Text Analytics on Google App

Nan (Miya) Wang, Chang Yin, Ellin Iqra, Anshu Vyas

INTRODUCTION

FORDHAM UNIVERSITY
THE JESUIT UNIVERSITY OF NEW YORK

- Application development is an extremely profitable business
    - 50 Billion Dollars revenue by 2016
- About Half of developers make less than $100
- Using the openly data available on Google Play, we extracted some features as input to help:
    - Classify applications
    - Label updations
    - Predict success

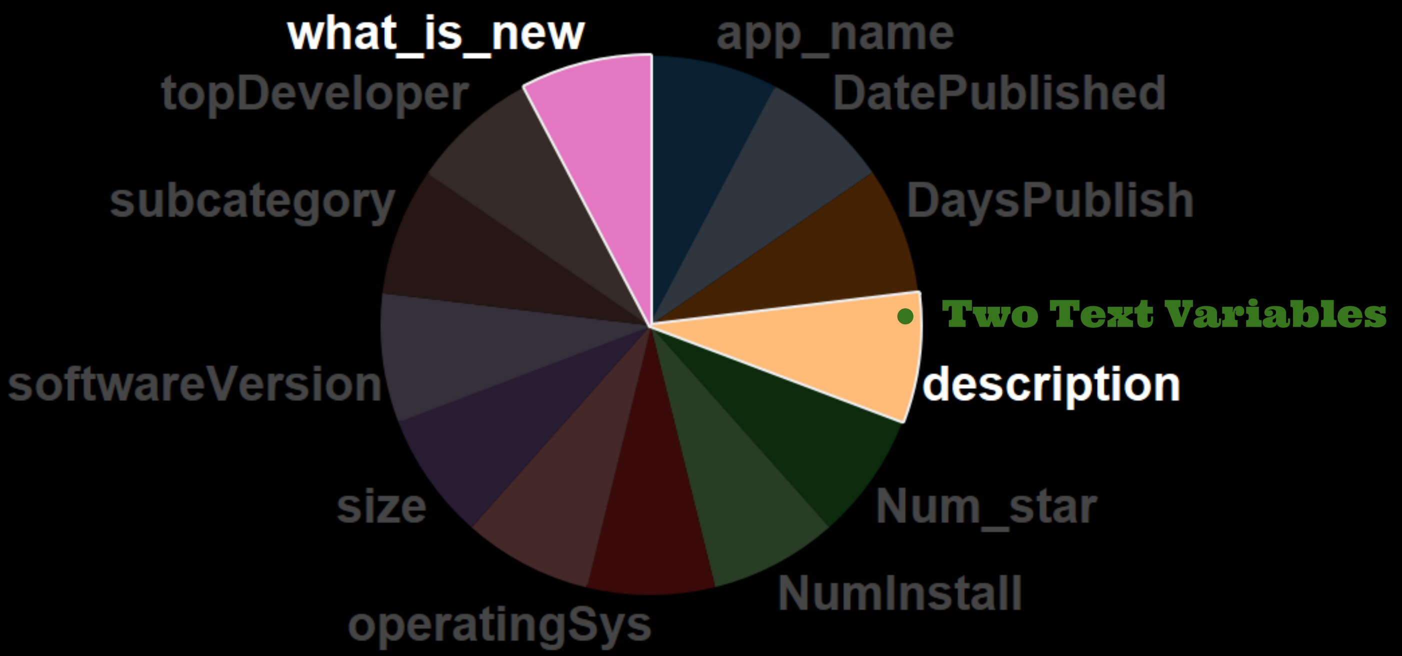
ANALYSIS OVERVIEW

TEXT CLUSTERING (Python)

TEXT CLASSIFICATION (Python)

