This is Major Tom to Ground Control :
"I'm stepping through the door
And I'm floating in a most peculiar way
And the stars look very different today "

Sub-Reddits Classification Analysis

DSI-18 Project 3

Robby Sim

7 Dec 2020

Photo by Andrew Xu, taken from https://digital-photography-school.com/lake-tekapo-stars/

Excel & VBA Galaxy

Planet Excel

Excel Knowledge Specialist

Planet VBA

VBA Knowledge Specialist

Excel Queries

VBA Queries

Potential Visitors

Excel & VBA Galaxy

Planet Excel

Planet VBA

Excel Queries

?Actual VBA Queries?

VBA Queries

?Actual Excel Queries?

Potential Visitors

Excel & VBA Galaxy

Planet Excel

Planet VBA

Excel Queries

VBA Queries

?Actual VBA Queries?

?Actual Excel Queries?

Problem Statement: Minimize the mis-classification of VBA queries (True) as Excel Queries (Predicted)

1,000 Excel queries received daily

20% mis-classified

200 mis-classified queries (VBA)

10 mins to correct

Loss time: 2,000 mins / day

11,000 hours annually

$50 / man-hour

$500,000 annually!

$500,000 resources wasted based on assumed 1,000 Excel queries received and 20% mis-classification rate

Scrape r/excel r/vba

Dropped duplicates (>1000 dups)

Dropped mis-classified posts (7%)

Basic analysis of postings

```
excel_df.shape
```

```
(1964, 105)
```

```
vba_df.shape
```

```
(1993, 105)
```

```
excel_df.drop_duplicates(inplace=True)
excel_df.reset_index(drop=True, inplace=True)
```

```
vba_df.drop_duplicates(inplace=True)
vba_df.reset_index(drop=True, inplace=True)
```

```
excel_df.drop_duplicates(subset = ['title', 'selftext', 'author_fullname'], keep='first', inplace=True)
```

```
vba_df.drop_duplicates(subset = ['title', 'selftext', 'author_fullname'], keep='first', inplace=True)
```

```
print(excel_df.shape)
print(vba_df.shape)
```

