For Reddit Grammaticality Analysis

Camryn Simons

Background

- The internet's influence on speech and writing
- Abbreviations
  - "Tix" instead of "Tickets"
- Acronyms
  - "LOL" meaning "Laugh Out Loud"
- Use of punctuation
  - "I'm so excited for the summer!!!!!!!!"
- Capitalization
  - "i don'T WANT TO GO TO CLASS"
- Sentence structure
  - "going to bed"
- Emojis
  - 😄

BTW

TTYL

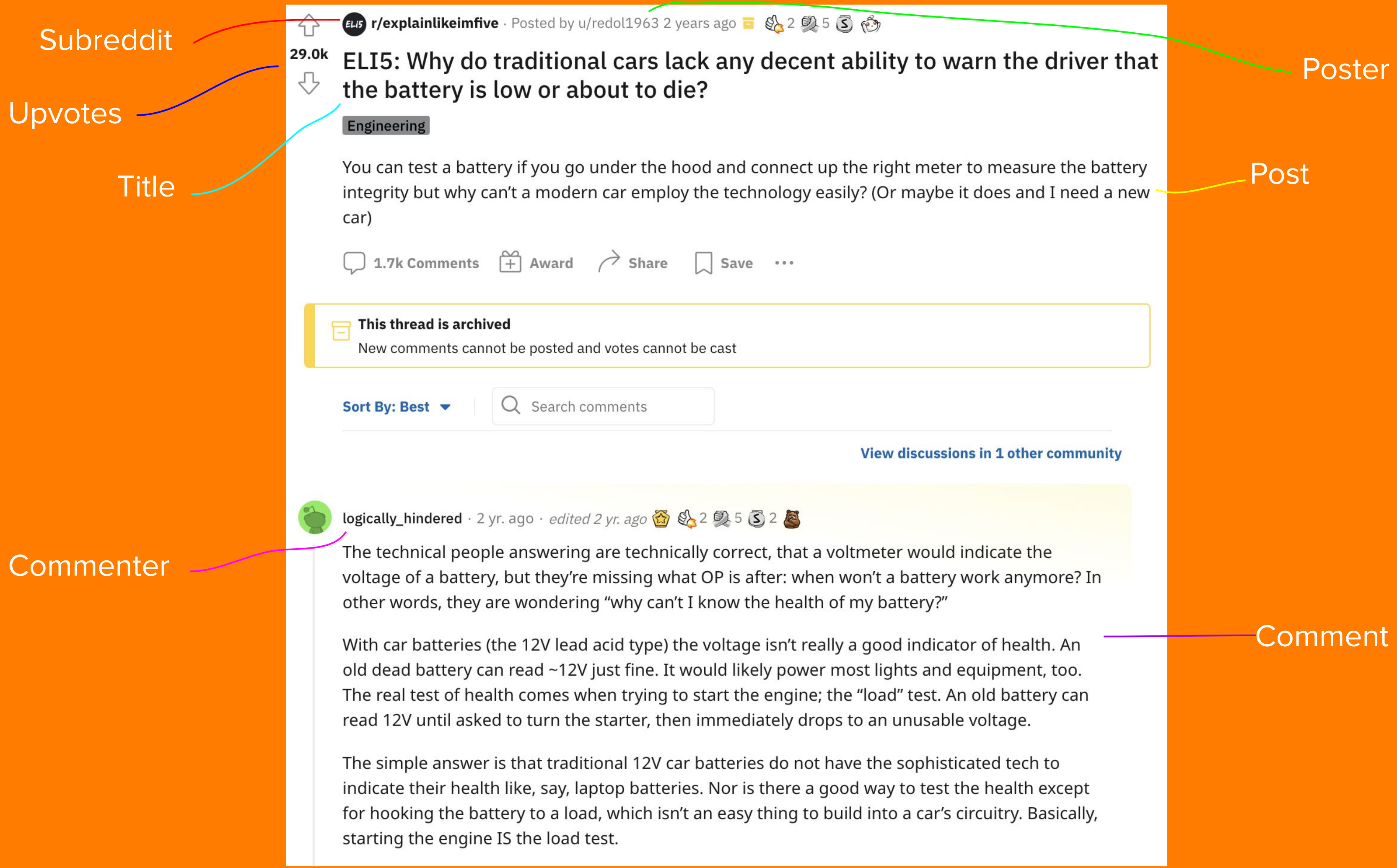
NVM

Background - Formalized

- Internet linguistics
  - Study of how language has been changed by the Internet
  - Four lenses: sociolinguistic, education, stylistic, applied linguistics
- Internet’s impact often seen in a negative light
  - Bad grammar
  - Bad spelling
- *Because Internet: Understanding the New Rules of Language,* Gretchen McCulloch
  - Language is an “open-source project”
  - Changes shouldn’t be seen as negative
  - There is no right way to use a language

