

```
        ),
        default=1.0,
    )
global_scale_setting = FloatProperty(
    name="Scale",
    min=0.01, max=1000.0,
    default=1.0,
)

def execute(self, context):
    # get the folder
    folder_path = (os.path.dirname(self.filepath))

    # get objects selected in the viewport
    viewport_selection = bpy.context.selected_objects

    # get export objects
    obj_export_list = viewport_selection
    if self.use_selection_setting == False:
        obj_export_list = [i for i in bpy.context.scene.objects]

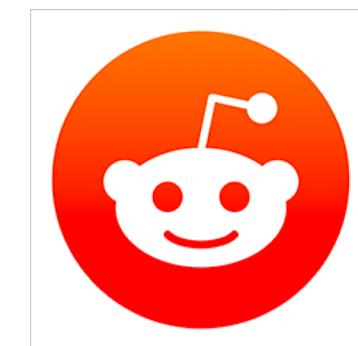
    # deselect all objects
    bpy.ops.object.select_all(action='DESELECT')

    for item in obj_export_list:
        item.select = True
        if item.type == 'MESH':
            file_path = os.path.join(folder_path, "{}.obj"
            bpy.ops.export_scene.obj(filepath=file_path,
                axis_forward=self.axis_forward,
                axis_up=self.axis_up,
                use_animation=self.use_animation,
                use_mesh_modifiers=self.use_mesh_modifiers,
                use_edges=self.use_edges,
                use_smooth_groups=self.use_smooth_groups,
                use_smooth_grouping=self.use_smooth_grouping,
                use_normals=self.use_normals,
                use_uvs=self.use_uvs,
                use_materials=self.use_materials,
                use_triangles=self.use_triangles,
                use_subdivision=self.use_subdivision,
                use_alpha=self.use_alpha,
                use_armature=self.use_armature,
                use_armature_nod
```

# NATURAL LANGUAGE PROCESSING

## A REDDIT STUDY

PROJECT 3 – BRYCE GILLESPIE



# THE PROBLEM

With approximately 14,000 reddit post titles and accompanying text, can we use natural language processing to build a classification model that will accurately predict whether the post was made by a bot?

OF COURSE I WILL BE UNDETECTABLY HUMAN.

*u/totallynotrobots\_SS*

A photograph showing a person from behind, wearing headphones and sitting at a desk with multiple computer monitors. One monitor displays what appears to be a coding interface or a complex data visualization. The scene is set in an office environment with other people visible in the background.

# DATA SOURCES

## S U B R E D D I T S I M U L A T O R

A subreddit entirely populated by bots. The main account posts a new thread every 6 hours generated by applying Markov Chains to the top 500 threads in r/all over the last 24 hours.

Every hour, a post is made by a single bot trained on one of several hundred subreddits in the same manner.

Every three minutes, a comment is posted by a similar bot.

## A S K R E D D I T

A general subreddit dedicated to asking and answering thought-provoking questions with nearly 21 million subscribers. The posts cover many different topics.