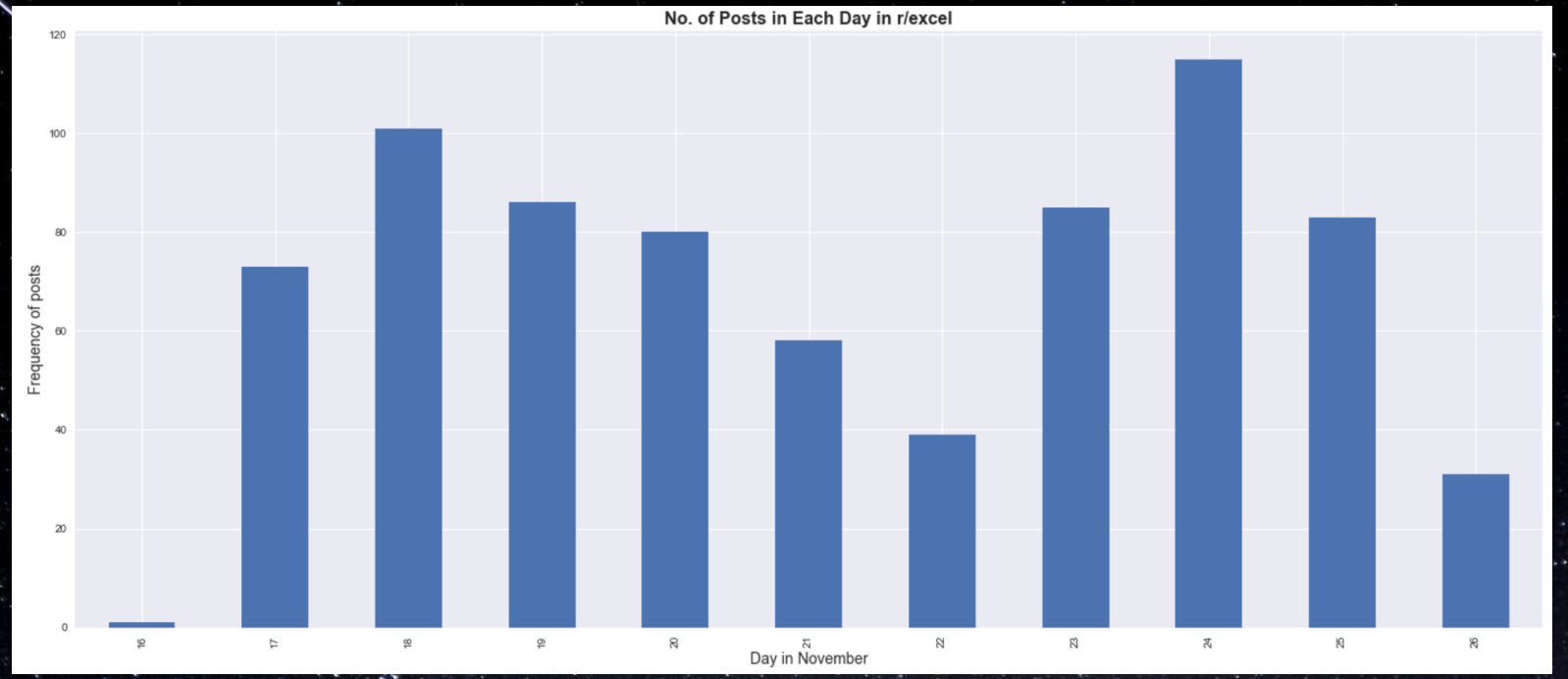
```
(807, 105)
(935, 105)
```

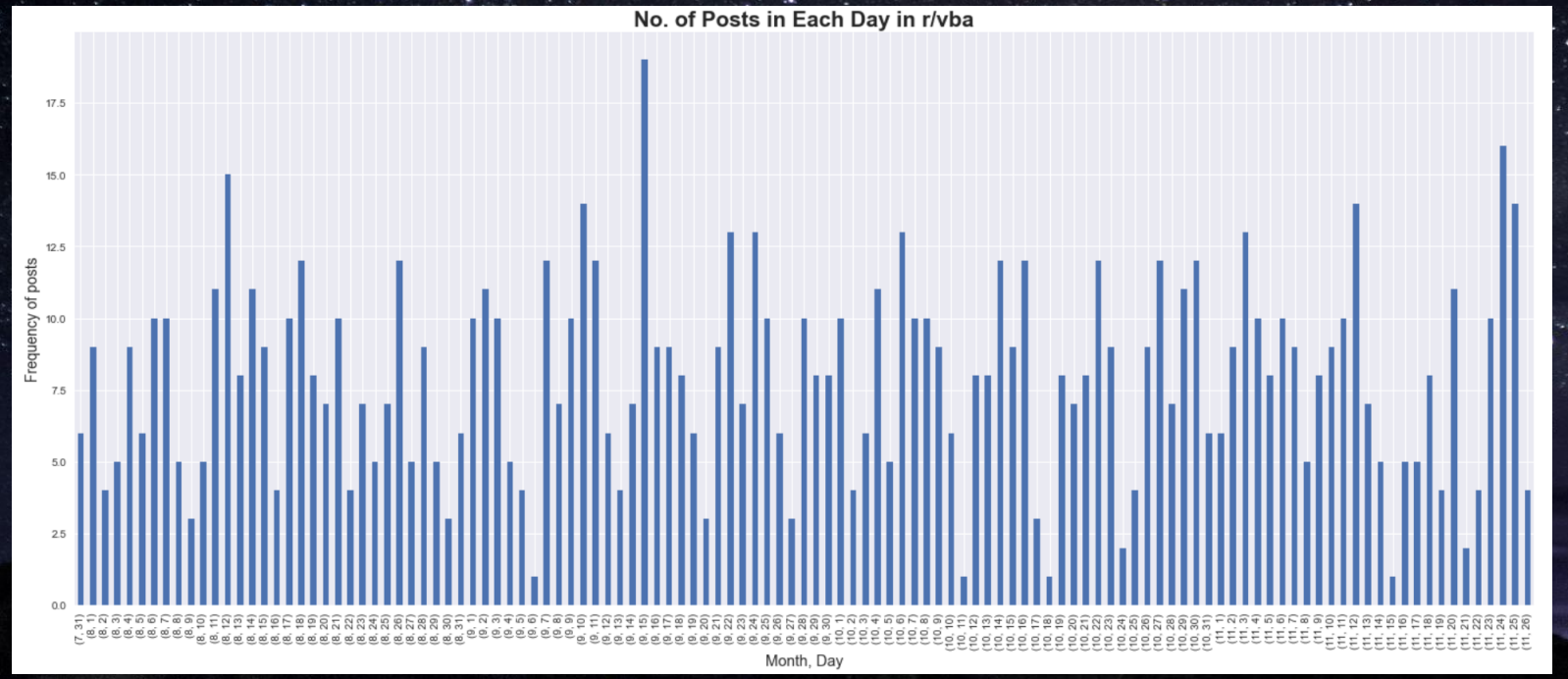
```
excel_df.loc[excel_df['title'].str.contains('vba')]
```

