
# Reddit Comment Text Analysis

## MSCA 31013 Final Project
## Team 1

Caixin Yang, Yue Wu, Chuyu Chen, Aman Gupta, Yizhou Zhang

# Agenda

1. Business Problem
2. Exploratory Data Analysis & Preparation
3. Machine Learning Models
4. Sentiment Analysis
5. Regression Analysis
6. Path Forward


# Business Problem

# Executive Summary

To help Reddit understand their topics, categorize comments attitude, and predict comments' likes and dislike scores.
We would like to conduct text analysis to better understand the current Reddit community through the subreddits.

1. **Target popular topics** using word clouds.
2. **Categorize the attitude** based on the comments.
3. **Predict scores** for each comment accordingly.

# Business Problem

## Sentiment Analysis

- Understand people's opinions from a post
- Potentially help Reddit gain an overview of the wider public opinion behind certain topics

## Regression Analysis

- Based on the body of the post, predict a post’s success before it’s submitted
- Potentially help Redditors gain upvotes, and predict which posts will get popular enough to hit the front page

# Challenges

## Data Size

- 8GB data with 21 variables
- Storage

## Computation Time

- NLP
- Regression
- Classification

## Memory

- Preprocessing
- Featurization

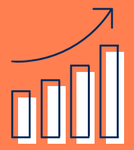

# EDA & Data Preparations

# Data Overview

A small portion of publicly available reddit comment datasets
-- comments posted in May 2015

Breaking an enormous dataset to explore

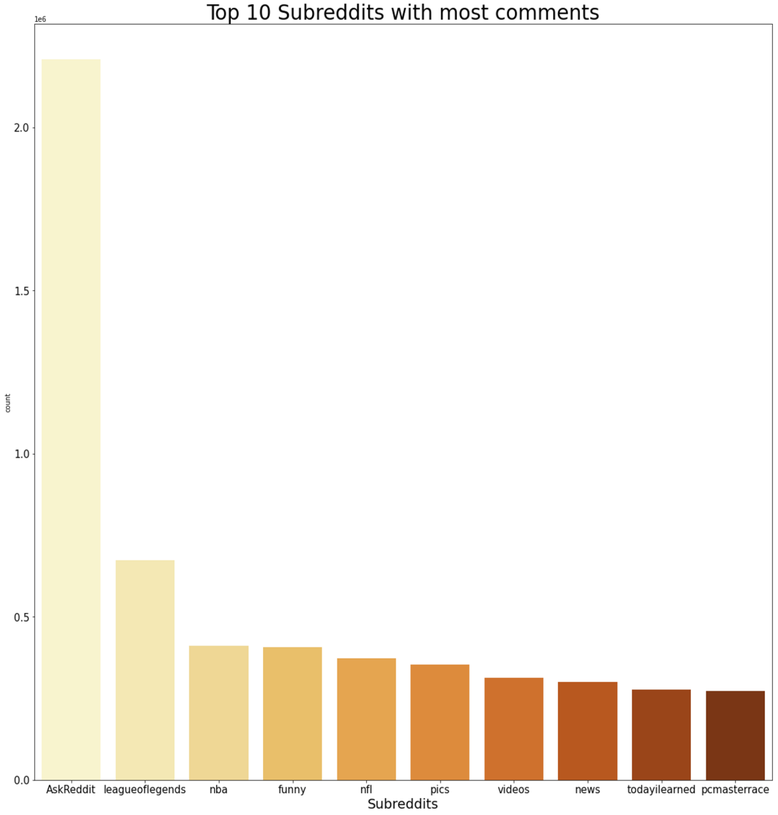
- Original Dataset released by Reddit: 1+ TB with 1.7 billion comments
- Practical subset: 8G with ~27 million comments
- cleaned comment dataset: 1.5+ GB with ~4 million comments