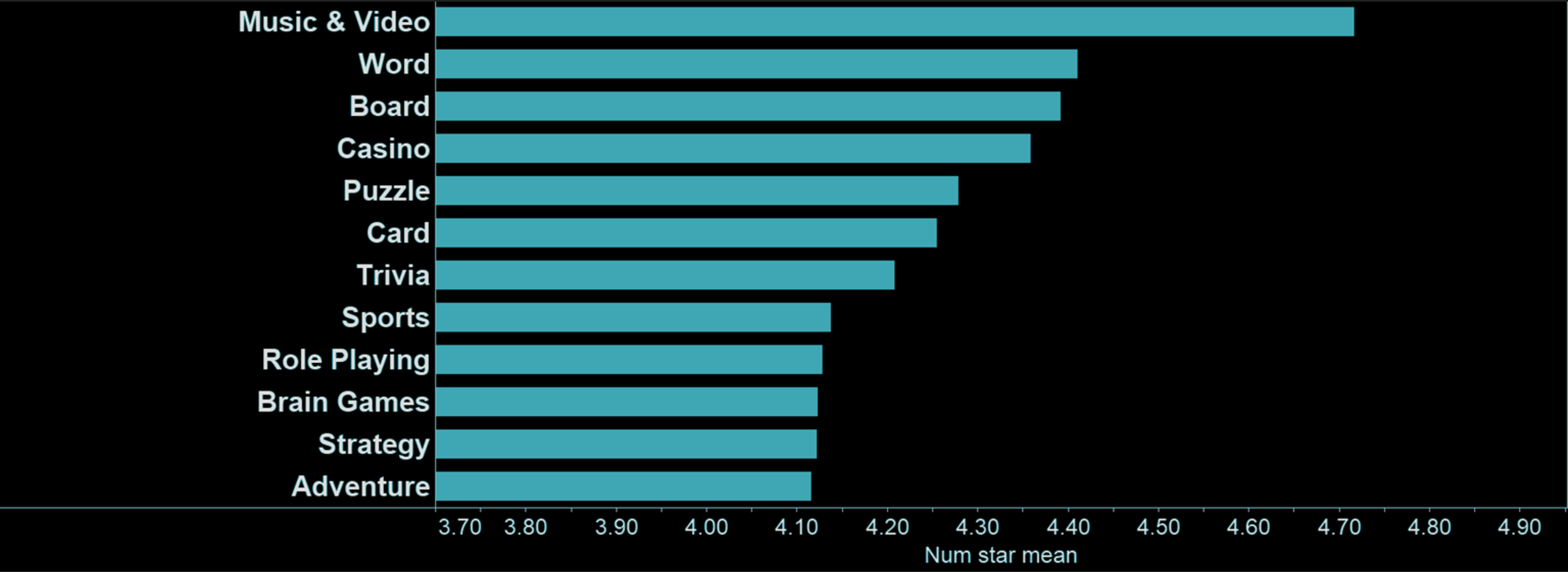
SUCCESS ANALYSIS (Tableau)

DATASET

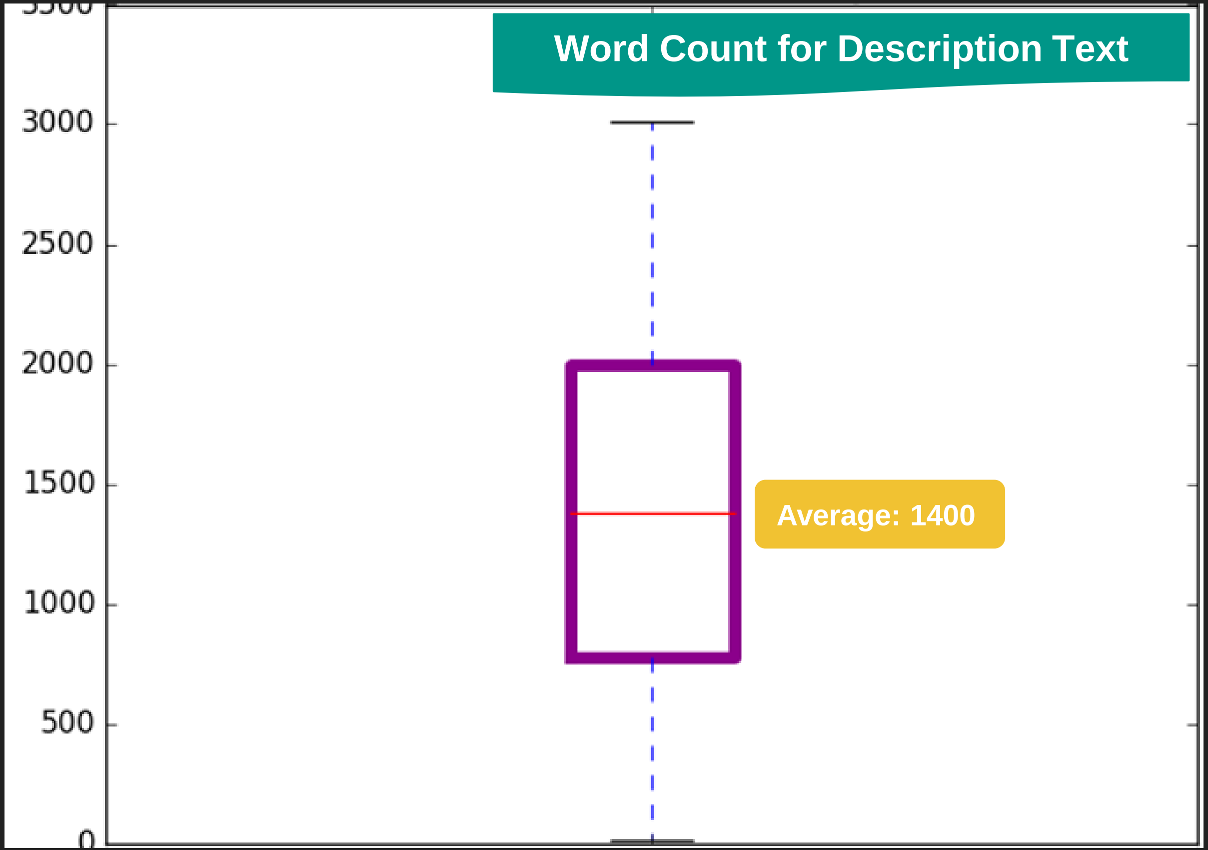


Source: https://play.google.com/store/apps/category/GAME/collection/topselling_new_free