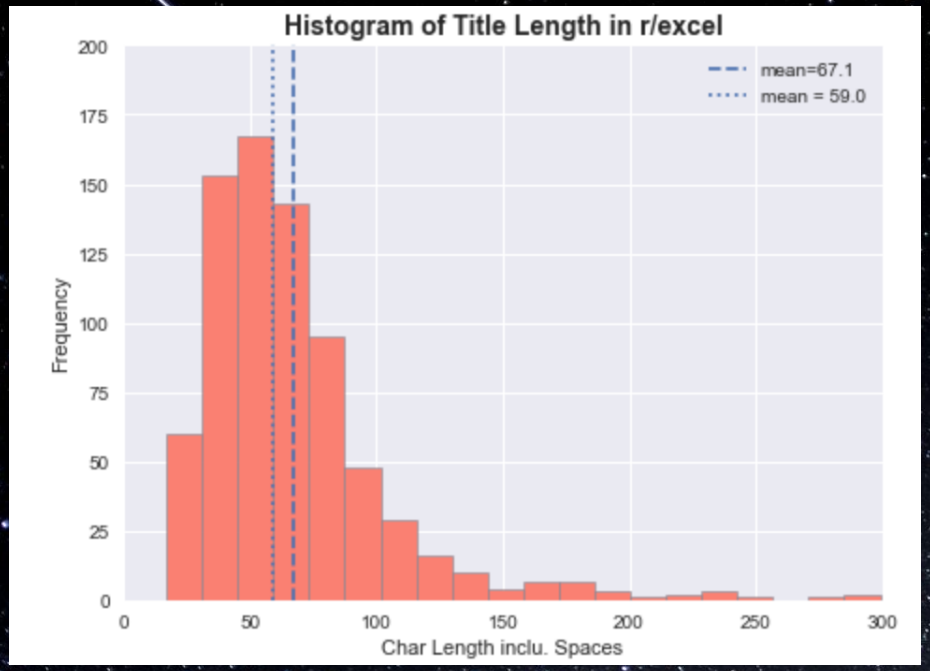
```
excel_df.loc[excel_df['selftext'].str.contains('vba')]
```

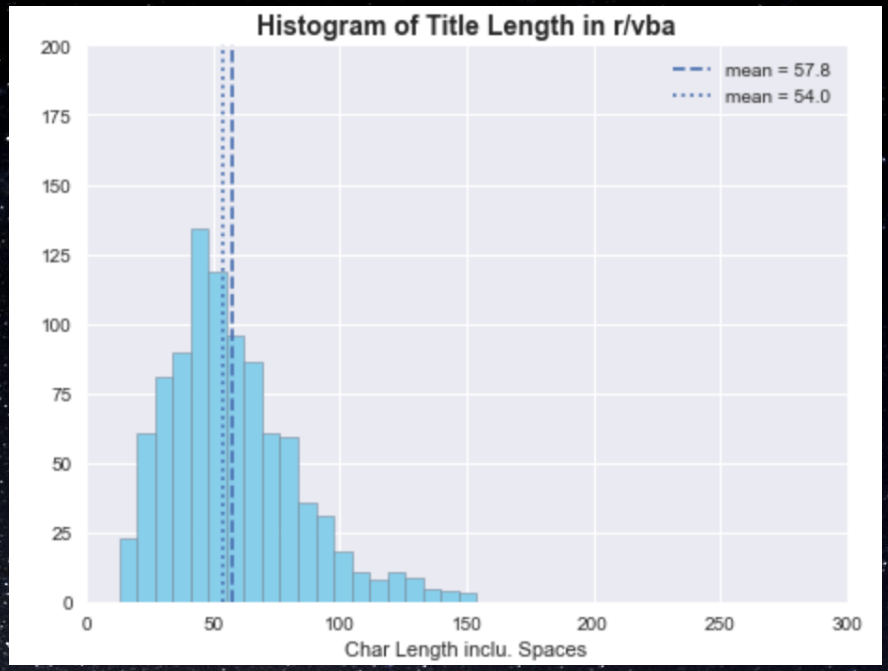
```
print(excel_df.shape)
print(vba_df.shape)
```

```
(752, 105)
(935, 105)
```


No. of Posts in Each Day in r/excel
Frequency of posts
Day in November
0
20
40
60
80
100
120
16
17
18
19
20
21
22
23
24
25
26


No. of Posts in Each Day in r/vba
Frequency of posts
Month, Day
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5





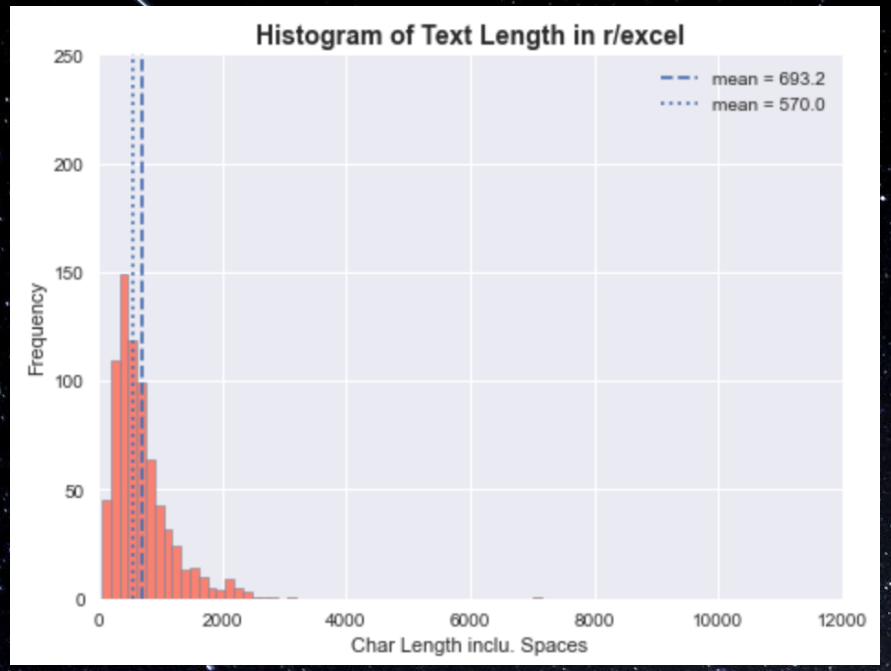
The over-enthusiastic title post (r/excel)