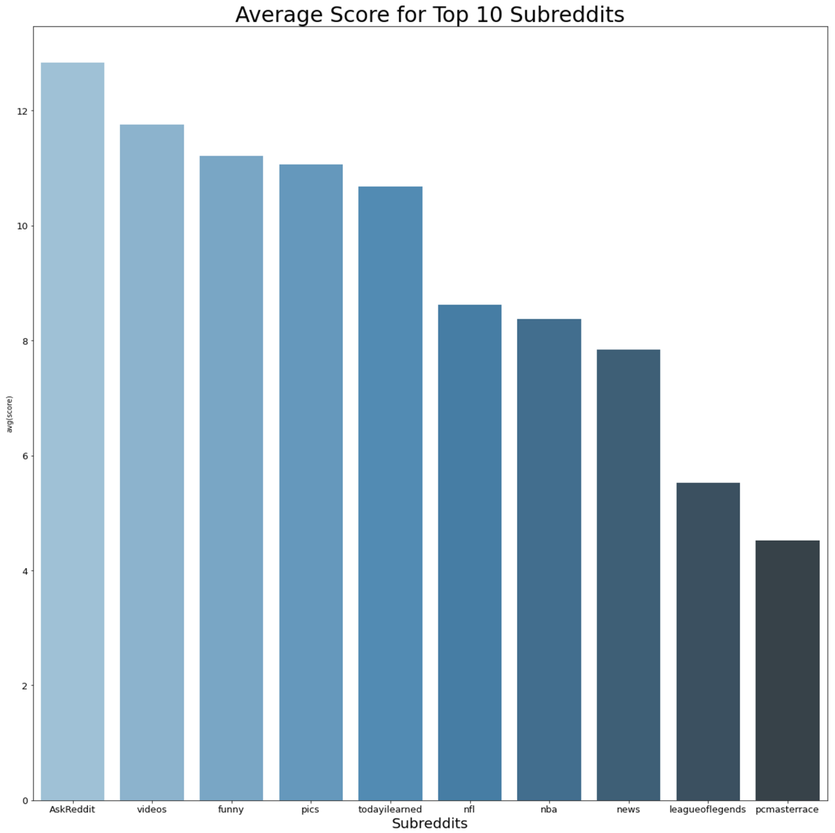


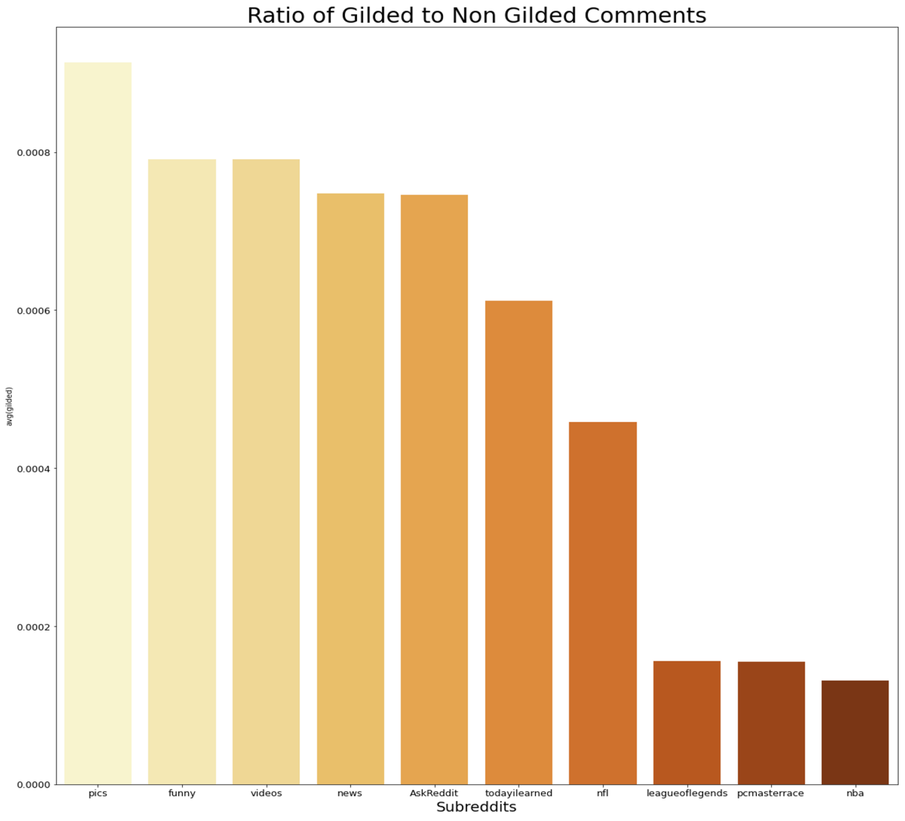
An aggregated table with 21 variables and 30K records.