EXPLANATORY ANALYSIS



EXPLANATORY ANALYSIS



TEXT PREPARATION

```
 u'play one success la vega slot game comfort home go mobil devic impress princess win wealth
glori golden knight scatter buck stack wild give quest rich reward download free collect welc
om bonu get start',
 u'kitti pawp avail play android',
 u'nan',
 u' new everi tile get bjoker tileb use wise onlin leaderboard googl play game bugfixesani su
ggest bug report welcomepleas write bad review contact us email ijdpappsgmailcomi',
```

```
%timeit
def remove_numbers(s):
    return s.translate(None, string.digits)

def lowercase_remove_punctuation(s):
    s = s.lower()
    s = s.translate(None, string.punctuation)
    return s
```

- Remove Numbers/Stopwords
- Lowercase
- Tokenize
- Stemming
- **Feature Engineering**

```
earn advanc concept game',
 u'graphic engin updatedfix imag doubl oneplu devic graphic artifact work gyroscop',
 u'first thank much play lost harmoni messag kind reviewson new version support nexu bug fix
tweak improv',
 u'ad levelsadjust difficulti levelsbug fixesremov ad',
 u'perform enhanc critic bug fixesthank play droppi ball',
 u'thank love continu work game adjust reviv mechan',
 u'fix certain bug',
```

```
    token_list = stem_token_list(token_list)
    return restring_tokens(token_list)

def all_work_for_whatisnew(s):
    s = remove_numbers(s)
    s = lowercase_remove_punctuation(s)
    s = remove_stopwords(s)
    token_list = word_tokenize(s)
    token_list = tag(token_list)
    token_list = stem_token_list(token_list)
    return restring_tokens(token_list)
```

TEXT CLUSTERING

FORDHAM UNIVERSITY
THE JESUIT UNIVERSITY OF NEW YORK

Feature Engineering—TFIDF

“min_df= 4”
“ngram range” =(1,5)
135 Features

```
# vectorize and re-weight
desc_vect = tfidf_vectorizer.fit_transform(what_is_new)
coo = desc_vect.tocoo(copy = False)
df_word_matrix = pd.DataFrame({'index': coo.row, 'feature': coo.col, 'data': coo.data}
               )[['index', 'feature', 'data']].sort_values(['index', 'feature']).reset_index(dr
word_matrix = df_word_matrix.pivot('index', 'feature', 'data')
word_matrix = word_matrix.fillna(0)
```

```
word_matrix.head()
```

feature	0	1	2	3	4	5	6	7	8	9	...	125	126	127	128	129	130	131	132	133	134
index																					
1	0	0	0	0	0	0.000000	0	0	0.000000	0	...	0	0	0	0	0.000000	0.000000	0	0.290277	0	0.000000
2	0	0	0	0	0	0.652527	0	0	0.627365	0	...	0	0	0	0	0.000000	0.000000	0	0.000000	0	0.000000
4	0	0	0	0	0	0.000000	0	0	0.000000	0	...	0	0	0	0	0.000000	0.000000	0	0.000000	0	0.000000
6	0	0	0	0	0	0.000000	0	0	0.000000	0	...	0	0	0	0	0.129226	0.493974	0	0.000000	0	0.164658
7	0	0	0	0	0	0.000000	0	0	0.000000	0	...	0	0	0	0	0.000000	0.000000	0	0.000000	0	0.000000

5 rows × 135 columns

Top 20 TFIDF Terms

```
[(u'minor bug fix', 1.0),
 (u'us', 0.92468964501509099),
 (u'player', 0.80887092392482673),
 (u'multipl', 0.77569328502893675),
 (u'spin', 0.76812347947729964),
 (u'updat new', 0.74707184593519893),
 (u'find', 0.67876051148278016),
 (u'updat', 0.66474330159165551),
 (u'bonu', 0.65252742204348113),
 (u'sound', 0.64030174158414288),
 (u'stabil', 0.63363569815498977),
 (u'mission', 0.63111007562945531),
 (u'bug', 0.62736513460690246),
 (u'first', 0.62363360382935040),
 (u'user experi', 0.60771003503531418),
 (u'increas', 0.60470548038951843),
 (u'user', 0.58710214295216057),
 (u'includ', 0.55980659470124550),
 (u'store', 0.55586785683546935),
 (u'thank', 0.51545319048405958)]
```