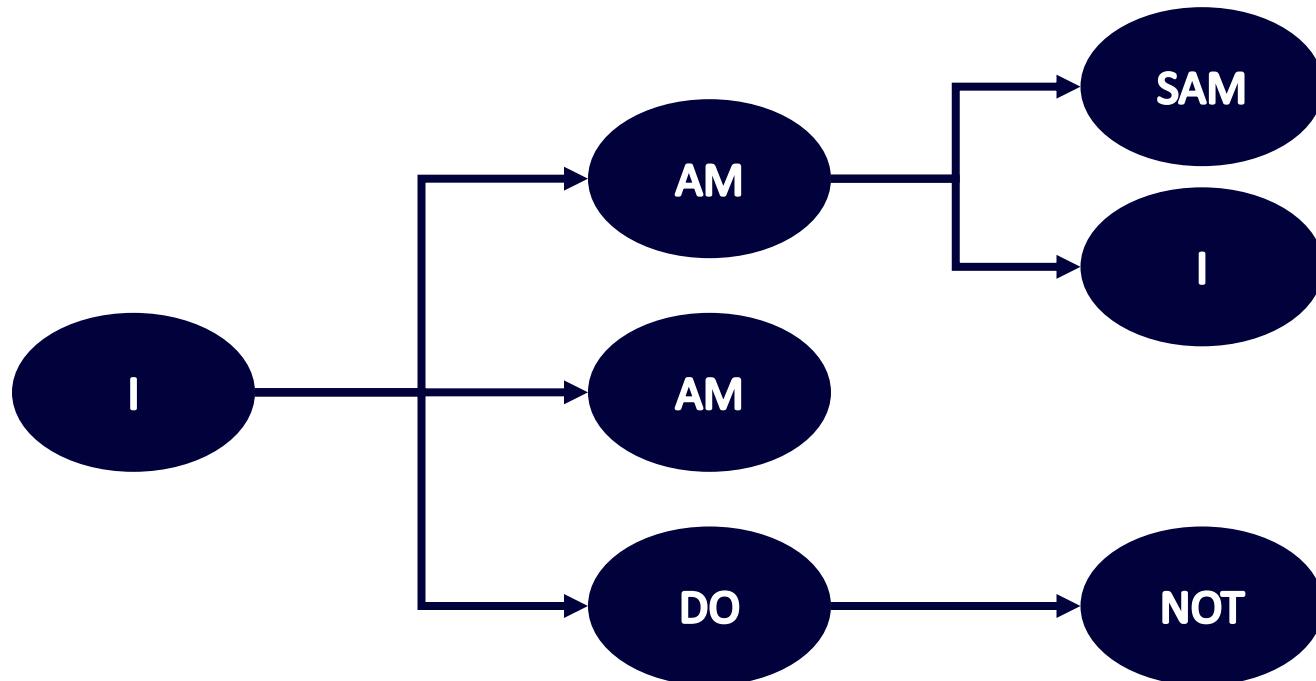
## S O U R C E S

Reddit API for recent posts

pushshift.io API for historical data

# A LITTLE ABOUT MARKOV CHAINS

I AM SAM. SAM I AM.. I DO NOT LIKE GREEN EGGS AND HAM..



# MODEL SELECTION AND PERFORMANCE

## SELECTION

Tested the following models

- Logistic Regression
- Naïve Bayes
- Random Forest
- AdaBoost
- Gradient boosting regression
- Support Vector Machine

with cross validation and grid searching over hyperparameters.

## PERFORMANCE

Logistic Regression performed the best, achieving **98%** accuracy on training data and **94%** accuracy on unseen data.

The final parameters used Lasso Regularization with a regularization strength of **0.5**

Vectorization was achieved using CountVectorizer with RegEx tokenization.

# MODEL PREDICTION EXAMPLES

What is the sexiest sex that sexes your sex???  
TIFU by putting toothpaste on my watch, bucko.

*u/circlejerk\_SS*

**53.9% chance this was written by a bot**

**3.61% chance this was written by a bot**

Florida Man injects himself with a machete  
and a plane and failed epically at both

*u/FloridaMan\_SS*

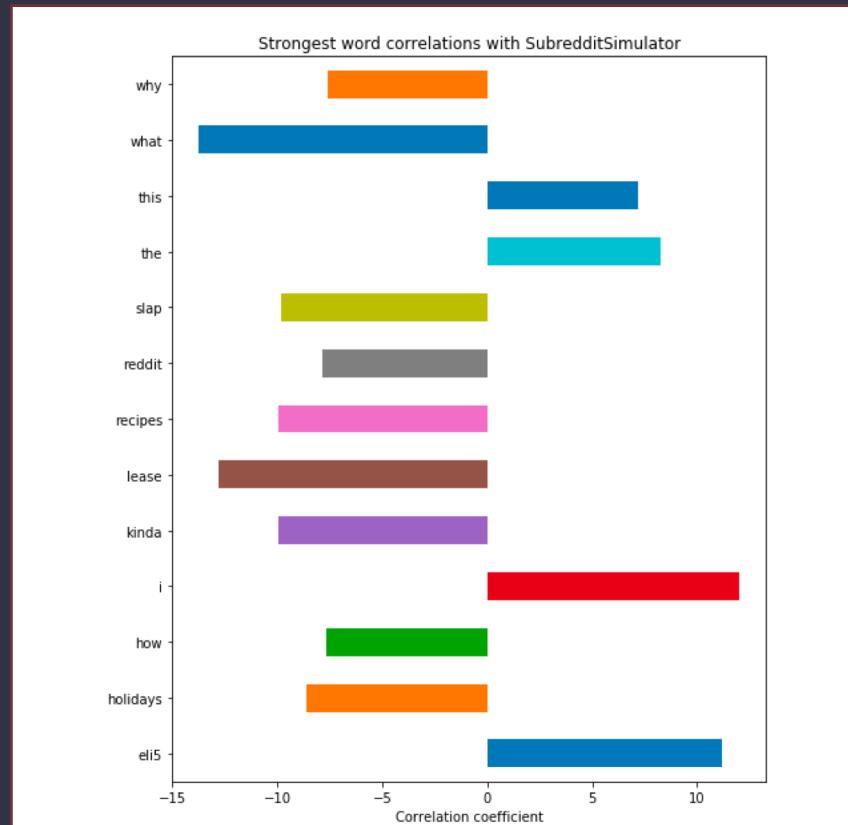
What is something that happened in history,  
that if it happened in a movie, people would  
call "plot hole"?

