Reddit Comment Classification

—

Shawn Mitchell MBA Msc

Data Science Question

Are the comments between two given sub-Reddits significantly different enough for a machine learning model to categorize the comments into their respective groups?

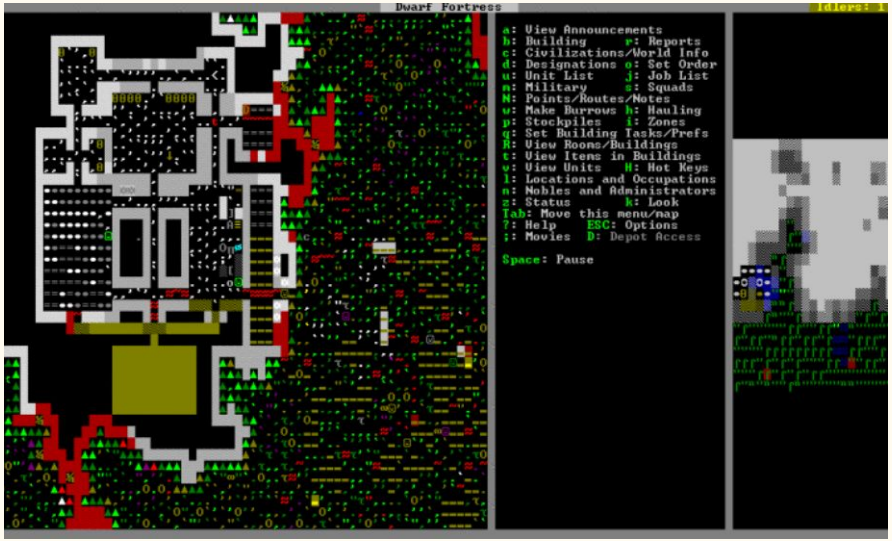
Factors that will influence this:

- Model selection/Hyperparameters/EDA
- Sub-Reddit selection
- Minimum criteria for accepting ‘success’ or viability