TEXT CLUSTERING

Feature Engineering—POS

```
d = (zip(tfidf_vectorizer.get_featu          ct.data))
df_clustering = DataFrame(d)
df_clustering.columns = ['feature',
df_clustering['Feature_NO'] = range
map(lambda x:x[1],nltk.pos_tag(df_c
df_clustering['tag'] = map(lambda x          df_clustering.feature))
df_clustering = df_clustering.sort          ing = False)
df_clustering.groupby('tag').count
```

```
df_clustering.loc[df_clustering['ta
```

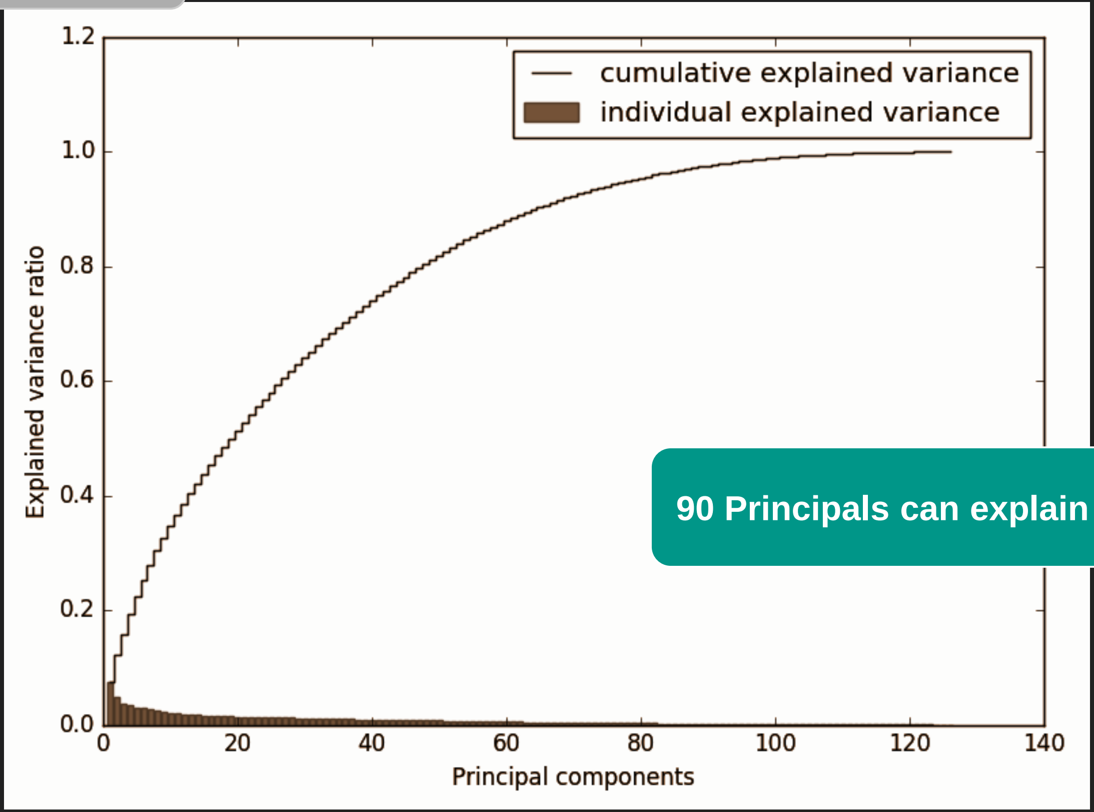
	feature	idf	Feature_NO	tag
50	find	0.678761	50	VBP
40	environ	0.493974	40	VBP
124	use	0.380722	124	VBP
72	issu	0.376592	72	VBP
22	collect	0.348295	22	VBP
103	reward	0.331995	103	VBP
77	make	0.312824	77	VBP
3	adjust	0.290277	3	VBP

CD	1
FW	1
IN	2
JJ	26
JJR	1
NN	75
NNS	3
PRP	1
RB	2
RBR	1
RBS	1
VB	6
VBD	1
VBN	2
VBP	12

'DT'
'JJ'
'JJS',
'JJR'
'NN'
'NNP'
'RB'
'VB'
'VBP'
'VBZ'
'RBR'
'VBD'
'VBN'

TEXT CLUSTERING

Feature Engineering—PCA



90 Principals can explain all the variance

TEXT CLUSTERING

FORDHAM UNIVERSITY
THE JESUIT UNIVERSITY OF NEW YORK

Model Building

```
sklearn_pca = PCA(n_components=9(
word_matrix_pca = sklearn_pca.fit                    _feat

from sklearn.cluster import KMear
km = KMeans(n_clusters = 4)
%time
km.fit(word_matrix_pca)
clusters = km.labels_.tolist()
df_what_is_new_ngram= DataFrame(                     index, 'cluster': clusters})
df_what_is_new_ngram.groupby('clu
content = []
for i in df_what_is_new_ngram.wha
    content.append(what_is_new[i]
df_what_is_new_ngram['content'] =
df_what_is_new_ngram.groupby('clu
```

	what_is_new
cluster	
0	18
1	87
2	26
3	24
4	42

KMeans

Only 197 unique text about update for 2227 APPs.