- Body
- subreddit
- score
- ups
- downs
- gilded

# Exploratory Data Analysis




**Calculated Top 10 subreddits based on number of comments**

**AskReddit was by far the most popular subreddit**

**These 10 reddits were used for the rest of the project**

**Calculated Average score of Top 10 subreddits**

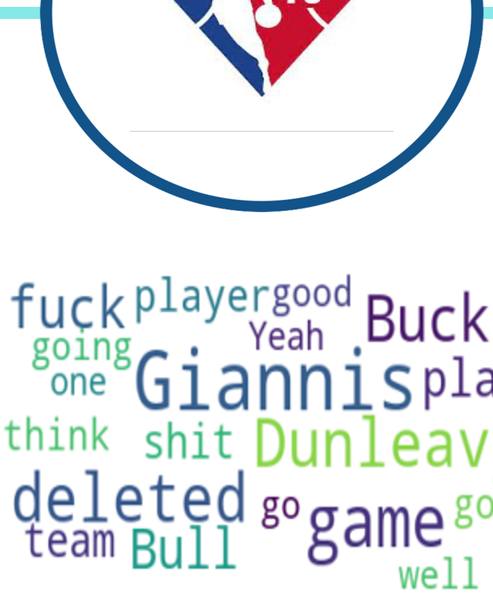
**Score = Ups - Downs**

**This helped us understand the overall positivity/negativity of comments in the subreddits**

**Gilding refers to awarding comments with Reddit Coins, a virtual currency**

**This helped us understand which subreddits had the greatest change of getting these rewards**

# Word Cloud Observations from Top 10 Subreddits




- Aggressive Language
- Frequent usage of Curse words
- Giannis was a popular figure


- More aggressive than NBA
- Frequent usage of Curse words, negative words


- Esports fans are less agressive
- Frequent reference to in game purchases and skins


- Frequent mention of police/cops
- Usage of opinion words such as think and probably

# Data Storage

| Data Storage Options | Upload Procedure | Pros | Cons |
|---|---|---|---|
| Google Cloud Platforms | Create project, bucket for storage, data proc cluster etc | More computing power, easier to run ML models | Complicated set up procedure compared to RCC |
| RCC | Create relevant directory using account, upload data using HDFS | Easy to set up, direct support from RCC to troubleshoot | Unreliable, often slow when running ML models |

# Data Preparation

| | body | clean_comment | category |
|---|---|---|---|
| 1 | gg this one's over. off to watch the NFL draft I guess | this one over off watch the nfl draft guess | 0 |
| 3 | No one has a European accent either because it doesn't exist. There are accents from Europe but not a European accent. | one has european accent either because doesn exist there are accents from europe but not european accent | 0 |
| 4 | That the kid "..reminds me of Kevin." so sad :-( | that the kid reminds kevin sad | -1 |
| 19 | NSFL | nsfl | 0 |
| 22 | Get back to your pott harry. | get back your pott harry | 0 |

## Drop Duplicates

Dropped duplicate rows
Shape of data: (28883304, 22) ->> (25622841, 22)

## Reduced Dataset

Filtered columns only include top 10 subreddits
['AskReddit','leagueoflegends','nba','funny','nfl','pics','videos','news','todayilearned','pcmasterrace']

## Text Cleaning

Including the removal of HTML special entities, tickers, hyperlinks,hashtags,punctuation, whitespace in front of or at the end of the text, new line characters, converting @ to AT_USER, change all words to lowercase, etc.

## Sentiment Labeling

Create a category column for labeling the sentiment of a comment using TextBlob in Python. +1 indicates a positive comment, 0 indicates neutral, -1 indicates a negative comment.