STELLARIS™
GRAND STRATEGY ON A GALACTIC SCALE

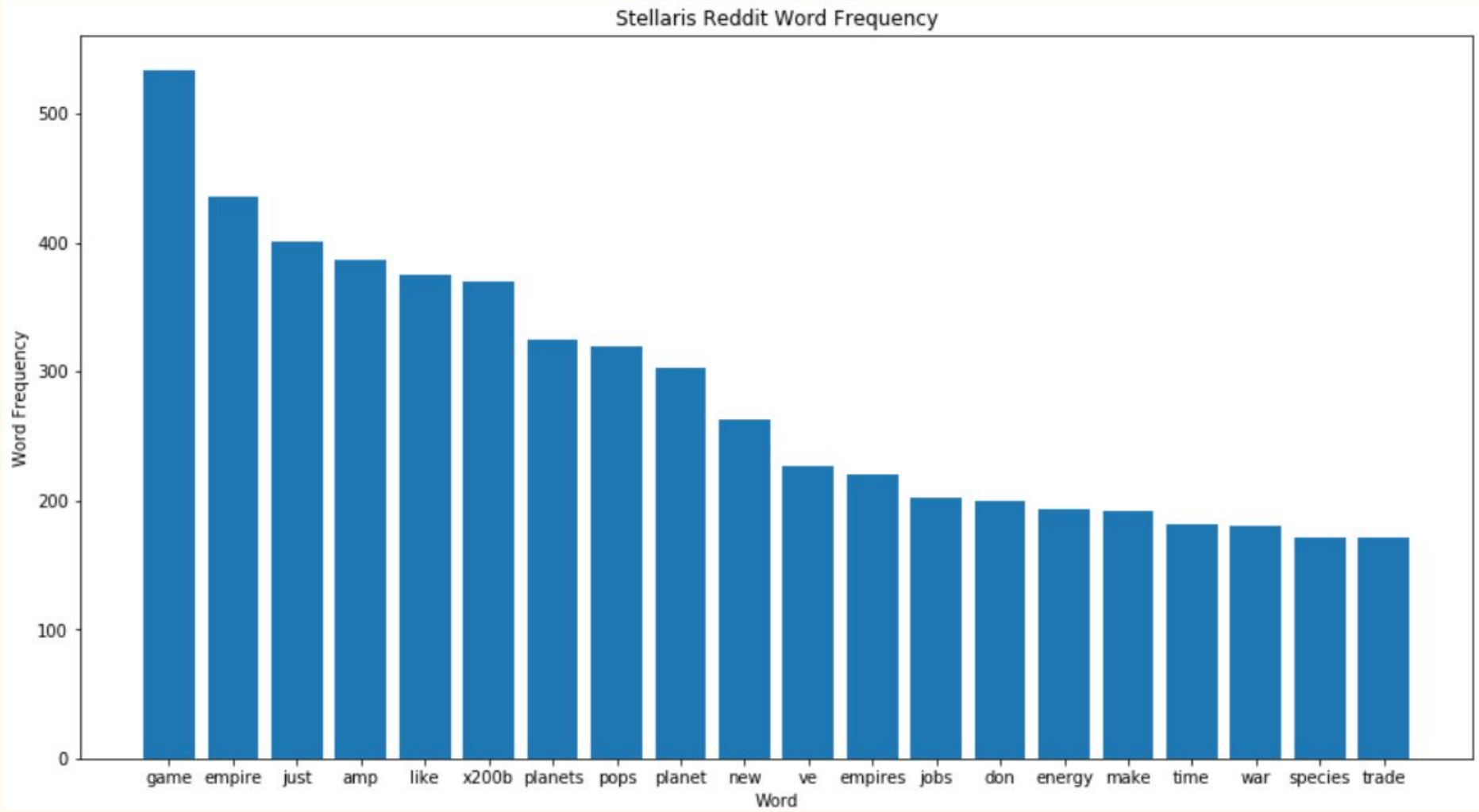

Dwarf
Fortress

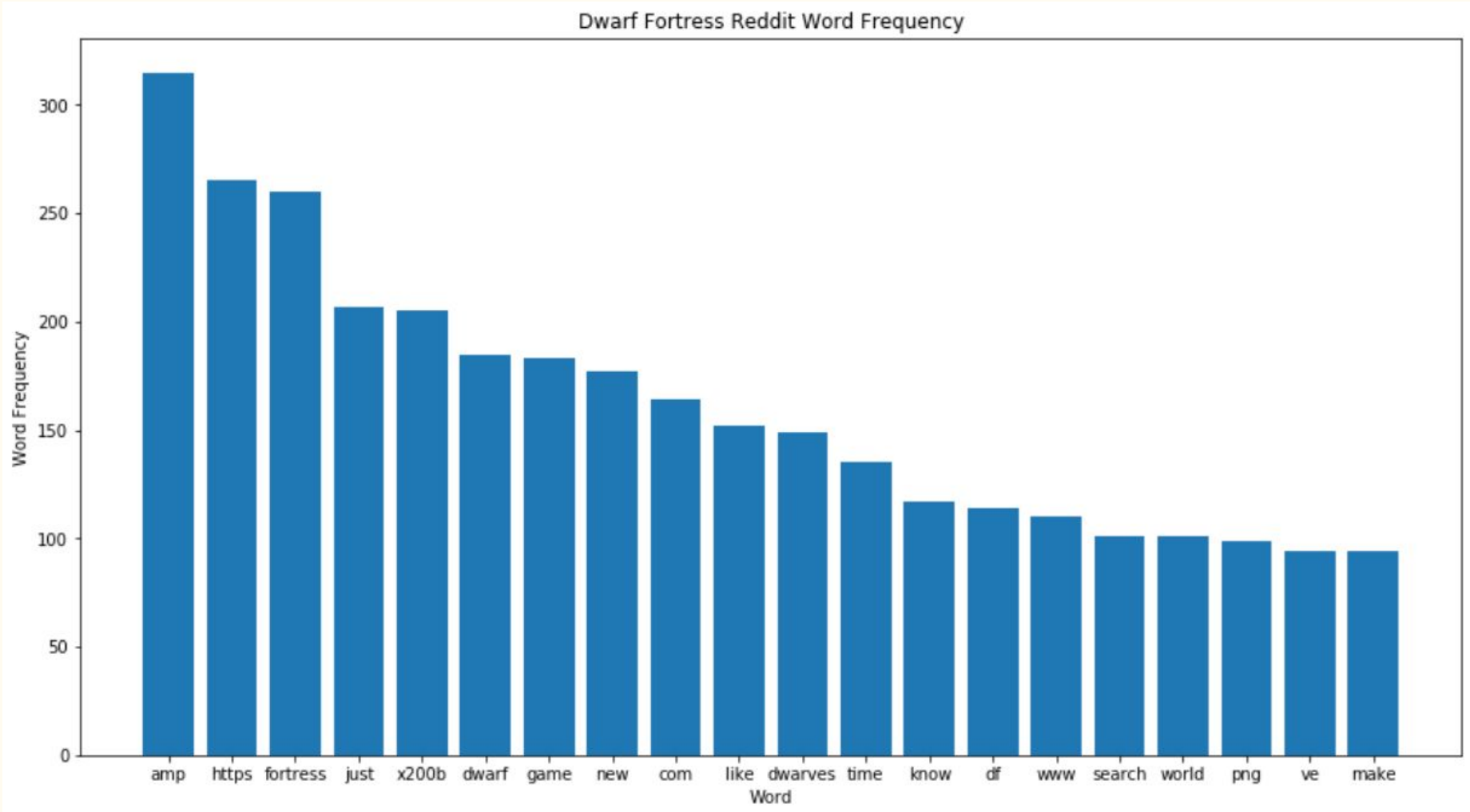

Gox Syndicate
Gox Syndicate
Cilxtac Hierarchy
Photecian Assembly
United Djomar Coalit
Zracon Combine
Outliner
Planets
Military Fleets
Civilian Ships
Armies
Observation
Descorax
Science Ship
Close Galaxy Map


Dwarf Fortress
a: View Announcements
b: Building r: Reports
c: Civilizations/World Info
d: Designations o: Set Order
u: Unit List j: Job List
m: Military s: Squads
N: Points/Routes/Notes
v: Make Burrows h: Hauling
p: Stockpiles i: Zones
q: Set Building Tasks/Prefs
R: View Rooms/Buildings
t: View Items in Buildings
v: View Units H: Hot Keys
l: Locations and Occupations
n: Nobles and Administrators
z: Status k: Look
Tab: Move this menu/map
?: Help ESC: Options
: Movies D: Depot Access
Space: Pause