Reddit

- Website and mobile application
- Composed of “Subreddits”
  - Forums for specific topics
  - Syntax: ‘r/[topic]’
- People can create, respond to, and react to posts and comments
- Posts and comments can be short or lengthy



r/explainlikeimfive · Posted by u/redol1963 2 years ago 2 5

29.0k

ELI5: Why do traditional cars lack any decent ability to warn the driver that the battery is low or about to die?

Engineering

You can test a battery if you go under the hood and connect up the right meter to measure the battery integrity but why can't a modern car employ the technology easily? (Or maybe it does and I need a new car)

1.7k Comments

This thread is archived
New comments cannot be posted and votes cannot be cast

Sort By: Best

Search comments

View discussions in 1 other community

logically_hindered · 2 yr. ago · *edited 2 yr. ago* 2 5 2

The technical people answering are technically correct, that a voltmeter would indicate the voltage of a battery, but they're missing what OP is after: when won't a battery work anymore? In other words, they are wondering "why can't I know the health of my battery?"

With car batteries (the 12V lead acid type) the voltage isn't really a good indicator of health. An old dead battery can read ~12V just fine. It would likely power most lights and equipment, too. The real test of health comes when trying to start the engine; the "load" test. An old battery can read 12V until asked to turn the starter, then immediately drops to an unusable voltage.

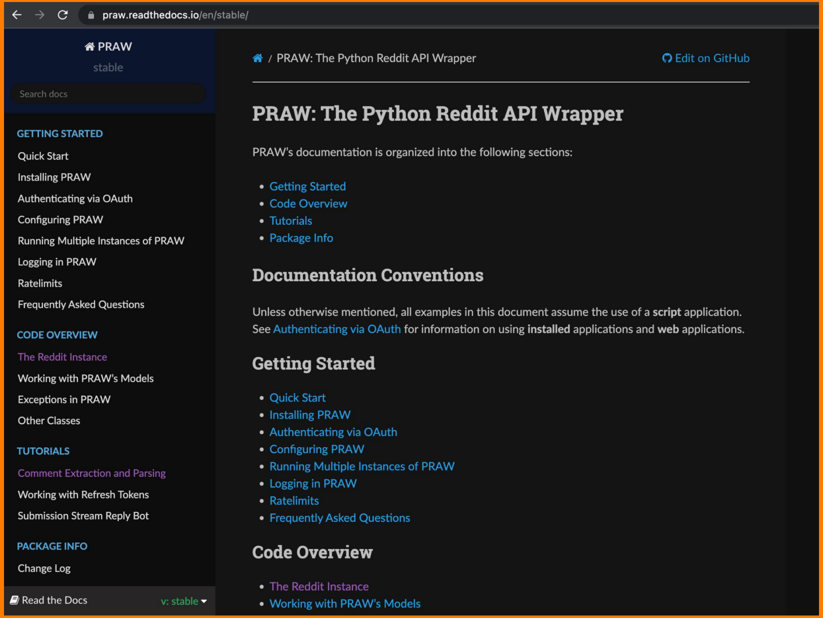
The simple answer is that traditional 12V car batteries do not have the sophisticated tech to indicate their health like, say, laptop batteries. Nor is there a good way to test the health except for hooking the battery to a load, which isn't an easy thing to build into a car's circuitry. Basically, starting the engine IS the load test.

Motivations

- **Overall Focus**: Grammaticality on Reddit
- **Question**: Is there a way to categorize common grammatical errors?
  - **Subquestion**: Which grammatical errors are most prevalent on Reddit?
  - **Subquestion**: Are there grammatical errors that are more common across certain subreddits?
  - **Subquestion**: Does the grammaticality of a post have an effect on its interactions(Comments, upvotes)?