# Sentiment Analysis

# Model Pipeline

**Extra text cleaning:**
- Removes punctuation
- Changes to lowercase
- Trim white spaces

**Categorical label -1, 0, 1 for negative, neutral, positive reviews**

**Regular Expression Tokenizer**

**StopWords Remover**

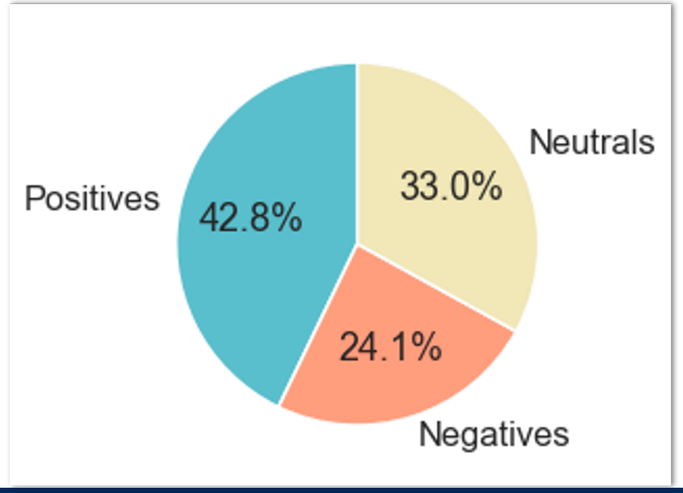
**CountVectorizer**

**StringIndexer**

| body | clean_comment | category | cleaned | words | filtered | features | label |
|---|---|---|---|---|---|---|---|
| gg this one's ove... | gg this one's ove... | 0 | gg this ones over... | [gg, this, ones, ... | [gg, this, ones, ... | (262144,[0,2,14,1... | 1.0 |
| No one has a Euro... | No one has a Euro... | 0 | no one has a euro... | [no, one, has, a,... | [no, one, has, a,... | (262144,[1,7,12,1... | 1.0 |
| That the kid "..r... | That the kid "..r... | -1 | that the kid remi... | [that, the, kid, ... | [that, kid, remin... | (262144,[4,5,28,3... | 2.0 |
| NSFL | NSFL | 0 | nsfl | [nsfl] | [nsfl] | (262144,[13710],[... | 1.0 |
| Get back to your ... | Get back to your ... | 0 | get back to your ... | [get, back, to, y... | [get, back, to, y... | (262144,[0,31,44,... | 1.0 |

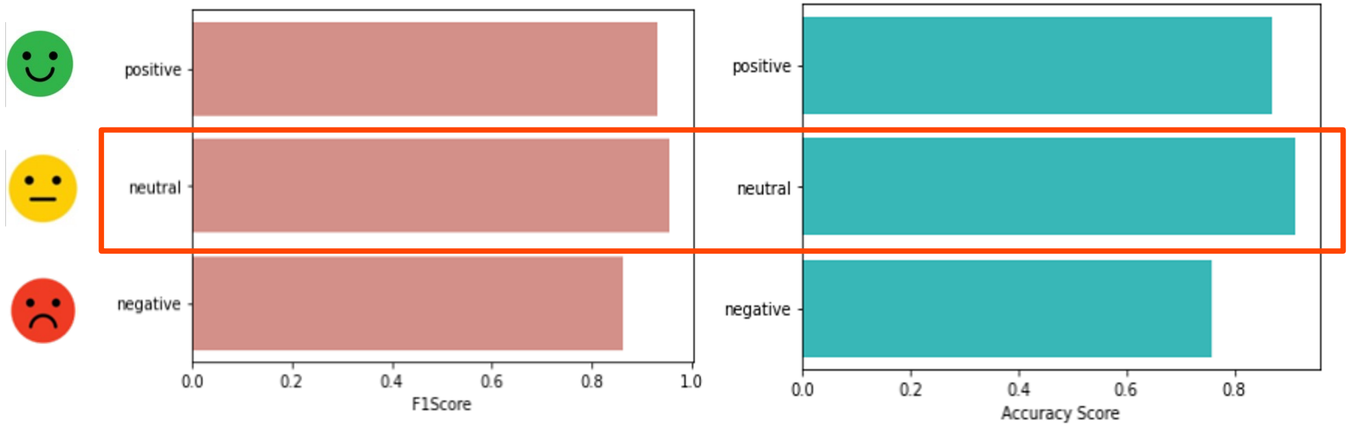
| filtered | rawFeatures | features | label |
|---|---|---|---|
| [gg, this, ones, ... | (10000,[35,352,48... | (10000,[35,352,48... | 1.0 |
| [no, one, has, a,... | (10000,[66,431,76... | (10000,[66,431,76... | 1.0 |
| [that, kid, remin... | (10000,[3723,4495... | (10000,[3723,4495... | 2.0 |
| [nsfl] | (10000,[7579],[1.0]) | (10000,[7579],[6.... | 1.0 |
| [get, back, to, y... | (10000,[488,1263,... | (10000,[488,1263,... | 1.0 |