*u/Tortured\_-soul*

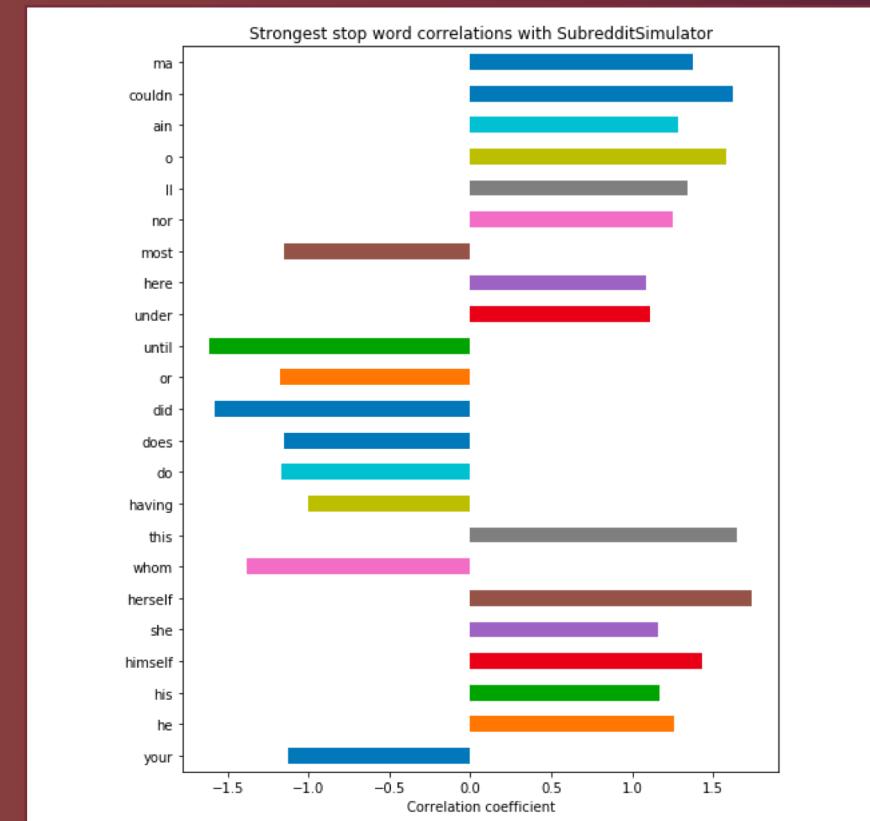
**99.9% chance this was written by a bot**

# CORRELATIONS WITH *SubredditSimulator*

## WORD CORRELATIONS



## STOPWORD CORRELATIONS



## SubredditSimulator Word Cloud

A word cloud visualization for the SubredditSimulator dataset. The words are arranged in a cluster, with larger fonts indicating higher frequency. Key words include: home, also, human, suresays, bad, place, job, put, gets, men, black, kids, next, hard, word, advice, around, asked, girl, still, many, play, country, last, song, old, school, left, two, days, post, working, going, years, actually, see, done, anyone, worst, believe, without, much, feel, real, way, anything, everyone, use, new, favorite, think, else, life, make, time, long, made, free, day, could, video, let, story, person, someone, looking, thing, give, serious, one, ever, something, best, experience, great, history, people, cat, things, good, would, friend, work, know, get, tell, never, like, happened, movie, found, say, game, lot, friends, looks, really, redditors, every, seen, today, need, year, right, want, moment, back, show, thought, kid, trump, question, better, first, world, used, man, look, ask, always, need, year, right, want, moment, back, show, thought, kid, trump, question, better, first, world, used, week, look, ask, always, need, year, right, want, moment, back, show, thought, kid, trump, question, better, first, world, used, dog, mrw, keep, head, find, wanted, money, little, change, house, love, live, even, getting, family, take, name, making, got, games, come, well, week, look, ask, always, need, year, right, want, moment, back, show, thought, kid, trump, question, better, first, world, used, dog, mrw, keep, head, find, wanted, money, little, change, house, love, live, even, getting, family, take, name, making, got, games, come, well.

## AskReddit Word Cloud

A word cloud visualization for the AskReddit dataset. The words are arranged in a cluster, with larger fonts indicating higher frequency. Key words include: day, good, see, every, way, tell, like, best, think, person, actually, new, find, never, post, say, job, serious, know, thing, would, made, people, something, trump, back, make, ever, time, worst, movie, game, life, someone, feel, first, work, going, want, much, take, year, years, story, need, got.

# WORD CLOUDS

## Observations:

- AskReddit had higher frequencies of fewer key words.
- Many of the highest frequency words are shared between the two subreddits.

## Filters:

- Word frequency of 5 or greater
- Excluded some common words ('reddit', 'redditor', 'removed', 'deleted')

## CONCLUSION

The very high accuracy of a logistic regression model implies that we can quite easily distinguish between these two subreddits using only bag of words methods. Upon reflection, perhaps `AskReddit` isn't as good of a stand-in for a general topic forum as hoped.

Given more time, I would instead take a sampling of each individual subreddit that `SubredditSimulator` pulls from, to get a more equivalent corpus of words.