Data Collection

- **Problem:** In order to analyze posts, posts need to be collected
  - Tens of thousands of posts
- **Solution:** PRAW
  - **P**ython **R**eddit **A**PI **W**rapper
  - Easier way of using the Reddit API
  - Variety of functions
    - Collecting posts
    - Collecting comments
    - So much more…



PRAW: The Python Reddit API Wrapper

PRAW's documentation is organized into the following sections:

- Getting Started
- Code Overview
- Tutorials
- Package Info

Documentation Conventions

Unless otherwise mentioned, all examples in this document assume the use of a **script** application. See Authenticating via OAuth for information on using **installed** applications and **web** applications.

Getting Started

- Quick Start
- Installing PRAW
- Authenticating via OAuth
- Configuring PRAW
- Running Multiple Instances of PRAW
- Logging in PRAW
- Ratelimits
- Frequently Asked Questions

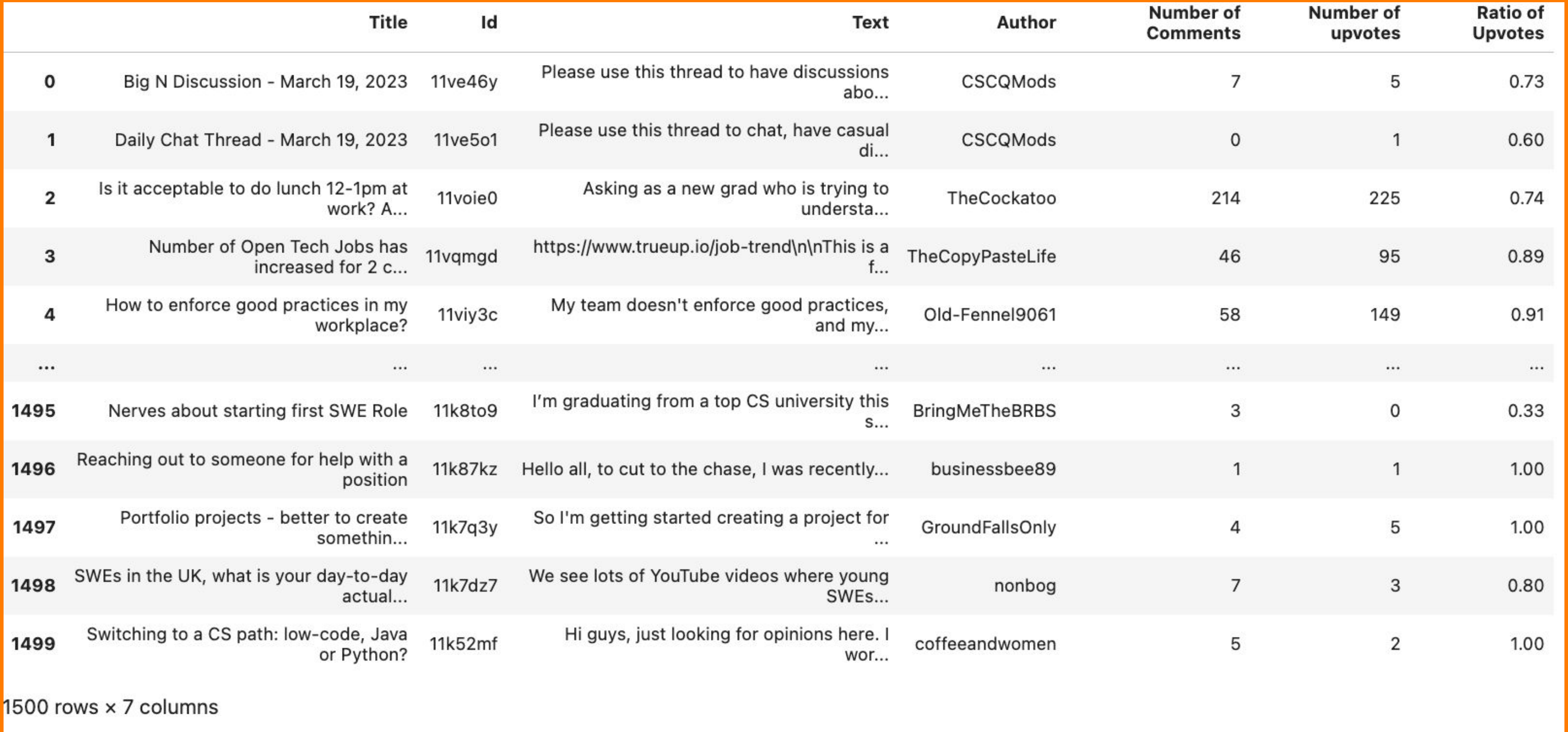
Code Overview

- The Reddit Instance
- Working with PRAW's Models

Data Collection: Limitations

- Actual problems with PRAW itself
  - **Problem:** Would not collect posts past a certain threshold regardless of input
    - **Solution:** Had to request batches of posts at different times of day or different days
- Data collection in general
  - **Problem:** Collecting posts took multiple minutes to run at a time
    - **Solution:** No solution… just patience
  - **Problem:** Large dataset, hard to find potential issues
    - **Solution:** Lots of printing

Data



	Title	Id	Text	Author	Number of Comments	Number of upvotes	Ratio of Upvotes
0	Big N Discussion - March 19, 2023	11ve46y	Please use this thread to have discussions abo...	CSCQMods	7	5	0.73