**HashingTF** **IDF**

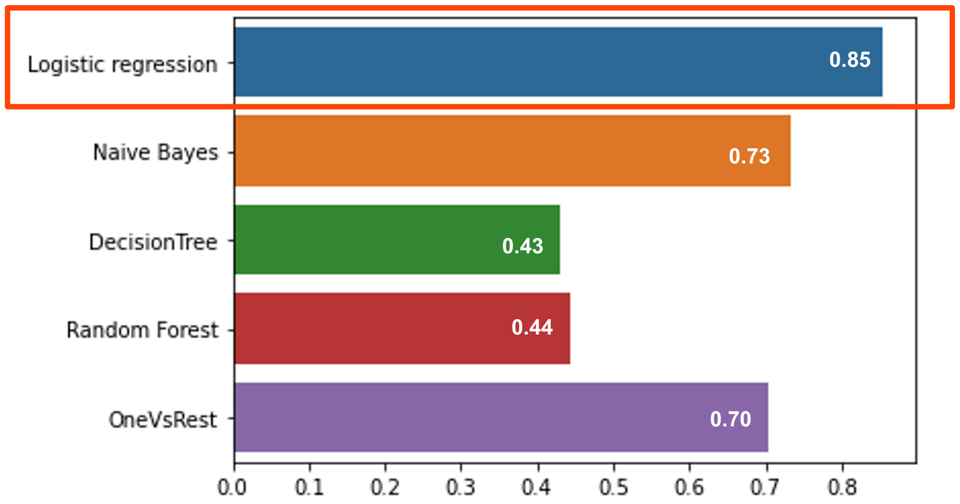



**Positive comments: 2083852**
**Neutral comments: 1606259**
**Negative comments: 1174577**

# Validation : Sentiment Analysis



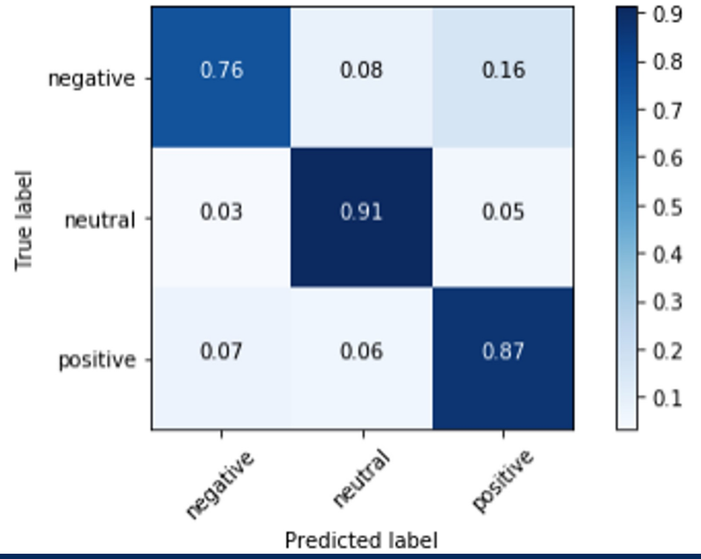

## 1. Model Validation - Accuracy




Logistic Regression: 85%

## 2. Logistic Regression F1 & Accuracy Scores

## 3. Logistic Regression Confusion Matrix




# Regression Analysis

# Model Flow

Data Cleaning

NLP Pipeline

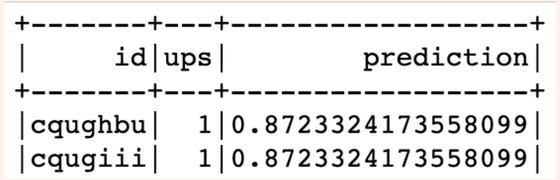
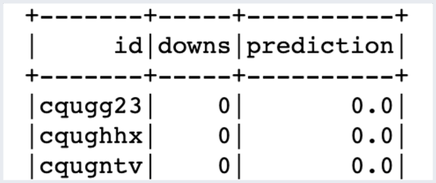
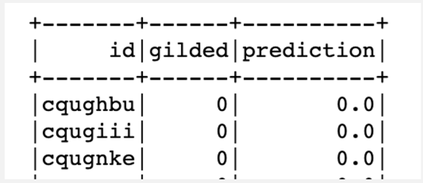
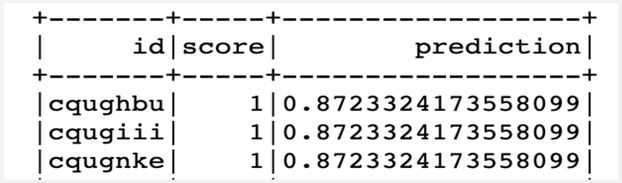
Random Forest

| | Regular Expression Tokenizer | Stop Words Remover | Featurization (CountVectorizer) |
|---|---|---|---|

| clean_comment | words | filtered_words | features |
|---|---|---|---|
| kinda the reason ... | [kinda, the, reas... | [kinda, reason, d... | (7082,[31,54,55,1... |
| believe whatever ... | [believe, whateve... | [believe, whateve... | (7082,[19,22,38,4... |
| this isn the fuck... | [this, isn, the, ... | [isn, fucking, pa... | (7082,[53,58,71,8... |
| lack serious you ... | [lack, serious, y... | [lack, serious, h... | (7082,[280,281,38... |
| entirely the faul... | [entirely, the, f... | [entirely, fault,... | (7082,[0,1,6,15,2... |
| love how bothered... | [love, how, bothe... | [love, bothered, ... | (7082,[0,1,4,22,2... |
| what have you bee... | [what, have, you,... | [christianity, ph... | (7082,[17,67,116,... |