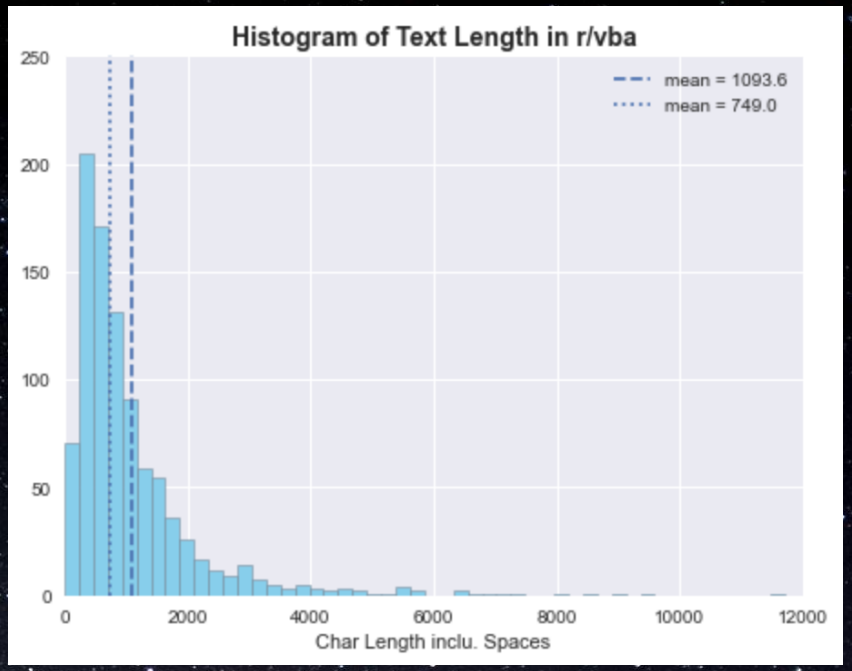
```
excel_df['title'][142]
```

```
]: 'three-line address is formatted fine in cell but appears all in one row... unless i double click on the cell, in w
hich case it goes back to normal. is there a way to "double click" a whole bunch of cells at once or avoid the exer
cise altogether?'
```

```
excel_df['selftext'][142]
```

```
]: "I have addresses in the following format:\n\n123 Fake Street  \nSuite 106  \nSpringfield, MA \n\nFor some reason,
when I paste the column with the addresses over to a new sheet, they appear like so:\n\n123 Fake StreetSuite 106Spr
ingfield, MA \n\nUnless I double click into the cell and then leave it, in which case it goes back to normal.\n\nAn
y ideas on how to fix it so I don't need to do the double clicking?"
```