TEXT CLUSTERING

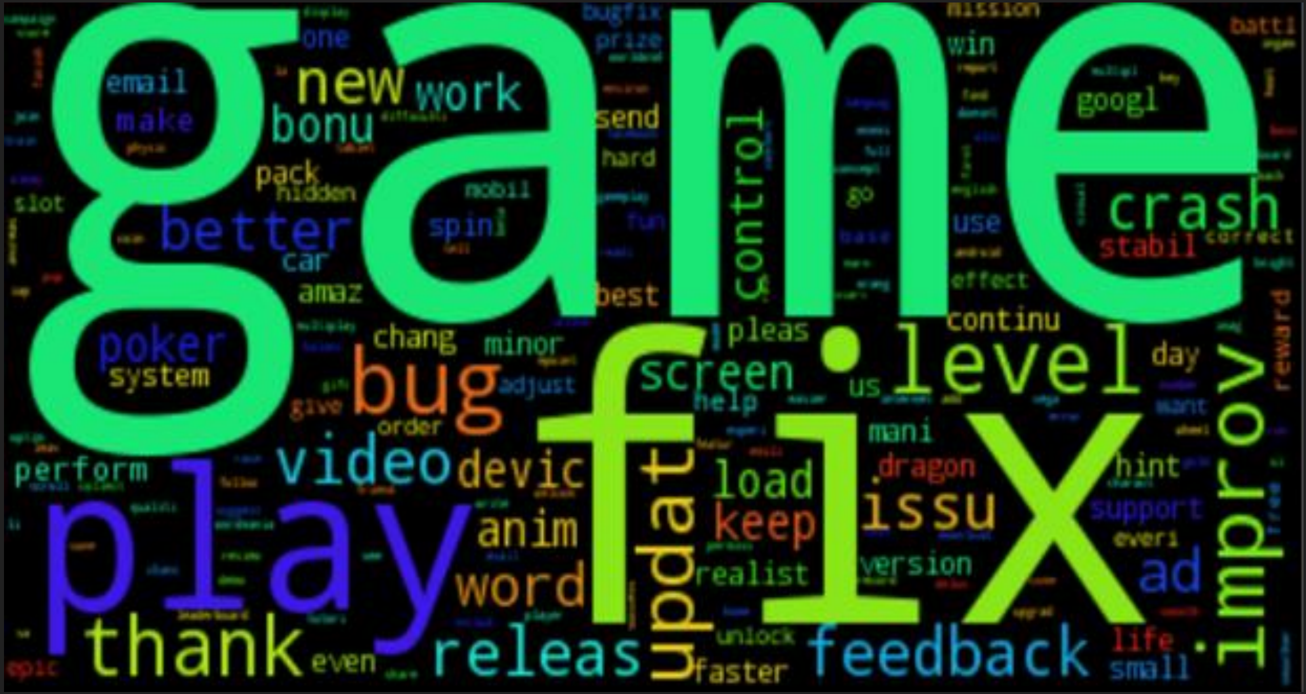
FORDHAM UNIVERSITY
THE JESUIT UNIVERSITY OF NEW YORK

Evaluation—Cluster 0

FORDHAM UNIVERSITY
THE JESUIT UNIVERSITY OF NEW YORK

TEXT CLUSTERING

Evaluation—Cluster 1



TEXT CLUSTERING

Evaluation—Cluster 2

TEXT CLUSTERING

FORDHAM UNIVERSITY
THE JESUIT UNIVERSITY OF NEW YORK

Evaluation—Cluster 3

TEXT CLUSTERING

Evaluation—Cluster 4

TEXT CLASSIFICATION

FORDHAM UNIVERSITY
THE JESUIT UNIVERSITY OF NEW YORK

Preprocessing

How Many App Categories

```
len(set(tuple(description_for_text.subcategory.tolist())))
```

```
23
```

Choose Just 10 out of 23 categories

Top Ten Categories

```
Ten_cate = description_for_text.groupby('subcategory').count().sort('index',ascending = False).i
Ten_cate
```

```
C:\Users\Miya\AppData\Local\Enthought\Canopy\User\lib\site-packages\ipykernel\__main__.py:1:
FutureWarning: sort(columns=....) is deprecated, use sort_values(by=.....)
  if __name__ == '__main__':
```

```
['Casual',
 'Puzzle',
 'Simulation',
 'Action',
 'Arcade',
 'Adventure',
 'Casino',
 'Racing',
 'Role Playing',
 'Sports']
```