| does living unhea... | [does, living, un... | [living, unhealth... | (7082,[115,168,29... |
| ambiguous pronoun... | [ambiguous, prono... | [ambiguous, prono... | (7082,[121,1525],... |
| someone wrote boo... | [someone, wrote, ... | [someone, wrote, ... | (7082,[4,13,49,11... |
| don understand th... | [don, understand,... | [understand, resp... | (7082,[1,7,8,14,4... |
| still believes hi... | [still, believes,... | [still, believes,... | (7082,[35,142,198... |
| that fucking stup... | [that, fucking, s... | [fucking, stupid,... | (7082,[86,222],[1... |

Training: 3 million records
Testing: 768k records

# Model Performance

| Label | Metrics | Result Snapshot | |
|---|---|---|---|
| Ups | RMSE: 126% |  | The original dataset is very imbalance with most comments got 1 upvote |
| Downs | RMSE: 0% |  | In the original dataset, very few comments got downvote and most were zero |
| Gilded | RMSE: 0% |  | In the original dataset, there was very small number of gilded comments and most were zero |
| Score | RMSE: 126% |  | The original dataset is very imbalance with most comments got a score of 1 |

- **SMOTE**
- **Model Tuning**
- **Other Modeling Techniques**

# Comparison of Model Performance

| Regression Algorithm | RMSE score |
|---|---|
| Logistic Regression | 28.37 |
| Linear Regression | 28.87 |
| Decision Tree | 26.07 |
| Gradient-Boosted Regression Tree | 26.09 |