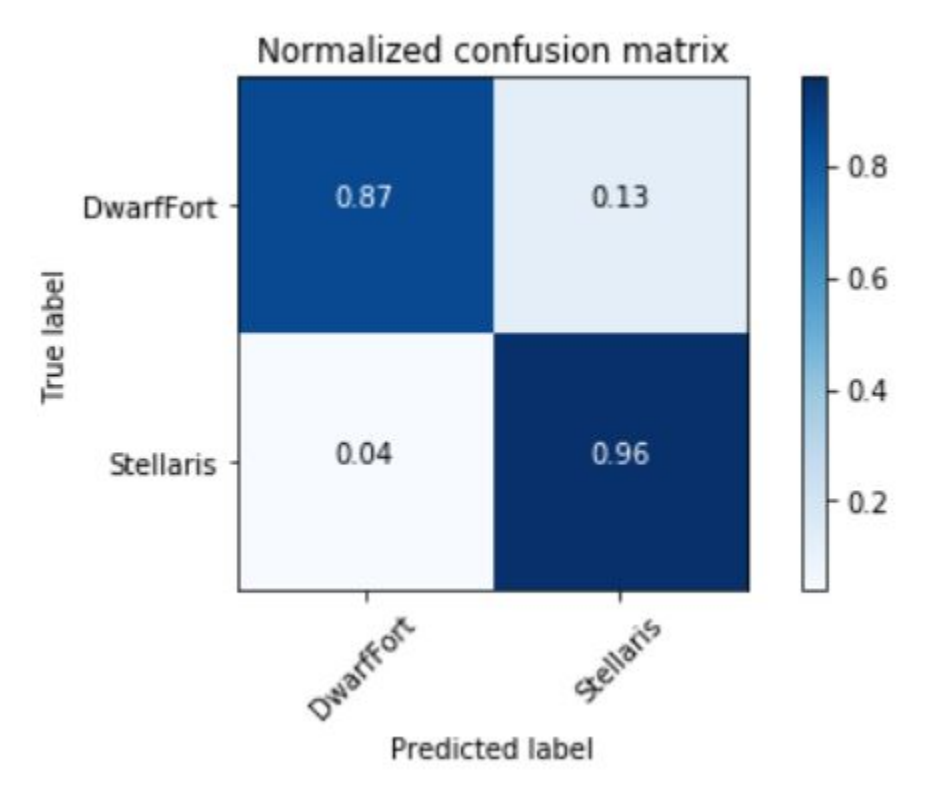
Word Commonality

[('amp', 568),
('game', 533),
('just', 474),
('x200b', 463),
('like', 409),
('new', 357),
('empire', 351),
('https', 306),
('planets', 258),
('pops', 256),
('ve', 255),
('time', 250),
('planet', 228),
('make', 226),
('fortress', 224),
('com', 208),
('don', 204),
('know', 197),
('empires', 186),
('jobs', 181)]


Stellaris Reddit Word Frequency
Word Frequency
0
100
200
300
400
500
game
empire
just
amp
like
x200b
planets
pops
planet
new
ve
empires
jobs
don
energy
make
time
war
species
trade
Word


Dwarf Fortress Reddit Word Frequency
Word Frequency
0
50
100
150
200
250
300
amp
https
fortress
just
x200b
dwarf
game
new
com
like
dwarves
time
know
df
www
search
world
png
ve
make
Word

Model Scoring - SGD Classifier, TFIDF



13% mislabelled as Stellaris

4% mislabelled as Dwarf Fortress

```
Accuracy Score train:
1.0
Accuracy Score test:
0.9301470588235294
```

Misclassification

Post incorrectly predicted to be: DwarfFort
['for example if i want to have an empire centered around human refugees from earth living on an ocean world. humans can live there, but they\'d rather be on a continental world.\r\n\r\nan alternative could be a "refugee regime" ethic, where you could create an original home world kind of like how syncretic evolution lets you design a syncretic species. \r\n\r\nmaybe both could work together, enabling your species to be native to a tomb world that used to be an continental world, but now live on a jungle moon somewhere.']

Misclassification

Post incorrectly predicted to be: DwarfFort
['\r\n\r\nhttps://i.redd.it/z4fn2i2p4l421.png']

Post incorrectly predicted to be: Stellaris
['yay! thanks to /u/insert_gnome_here for being the first to report it being back up!']

Post incorrectly predicted to be: Stellaris
['pretty much that. i find it disappointing i can have a top level cpu and gpu, yet play lowering the 4x4 starting grid to 2x2, and still have to optimise my builds and strategy to limit my fps plummeting']

Misclassification

Post incorrectly predicted to be: DwarfFort
["the reason for this is simple: castor is a hexinary star system and it's really neat.\r\n\r\na and b orbit each other and that whole group is orbited by c but! all 3 sets are actually binaries with white dwarves.\r\n\r\nwould be a really fun system to have in there.\r\n\r\ni've not tried my hand at modding or i'd try to make it myself."]

Conclusion - Business Case

- Identifying discussions relating to your business or competitors
- Sentiment analysis
- Spam filtering
- Email/communication organization
- Database lookups (support wiki)