The detailed explainer r/vba

'I started a discussion a few weeks ago regarding passing an array from a function to my main sub to minimize how many times I\'m re-writing the same section of code every time I need to access this data. See below link for background: https://www.reddit.com/r/vba/comments/is244n/passing_variant_array_from_one_sub_to_another/\n\nSkip to the two bold sentences for the actual piece of discussion I\'m looking for ideas/thoughts on.\n\nI ended up throwing the array into a dictionary to pass back to my main sub since that was the end goal anyways. It works great, no complaints at all, but I like to optimize and learn as much as I can and had a thought that I\'m trying to think through currently.\n\nI have built multiple "databases" which is just columns of data that I can reference. Each contains varying columns and data I need at different times. Two databases have 12 columns, one database has 4 columns, and I\'m sure down the road I\'ll be adding more with varying column #s.\n\nIn order to make life easier I have global public variables for the column # reference, so when I add new columns to the databases all I need to do is update the column number in one place instead of going through every code that touches it.\n\nAnyway, I currently have set up multiple select cases based on the data I need, and add all the data into a dictionary. As you\'ll see in the below code, I reference each column by it\'s public variable. **I\'m thinking it would be easier if I find the last column (not referencing the public variable), make that the array, and then somehow loop through from column 1 through last column adding to the dictionary items (again, instead of referencing the public variable)**.\n\nAn additional thought is, in my function I have an optional dictionaryLookupColumn which I was going to use later on with an additional select case if I needed to create a different key for the dictionary. **So my question would be, is there a line or two of code I can do to avoid having to make a separate select case for these one off dictionaries? I would assume this would work with the above thought on looping through the columns, and if I want column 3 to be the dictionary key, I would have to add columns 1 through end of range skipping column 3. But how could I do that?**\n\nI know that to utilize the dictionaryLookupColumn I am going to have to change my select case to something like; select case True and then case sheetName = xxxx and dictionaryLookupColumn = xx\n\nEDIT: Actually, if I looped through adding all columns to the items, I would want to keep every column (not skip over column 3) so that I can reference the correct column using the public variable when I split in the main code. So I guess is there an easy way to avoid doing a loop within a loop? \n\n\n\n   Function getDatafromDatabase(filewithPath As String, sheetName As String, Optional dictionaryLookupColumn As Byte) As Dictionary\n   \n   \'\'=========================\n   \'\' Variable Declarations\n   \'\'=========================\n   \n   Dim currentDatabase As Workbook\n   Dim currentSheet As Worksheet\n   \n   Dim currentArr As Variant\n   Dim currentDictionary As Dictionary\n   \n   Dim i As Long\n   \n   \'\'==============\n   \'\' Code Start\n   \'\'==============\n   \n   Set currentDatabase = Workbooks.Open(filewithPath)\n   Set currentSheet = currentDatabase.Sheets(sheetName)\n   Set currentDictionary = New Dictionary\n   \n   With currentSheet\n   \

```
n      \'If autofilter is currently on show all data\n          If .AutoFilterMode = False Then\n          ElseIf .AutoFilter.FilterMode Then\n              .ShowAllData\n          Else\n          End If\n      \n\'Find which pre-determined sheet we\'re looking at, add entire range to current array variable, then loop through array adding unique values to current dictionary\n          Select Case currentSheet.Name\n          Case Is = "Master Clients"\n              \n              currentArr = .Range(.Cells(2, 1), .Cells(.Cells(Rows.Count, 1).End(xlUp).Row, clients_dnrColumn)).Value\n              \n              For i = LBound(currentArr) To UBound(currentArr)\n                  If currentDictionary.Exists(currentArr(i, clients_endpointnumberColumn)) Then\n                  Else\n                      currentDictionary.Add _\n                      currentArr(i, clients_endpointnumberColumn), _\n                      currentArr(i, clients_campaignColumn) &amp; "," &amp; _\n                      currentArr(i, clients_endpointnameColumn) &amp; "," &amp; _\n                      currentArr(i, clients_clientnameColumn) &amp; "," &amp; _\n                      currentArr(i, clients_typeColumn) &amp; "," &amp; _\n                      currentArr(i, clients_payoutColumn) &amp; "," &amp; _\n                      currentArr(i, clients_talktimeColumn) &amp; "," &amp; _\n                      currentArr(i, clients_dedupeColumn) &amp; "," &amp; _\n                      currentArr(i, clients_sharerecordingColumn) &amp; "," &amp; _\n                      currentArr(i, clients_statusColumn) &amp; "," &amp; _\n                      currentArr(i, clients_reportlocationColumn) &amp; "," &amp; _\n                      currentArr(i, clients_hoursColumn) &amp; "," &amp; _\n                      currentArr(i, clients_dnrColumn)\n                  End If\n              Next i\n              \n          Case Is = "Master Emails"\n              \n              currentArr = .Range(.Cells(2, 1), .Cells(.Cells(Rows.Count, 1).End(xlUp).Row, clientEmails_headerColumn)).Value\n              \n              For i = LBound(currentArr) To UBound(currentArr)\n                  If currentDictionary.Exists(currentArr(i, clientEmails_clientnameColumn)) Then\n                  Else\n                      currentDictionary.Add _\n                      currentArr(i, clientEmails_clientnameColumn), _\n                      currentArr(i, clientEmails_clientnicknameColumn) &amp; "," &amp; _\n                      currentArr(i, clientEmails_toColumn) &amp; "," &amp; _\n                      currentArr(i, clientEmails_ccColumn) &amp; "," &amp; _\n                      currentArr(i, clientEmails_headerColumn)\n                  End If\n              Next i\n      \n          Case Is = "Master TFN"\n              \n              currentArr = .Range(.Cells(2, 1), .Cells(.Cells(Rows.Count, 1).End(xlUp).Row, tfns_notes2Column)).Value\n              \n              For i = LBound(currentArr) To UBound(currentArr)\n                  If currentDictionary.Exists(currentArr(i, tfns_trackingnumberColumn)) Then\n                  Else\n                      currentDictionary.Add _\n                      currentArr(i, tfns_trackingnumberColumn), _\n                      currentArr(i, tfns_sourcenameColumn) &amp; "," &amp; _\n                      currentArr(i, tfns_sourceColumn) &amp; "," &amp; _\n
```

“Cleaning Agents”