1	Daily Chat Thread - March 19, 2023	11ve5o1	Please use this thread to chat, have casual di...	CSCQMods	0	1	0.60
2	Is it acceptable to do lunch 12-1pm at work? A...	11voie0	Asking as a new grad who is trying to understa...	TheCockatoo	214	225	0.74
3	Number of Open Tech Jobs has increased for 2 c...	11vqmgd	https://www.trueup.io/job-trend\n\nThis is a f...	TheCopyPasteLife	46	95	0.89
4	How to enforce good practices in my workplace?	11viy3c	My team doesn't enforce good practices, and my...	Old-Fennel9061	58	149	0.91
...	...	...	...	...	...	...	...
1495	Nerves about starting first SWE Role	11k8to9	I'm graduating from a top CS university this s...	BringMeTheBRBS	3	0	0.33
1496	Reaching out to someone for help with a position	11k87kz	Hello all, to cut to the chase, I was recently...	businessbee89	1	1	1.00
1497	Portfolio projects - better to create somethin...	11k7q3y	So I'm getting started creating a project for ...	GroundFallsOnly	4	5	1.00
1498	SWEs in the UK, what is your day-to-day actual...	11k7dz7	We see lots of YouTube videos where young SWEs...	nonbog	7	3	0.80
1499	Switching to a CS path: low-code, Java or Python?	11k52mf	Hi guys, just looking for opinions here. I wor...	coffeeandwomen	5	2	1.00

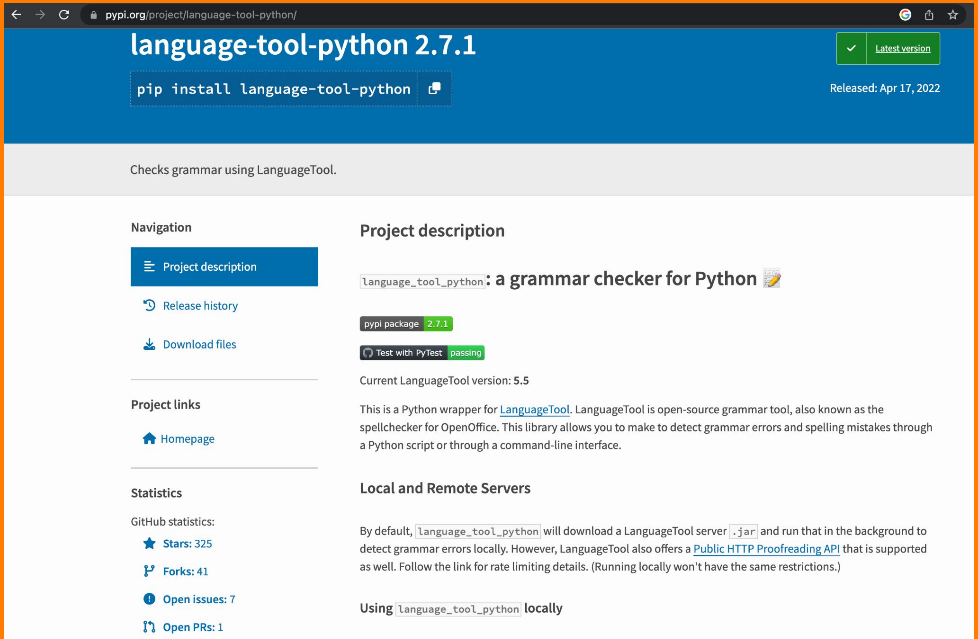
1500 rows × 7 columns

Analysis

- **Problem:** How do we determine grammaticality?
  - **Solution:** Analyzing every single post manually
    - Infeasible
    - Potentially subjective
  - **Solution:** Creating my own tool
    - Daunting task
    - Potentially subjective
  - **Solution:** Check for existing tools
    - Still potentially has issues
    - Takes the bulk of the issue away