TEXT CLASSIFICATION

FORDHAM UNIVERSITY
THE JESUIT UNIVERSITY OF NEW YORK

Feature Engineering —Parameter Setting

```
tfidf_vectorizer = TfidfVectorizer(
    min_df= 2,   # min count for relevant vocabulary
    max_features=100000,  # maximum number of features
    strip_accents='unicode',  # replace all accented unicode char
    # by their corresponding  ASCII char
    analyzer='word',  # features made of words
    token_pattern=u'[a-z]+',  # tokenize only words of 4+ chars
    ngram_range=(1, 2),  # features made of a single tokens
    use_idf=True,  # enable inverse-document-frequency reweighting
    smooth_idf=True,  # prevents zero division for unseen words
    sublinear_tf=False)
```

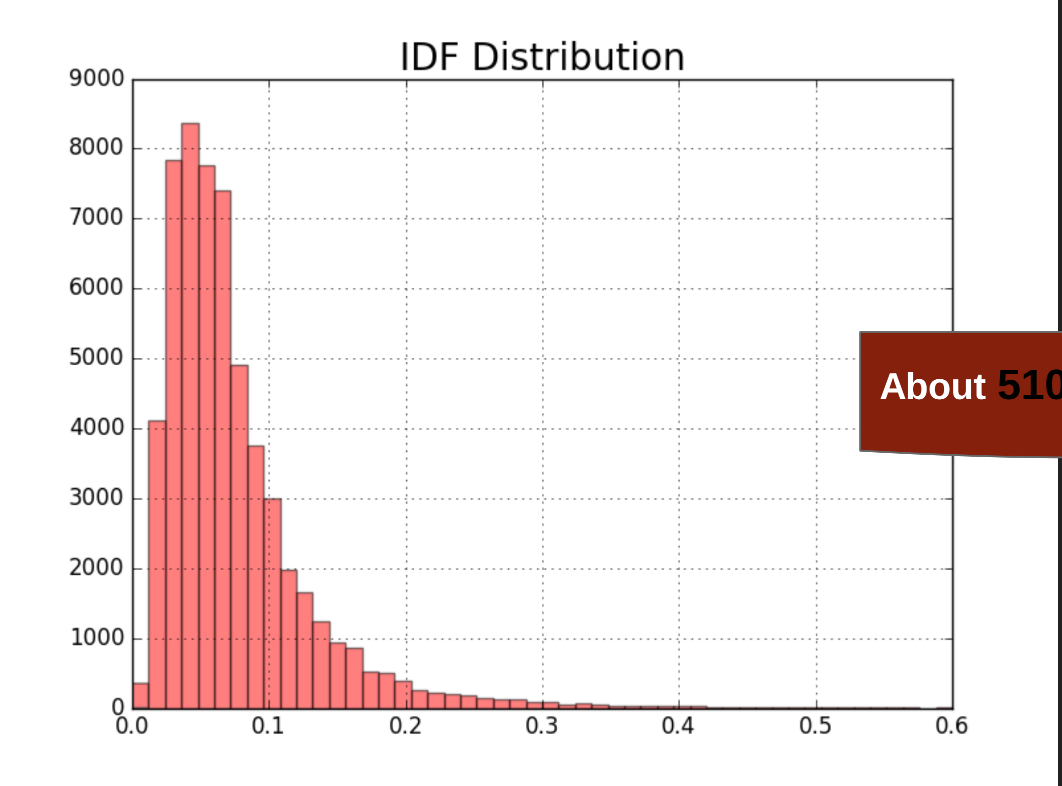
TEXT CLASSIFICATION

FORDHAM UNIVERSITY
THE JESUIT UNIVERSITY OF NEW YORK

Feature Engineering — POS

TEXT CLASSIFICATION

Feature Engineering — TFIDF



Drop Features with TFIDF larger than 0.074 and smaller than 0.3

About 51000 features reduced to 20040

TEXT CLASSIFICATION

Model Building

```
clf = GaussianNB()
#clf.fit(desc_matrix_less_features,description_for_text['subcategory'])
```