Combined title & text

Cleaned with regex, stopwords & lemmatizer

Checked cleaned text & Cvec

Tweak cleaning agents & check again…

```
# 1. convert text to lower case
lower_text = raw_text.lower()

# 2. Remove HTML type chars
review_text = BeautifulSoup(lower_text).get_text()

# 3. Remove http(s) type strings
review_text = re.sub("http.\S+", " ", review_text)

# 4. Remove pointer reference and zero-width spaces
review_text = re.sub("[&]\S+", " ", review_text)

# 5. removing \xa0 reference
review_text = re.sub("\\xa0", " ", review_text)

# 6. removing reference to excel and vba
review_text = re.sub("(vba)|(excel)", " ", review_text)

# 7. removing new line reference
review_text = re.sub("\\n", " ", review_text)

# 8. removing long words reference (more than 15 char)
review_text = re.sub("[\S]{15}\S+", " ", review_text)

# 9. removing non-char non-num reference and all numbers
review_text = re.sub("([^a-z])|([\d])", " ", review_text)

# 10. removing standalone characters and multiple spaces
review_text = re.sub("(^| ).(( ).)*( |$)", " ", review_text)

# 11. Split into individual words.
words = review_text.split()
```

```
# 12. Using stopwords and user-defined words that are common across the 2 reddits
stops = stopwords.words('english')
new_words = ['column', 'columns', 'rows', 'row', 'cell', 'cells',
             'data', 'formula', 'worksheets', 'worksheet', 'values', 'value',
             'sheet', 'sheets', 'table', 'tables', 'macro', 'work',
             'hi', 'hello', 'file', 'files', 'number', 'numbers',
             'texts', 'text', 'using', 'would', 'like', 'function',
             'functions']
stops.extend(new_words)
meaningful_words = [w for w in words if w not in stops]

# 13. Lemmatizing
meaningful_words = [lemmatizer.lemmatize(w) for w in meaningful_words]

# 14. Join the words back into one string separated by space
return(" ".join(meaningful_words))
```

Original

'extracting data from one long string - in various formats Need advise on whether I\'m going down the correct methodology to achieve the following...I\'m trying to split out the product info into seperate columns. It becomes difficult when there\'s more than one item as well as the different format if the product is a cake.\n\nAt present, HTML emails are received. I then use Power Automate to add row to excel which is solely the first two columns of data - Order Number and Products\n\nFor Products List (column C), I remove the return spaces and replace it with || to make it easier to work with\n\n   =SUBSTITUTE(SUBSTITUTE(B2,CHAR(13),""),CHAR(10),"|")\n\nFor PLU / Items / Quantity / Price columns, I search for the || to return the relavent information : for example Items formula :\n\n   =TRIM(MID(SUBSTITUTE($C2,"||",REPT(" ",LEN($C2))),(1-1)*LEN($C2)+1,LEN($C2)))\n\n[The above process works if there\'s only one product listed.](https://preview.redd.it/gqp8emllrk161.png?width=1510&format=png&auto=webp&s=ade6efd14243abd3885c4ac7edac204ebb2c9130)\n\n[It get\'s more complicated when there\'s more than one item purchased or if a cake is purchased \\(additional lines\\)](https://preview.redd.it/p9dpg8nxtk161.png?width=1514&format=png&auto=webp&s=1a9a74f72885c8e493618bf2b53e72daf960395e)\n\n[Desired output for multiple products :](https://preview.redd.it/sld0spdtuk161.png?width=1229&format=png&auto=webp&s=1239f96bb5233e0cb2af5f1db762eced97eb9051)\n\n[Desired output for cake products :](https://preview.redd.it/eaf6kfz7vk161.png?width=1183&format=png&auto=webp&s=ea54c86d321b584fa6d71d5353ff93875e14a599)\n\n\n\nSample file : https://we.tl/t-2y6j02I9Az'

'extracting one long string various format need advise whether going correct methodology achieve following trying split product info seperate becomes difficult one item well different format product cake present html email received use power automate add solely first two order product product list remove return space replace make easier plu item quantity price search return relavent information example item process work one product listed get complicated one item purchased cake purchased additional line desired output multiple product desired output cake product sample'

Cleaned


Word Cloud of Cleaned Title & Text in r/excel
need
want
change
spreadsheet
name
date
different
thank
use
example
make
range
multiple
onetime
knowlist
trying
way
help
possible
pivot
first
problem
workbook
based
find
day
next
look
set


Word Cloud of Cleaned Title & Text in r/vba
code
sub
end
dim
string
workbook
set
range
error
next
time
false
elseif
variable
nothing
anyone
use
userform
thank
document
help
name
true
integer
msgbox
private

Log Reg

```
Scoring based on Count Vectorizing:
Log Reg Average Cross-Val-Score, 5-folds: 0.7983050847457627
Log Reg Training Score: 0.9991525423728813
Log Reg Validation Score: 0.8244575936883629

Scoring based on TF-IDF Vectorizing:
Log Reg Average Cross-Val-Score, 5-folds: 0.8381355932203389
Log Reg Training Score: 0.9516949152542373
Log Reg Validation Score: 0.8422090729783037
```