Language-Tool-Python

- Python wrapper for a grammar tool
- Takes in input of a string
- If there are no issues:
  - Outputs nothing
- If there are issues:
  - Detailed output with explanation



language-tool-python 2.7.1

```
pip install language-tool-python
```

Released: Apr 17, 2022

Checks grammar using LanguageTool.

Statistics

GitHub statistics:

- Stars: 325
- Forks: 41
- Open issues: 7
- Open PRs: 1

Project description

`language_tool_python`: a grammar checker for Python 📝

Current LanguageTool version: **5.5**

This is a Python wrapper for LanguageTool. LanguageTool is open-source grammar tool, also known as the spellchecker for OpenOffice. This library allows you to make to detect grammar errors and spelling mistakes through a Python script or through a command-line interface.

Local and Remote Servers

By default, `language_tool_python` will download a LanguageTool server `.jar` and run that in the background to detect grammar errors locally. However, LanguageTool also offers a Public HTTP Proofreading API that is supported as well. Follow the link for rate limiting details. (Running locally won't have the same restrictions.)

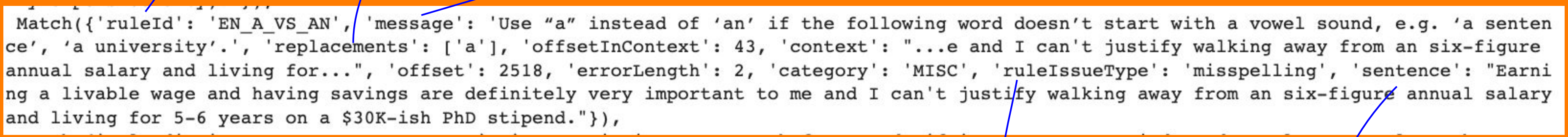
Using `language_tool_python` locally

Example Output

Replacements to fix the issue

The rule that was broken

Detailed explanation



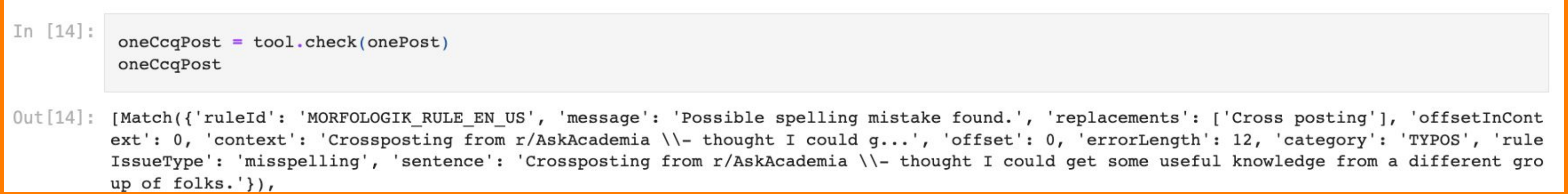
```
 Match({'ruleId': 'EN_A_VS_AN', 'message': 'Use “a” instead of ‘an’ if the following word doesn’t start with a vowel sound, e.g. ‘a senten
ce’, ‘a university’.', 'replacements': ['a'], 'offsetInContext': 43, 'context': "...e and I can't justify walking away from an six-figure
annual salary and living for...", 'offset': 2518, 'errorLength': 2, 'category': 'MISC', 'ruleIssueType': 'misspelling', 'sentence': "Earni
ng a livable wage and having savings are definitely very important to me and I can't justify walking away from an six-figure annual salary
and living for 5-6 years on a $30K-ish PhD stipend."}),
```