# Perhaps Next Step to Improve Prediction Score...

nba

```
[('the', 216), ('and', 80), ('that', 74), ('you', 48), ('this', 46), ('game', 40), ('was', 3
9), ('for', 39), ('just', 32), ('like', 31), ('but', 30), ('not', 27), ('his', 27), ('have',
25), ('can', 25), ('they', 24), ('him', 23), ('with', 22), ('are', 22), ('when', 21), ('tea
m', 18), ('good', 16), ('think', 15), ('some', 15), ('lebron', 15), ('would', 15), ('what', 1
4), ('curry', 14), ('has', 14), ('don', 14), ('out', 14), ('year', 13), ('one', 13), ('been',
13), ('back', 13), ('better', 13), ('only', 13), ('see', 12), ('them', 12), ('how', 12), ('mo
re', 12), ('could', 11), ('harden', 11), ('get', 11), ('from', 11), ('play', 11), ('lol', 1
0), ('got', 10), ('really', 10), ('their', 10)]
```

funny

```
[('the', 150), ('you', 76), ('and', 69), ('that', 67), ('this', 38), ('but', 33), ('are', 3
2), ('just', 31), ('for', 29), ('was', 27), ('like', 25), ('have', 25), ('not', 22), ('can',
20), ('all', 20), ('about', 18), ('what', 18), ('don', 17), ('people', 17), ('with', 17), ('t
hink', 16), ('your', 15), ('from', 15), ('them', 15), ('really', 14), ('they', 14), ('would',
13), ('see', 13), ('didn', 12), ('get', 11), ('his', 10), ('same', 10), ('way', 10), ('when',
10), ('had', 10), ('who', 10), ('out', 10), ('funny', 9), ('actually', 9), ('has', 9), ('ther
e', 9), ('know', 8), ('why', 8), ('how', 8), ('stuff', 8), ('here', 8), ('some', 8), ('thin
g', 8), ('their', 8), ('did', 7)]
```

```
'and': 0.8266785731844679,
'the': 0.9267620317414504,
'you': 0.9808292530117262,
'that': 1.037987666851675,
'way': 1.037987666851675,
'are': 1.037987666851675,
'for': 1.037987666851675,
'not': 1.037987666851675,
'but': 1.037987666851675,
'really': 1.0986122886681098,
```

```
'fired': 2.4849066497880004,
'talent': 2.4849066497880004,
'release': 2.4849066497880004,
'protect': 2.4849066497880004,
'math': 2.4849066497880004,
'correct': 2.4849066497880004,
'science': 2.4849066497880004,
'card': 2.4849066497880004,
'tons': 2.4849066497880004,
'compare': 2.4849066497880004,
'bank': 2.4849066497880004,
```

# Thank you!

# Appendix

# Text Cleaning in Python:

In [35]:
```
# helper function to clean comments
def processComment(comment):
    # Remove HTML special entities (e.g. &amp;)
    comment = re.sub(r'\&\w*;', '', str(comment))
    #Convert @username to AT_USER
    comment = re.sub('@[^\s]+','', comment)
    # Remove tickers
    comment = re.sub(r'\$\w*', '', comment)
    # To lowercase
    comment = comment.lower()
    # Remove hyperlinks
    comment = re.sub(r'https?:\/\/.*\/\w*', '', comment)
    # Remove hashtags
    comment = re.sub(r'#\w*', '', comment)
    # Remove Punctuation and split 's, 't, 've with a space for filter
    comment = re.sub(r'[' + punctuation.replace('@', '') + ']+', ' ', comment)
    # Remove words with 2 or fewer letters
    comment = re.sub(r'\b\w{1,2}\b', '', comment)
    # Remove whitespace (including new line characters)
    comment = re.sub(r'\s\s+', ' ', comment)
    # Remove single space remaining at the front of the comment.
    comment = tweet.lstrip('')
    # Remove characters beyond Basic Multilingual Plane (BMP) of Unicode:
    comment = ''.join(c for c in comment if c <= '\uffff')
    return comment
#

df1['clean_comment'] = df1['body'].apply(processComment)
df1[['body','clean_comment']].head()
```

Out[35]:

| | body | clean_comment |
|---|---|---|
| **1** | gg this one's over. off to watch the NFL draft I guess | this one over off watch the nfl draft guess |
| **3** | No one has a European accent either because it doesn't exist. There are accents from Europe but not a European accent. | one has european accent either because doesn exist there are accents from europe but not european accent |
| **4** | That the kid "..reminds me of Kevin." so sad :-( | that the kid reminds kevin sad |
| **19** | NSFL | nsfl |
| **22** | Get back to your pott harry. | get back your pott harry |