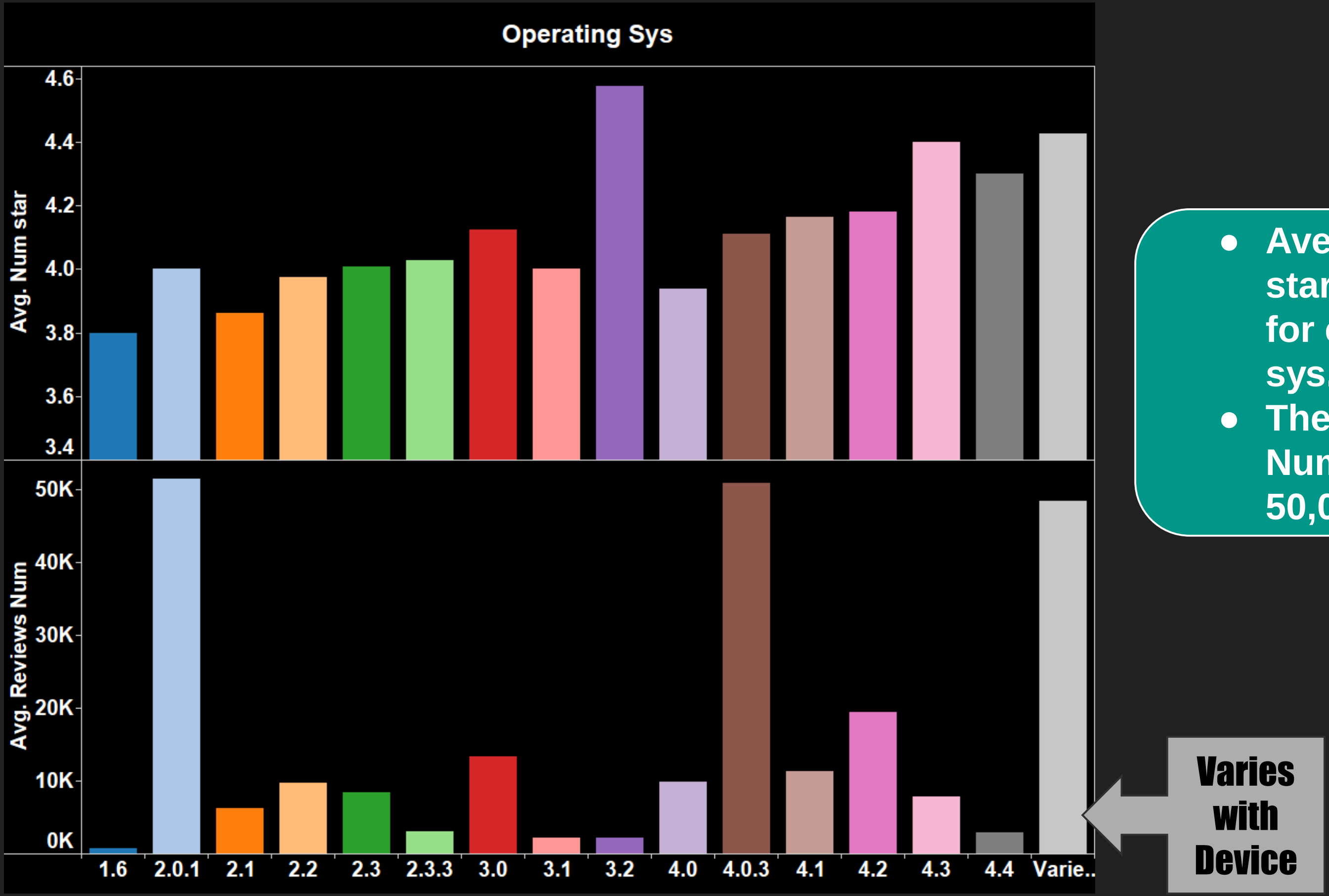
```
%timeit
scores_gb = cross_val_score(clf,
                            desc_matrix_less_features,description_for_text['subcategory'],c
                            v =10, scoring = 'accuracy')
```

```
scores_gb.mean()
```

```
0.71821866138887236
```