Naïve Bayes

```
Scoring based on Count Vectorizing:
Naive Bayes Average Cross-Val-Score, 5-folds: 0.8076271186440678
Naive Bayes Training Score: 0.9203389830508475
Naive Bayes Validation Score: 0.8284023668639053

Scoring based on TF-IDF Vectorizing:
Naive Bayes Average Cross-Val-Score, 5-folds: 0.8050847457627119
Naive Bayes Training Score: 0.9432203389830508
Naive Bayes Validation Score: 0.8027613412228797
```

Random Forest

```
Scoring based on Count Vectorizing:
Random Forest Average Cross-Val-Score, 5-folds: 0.8288135593220339
Random Forest Training Score: 1.0
Random Forest Validation Score: 0.8382642998027613

Scoring based on TF-IDF Vectorizing:
Random Forest Average Cross-Val-Score, 5-folds: 0.8296610169491526
Random Forest Training Score: 1.0
Random Forest Validation Score: 0.8382642998027613
```

Extra Trees

```
Scoring based on Count Vectorizing:
Extra Trees Average Cross-Val-Score, 5-folds: 0.8203389830508474
Extra Trees Training Score: 1.0
Extra Trees Validation Score: 0.8441814595660749

Scoring based on TF-IDF Vectorizing:
Extra Trees Average Cross-Val-Score, 5-folds: 0.8288135593220339
Extra Trees Training Score: 1.0
Extra Trees Validation Score: 0.8382642998027613
```

Support Vector Machine Classifier

```
Scoring based on Count Vectorizing:
Support Vector Machine Average Cross-Val-Score, 5-folds: 0.8025423728813559
Support Vector Machine Training Score: 0.9330508474576271
Support Vector Machine Validation Score: 0.8303747534516766

Scoring based on TF-IDF Vectorizing:
Support Vector Machine Average Cross-Val-Score, 5-folds: 0.8347457627118644
Support Vector Machine Training Score: 0.9957627118644068
Support Vector Machine Validation Score: 0.8500986193293886
```

Conclusions:

Over-fitted!

tvec > cvec

Model	Best Training Accuracy	Validation Accuracy	Precision	Recall	f1-score	TN	FP	FN	TP
Log Reg	0.84	0.842	0: 0.81 1: 0.87	0: 0.84 1: 0.84	0: 0.83 1: 0.86	190	36	44	237
Naïve Bayes	0.831	0.836	0: 0.84 1: 0.84	0: 0.79 1: 0.88	0: 0.81 1: 0.86	178	48	35	246
Random Forest	0.842	0.838	0: 0.79 1: 0.88	0: 0.86 1: 0.82	0: 0.83 1: 0.85	195	31	51	230
Extra Trees	0.841	0.866	0: 0.82 1: 0.91	0: 0.90 1: 0.84	0: 0.86 1: 0.87	204	22	46	235
SVM	0.84	0.85	0: 0.81 1: 0.88	0: 0.86 1: 0.84	0: 0.84 1: 0.86	195	31	45	236
Gradient Boost	0.814	0.836	0: 0.76 1: 0.92	0: 0.92 1: 0.77	0: 0.83 1: 0.84	207	19	64	217
AdaBoost	0.814	0.826	0: 0.76 1: 0.90	0: 0.89 1: 0.77	0: 0.82 1: 0.83	202	24	64	217

Problem Statement: Minimize the mis-classification of VBA queries (True) as Excel Queries (Predicted)

Model	Best Training Accuracy	Validation Accuracy	Precision	Recall	f1-score	TN	FP	FN	TP
Log Reg	0.84	0.842	0: 0.81 1: 0.87	0: 0.84 1: 0.84	0: 0.83 1: 0.86	190	36	44	237
Naïve Bayes	0.831	0.836	0: 0.84 1: 0.84	0: 0.79 1: 0.88	0: 0.81 1: 0.86	178	48	35	246
Random Forest	0.842	0.838	0: 0.79 1: 0.88	0: 0.86 1: 0.82	0: 0.83 1: 0.85	195	31	51	230
Extra Trees	0.841	0.866	0: 0.82 1: 0.91	0: 0.90 1: 0.84	0: 0.86 1: 0.87	204	22	46	235
SVM	0.84	0.85	0: 0.81 1: 0.88	0: 0.86 1: 0.84	0: 0.84 1: 0.86	195	31	45	236
Gradient Boost	0.814	0.836	0: 0.76 1: 0.92	0: 0.92 1: 0.77	0: 0.83 1: 0.84	207	19	64	217
AdaBoost	0.814	0.826	0: 0.76 1: 0.90	0: 0.89 1: 0.77	0: 0.82 1: 0.83	202	24	64	217

Naïve Bayes Classifier Accuracy is

28.2 ppts

higher than baseline

```
: train['subreddit'].value_counts(normalize=True)

: vba      0.554238
  excel    0.445762
  Name: subreddit, dtype: float64
```


