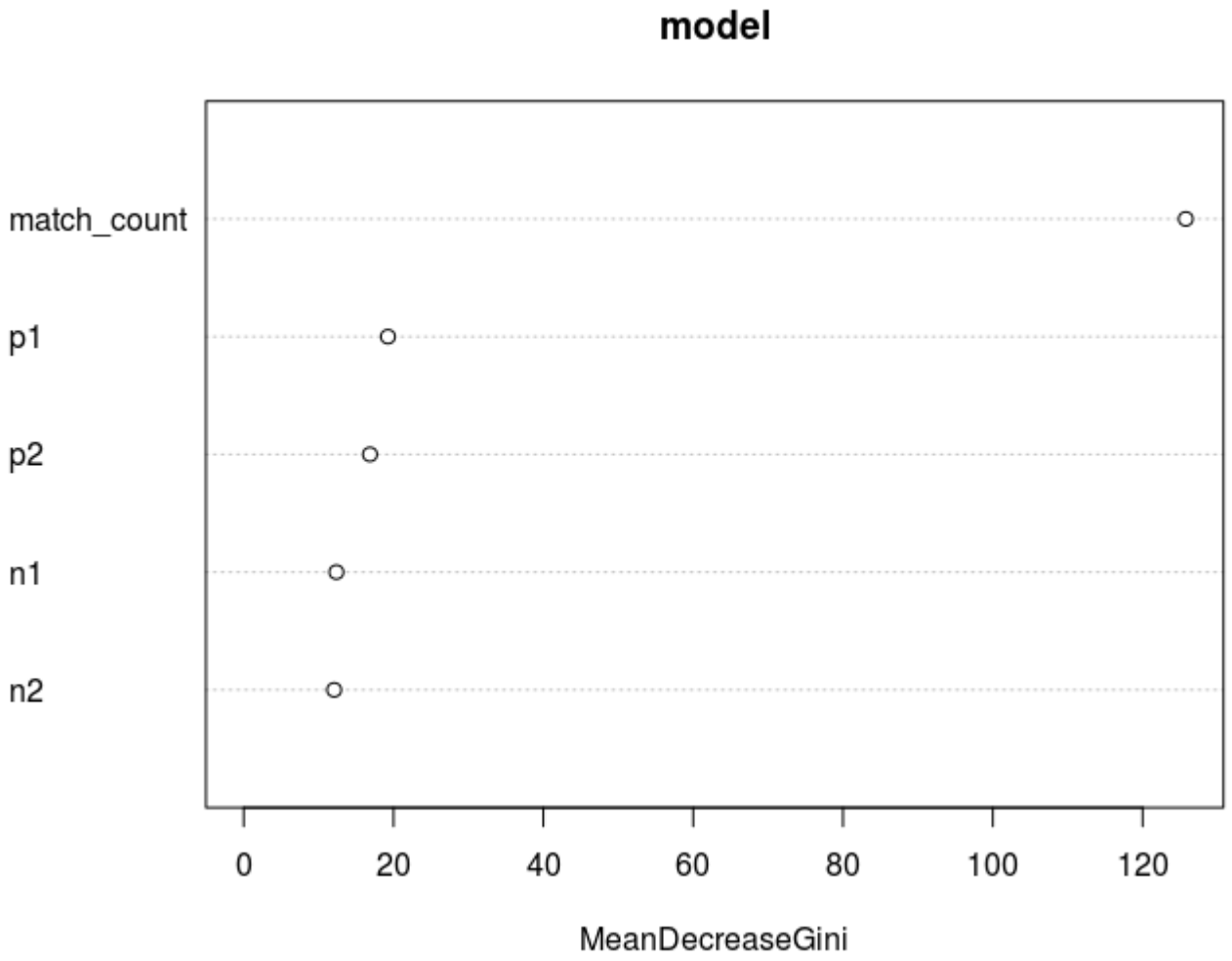
- Random Forests improve variance by reducing correlation between trees, this is accomplished by random selection of feature-subset for split at each node.

```
Call:
 randomForest(formula = is_duplicate ~ ., data = tr, n.tree = 1000)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 32%
Confusion matrix:
    0   1 class.error
0 444 158   0.2624585
1 162 236   0.4070352
```

- Misclassification Rate for train : 22.12%
- Misclassification Rate for test : 20.6%

SIGNIGICANCE OF PREDICTOR VARIABLES



APPROACH 3: XG BOOST

PARAMETER TUNING FOR TREE BOOSTER

- **eta** (Learning rate):
  - Step size shrinkage used to prevent overfitting.
  - 0.1
- **max_depth**
  - Maximum depth of a tree, increasing this value will make the model more complex / likely to be overfitting.
  - 5
- **nrounds**
  - The number of iterations.
  - 100

TUNING LEARNING TASK PARAMETERS

- These parameters are used to define the optimization objective the metric to be calculated at each step.
- **Objective:**
  - binary:logistic – logistic regression for binary classification, returns predicted probability
- **eval_metric :**
  - The metric to be used for validation data.
  - logloss – negative log-likelihood

XG BOOST

```
  cb.print.evaluation(period = print_every_n)
  cb.evaluation.log()
  cb.save.model(save_period = save_period, save_name = save_name)
niter: 100
evaluation_log:
    iter train_logloss
       1      0.668746
       2      0.648710
---
      99      0.512586
     100      0.512419
```

COMPARISON

- Metric: Log Loss

$$-\frac{1}{N}\sum_{n=1}^{N}\left[y_n \log \hat{y}_n + (1-y_n)\log(1-\hat{y}_n)\right]$$

Model	Log Loss
Random Forest	0.5151
XGBoost	0.5124

CHALLENGES

- **Scalability**:

  We resampled data set to 10,000 rows.

- **Class Imbalance**:

  Higher priority to classes that are important.

  is_duplicate=0

University at Buffalo The State University of New York

THANK YOU!!!