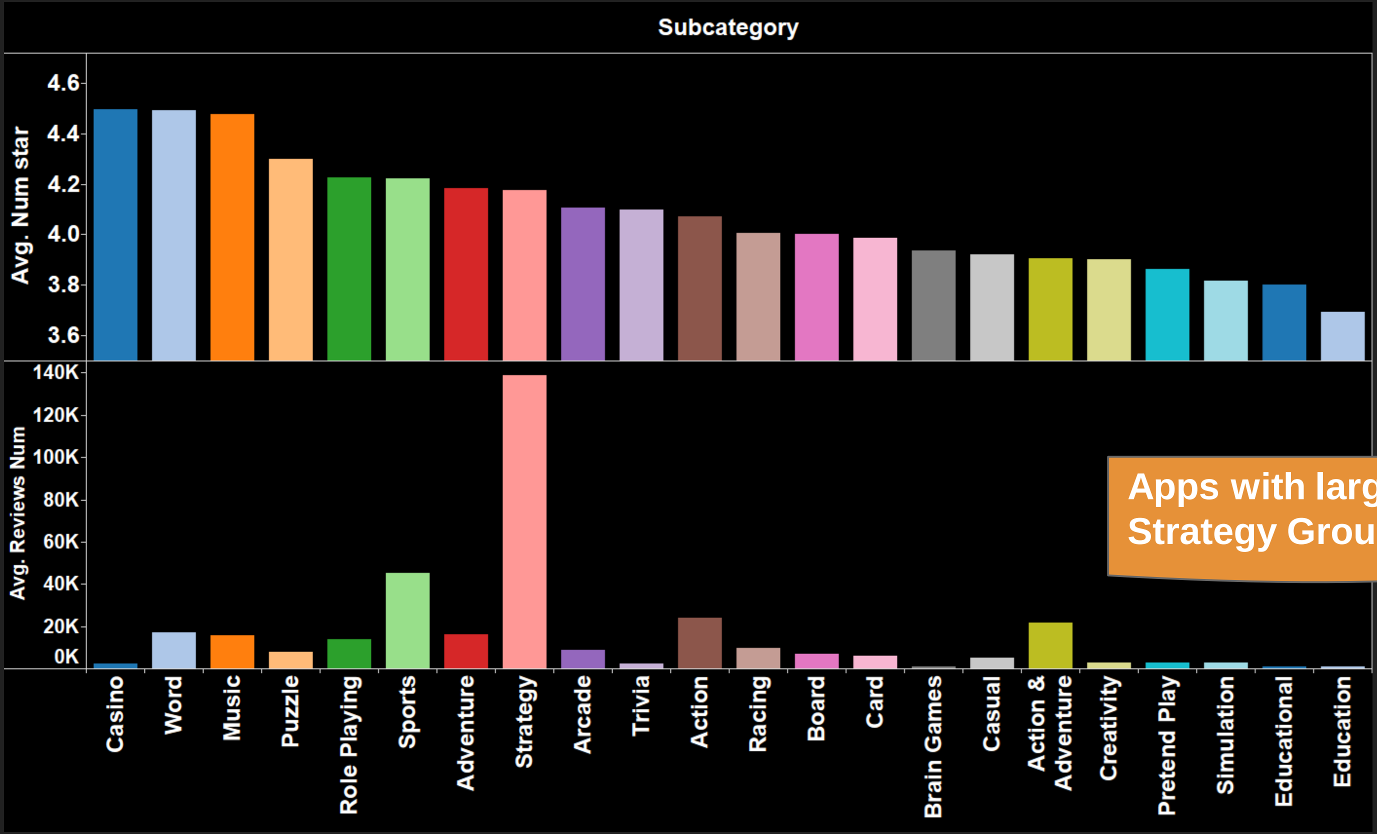
SUCCESS ANALYSIS

FORDHAM UNIVERSITY
THE JESUIT UNIVERSITY OF NEW YORK



- Average of Num of star and review num for each operating sys.
- The data is filtered on Num Install over 50,000

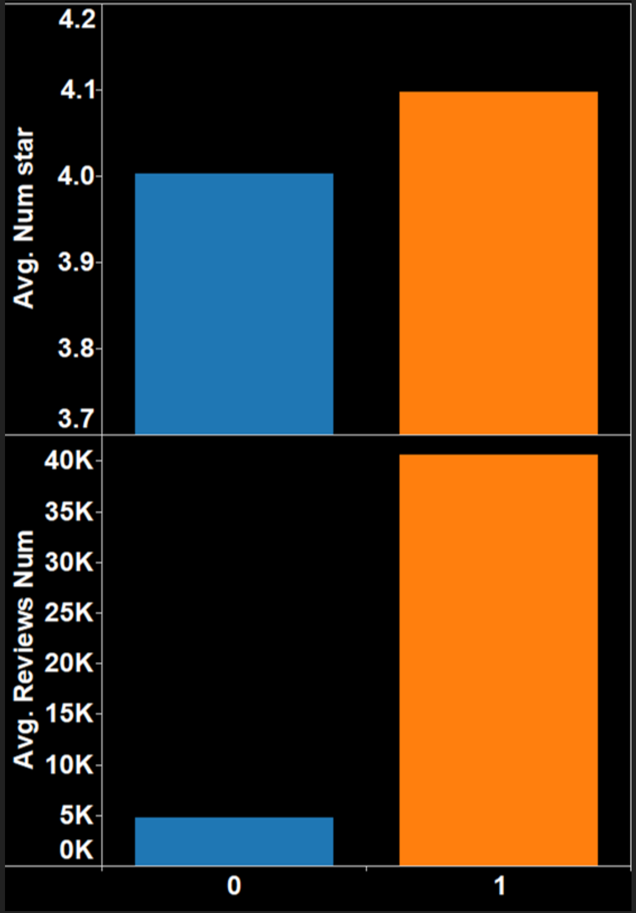
SUCCESS ANALYSIS



- Average of Num of star and review num for each subcategory.
- Data is filtered on Num Install over 50,000

Apps with largest installations are all in Strategy Group.

SUCCESS ANALYSIS



- Average of Num of star and review num for Top Developer or not.
- Data is filtered on Num Install over 50,000

CONCLUSION

FORDHAM UNIVERSITY
THE JESUIT UNIVERSITY OF NEW YORK

- **Five labels available for application updation:**
  - ➢ Add New Features
  - ➢ GamePlay Modification
  - ➢ Improve Levels
  - ➢ Fix Bugs & New Versions
  - ➢ Fix Minor Bugs & Optimization

- **71% of Accuracy for classify apps into 10 subcategories.**

- **App Subcategory, Developer, Operating System have correlation relationship with APP success.**