Best Parameters: Multinomial Bayes Classifier
Alpha = 1
Max df
= 0.85
Max features
= 4800
Min df
= 2
ngram range
= (1, 3)

Top predictive words

conditional formatting rule

find sum

Planet Excel

countifs

power pivot

power query editor
powerquery

show total

make chart

create conditional formatting

sumifs

analysis toolpak

error

end sub

sub

Planet VBA

end

true

dim

code

run

loop

string

Examples of Mis-Classified Post (False Negatives)

Original

'changing pivot chart axes. I have a pivot chart I created from my power pivot data model. I want to write a macro to change the axes from the "date" field to the "month" field on the calendar table. I'm having trouble assigning the axis object and figuring out which method to use to change this so the chart shows data by month instead of by date.'

'change excel formula based on another cell I cannot wrap my brain around how to use VBA to change the formula based on a value in another cell. For example, if cell A1 = \'XYZ\', use formula "=B1+C1". If cell A1 = \'ABC", use formula "=B1+D1".\n\nI was thinking I would put the formulas in another sheet and use VLOOKUP to get the formula to use\n\nXYZ "=B+C"\n\nABC ="B+D"\n\nThe part I cannot figure out is how to get the row numbers into the formula using VBA (ie, =B1+C1")\n\nAny help would be appreciated.\n\nThanks.'

'excel dropdown list based on another cell's value Hey guys, I really need any advice here!\n\nI've done with google but literally found nothing relevant for what I need. It's like not much people use drop down list.\n\n\nSo basically I need to create a dropdown list for excel that the items in the list gets added and removed dynamically based on another cell's value. Here is a rough example:-\n\nITEM NUMBER\n\n A 10\n\n B 4\n\n C 0\n\n D 1\n\nIt's on table.\n\n\nThe DropDown list is referenced to the ITEM column and if the values in the adjacent column NUMBER is greater than 0 the ITEM adjacent to it must be shown in DropDown list. \n\nand ITEM with value 0 in it's adjacent NUMBER column must be excluded from the DropDown list, \n\nHence ITEM C should be removed from the DropDown List once it reaches 0 and the remaining A B D will not be affected. \n\n\nI've tried VBA but I'm not good at it and neither google was any useful, so excuse me if I couldn't provide with an existing code I worked on,\n\nI've managed to do this in excel with formulas \nbut it ended up removing the last row every time instead of the cell C, hence no success here too.'

Cleaned

The 'Pivot-ed-to-become-predicted-excel' post

'changing pivot chart ax pivot chart created power pivot model want write change ax date field month field calendar trouble assigning axis object figuring method use change chart show month instead date'

The 'one-keyword-that-changed-everything' post

'change based another cannot wrap brain around use change based another example xyz use abc use thinking put formula another use vlookup get use xyz abc part cannot figure get ie help appreciated thanks'

The 'list-abuser' post

'dropdown list based another hey guy really need advice done google literally found nothing relevant need much people use drop list basically need create dropdown list item list get added removed dynamically based another rough example item dropdown list referenced item adjacent greater item adjacent must shown dropdown list item adjacent must excluded dropdown list hence item removed drop down list reach remaining affected tried good neither google useful excuse provide existing code worked managed formula ended removing last every time instead hence success'

Excel predictive words

VBA predictive words

Can we do better?

```
# 1. convert text to lower case
lower_text = raw_text.lower()

# 2. Remove HTML type chars
review_text = BeautifulSoup(lower_text).get_text()

# 3. Remove http(s) type strings
review_text = re.sub("http.\S+", " ", review_text)

# 4. Remove pointer reference and zero-width spaces
review_text = re.sub("[&]\S+", " ", review_text)

# 5. removing \xa0 reference
review_text = re.sub("\\xa0", " ", review_text)

# 6. removing reference to excel and vba
review_text = re.sub("(vba)|(excel)", " ", review_text)

# 7. removing new line reference
review_text = re.sub("\\n", " ", review_text)

# 8. removing long words reference (more than 15 char)
review_text = re.sub("[\S]{15}\S+", " ", review_text)

# 9. removing non-char non-num reference and all numbers
review_text = re.sub("([^a-z])|([\d])", " ", review_text)

# 10. removing standalone characters and multiple spaces
review_text = re.sub("(^| ).(( ).)*( |$)", " ", review_text)

# 11. Split into individual words.
words = review_text.split()
```

Model	Best Training Accuracy	Validation Accuracy	Precision	Recall	f1-score	TN	FP	FN	TP
(Old) Naïve Bayes	0.831	0.836	0: 0.84 1: 0.84	0: 0.79 1: 0.88	0: 0.81 1: 0.86	178	48	35	246
(New) Naïve Bayes	0.848	0.876	0: 0.89 1: 0.89	0: 0.82 1: 0.92	0: 0.86 1: 0.89	186	40	23	258

34% reduction in False Negatives

Thank You