Type of rule

Sentence in which it occurs

Top Error

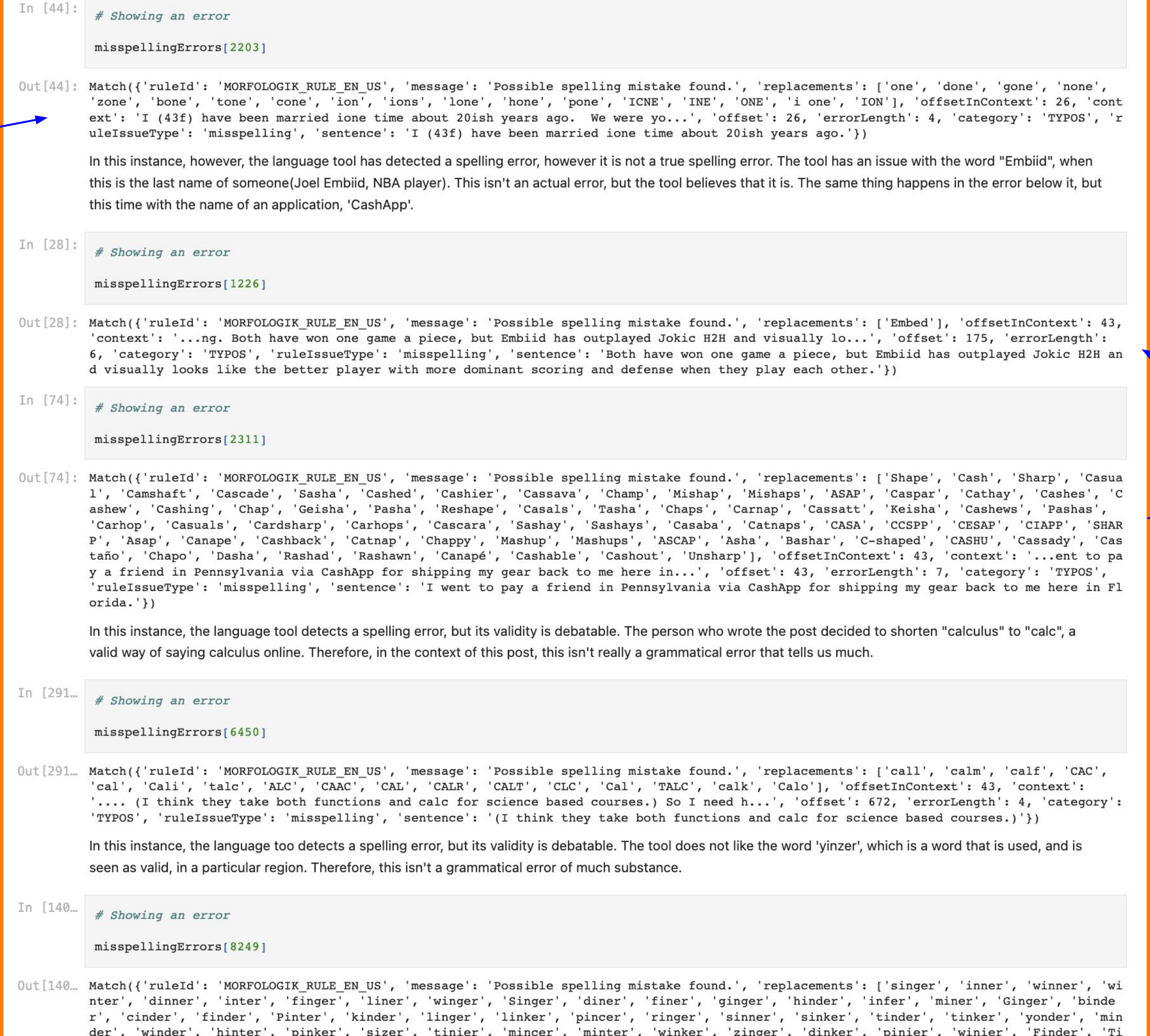
- **MORFOLOGIK_RULE_EN_US** in 12 out of 15 subreddits
  - 2nd top error in 3/15 subreddits
- What does this error mean?
  - Potential spelling error found
    - Key word: **Potential**
- There are several issues with this….



```
In [14]:  oneCcqPost = tool.check(onePost)
          oneCcqPost
```

```
Out[14]: [Match({'ruleId': 'MORFOLOGIK_RULE_EN_US', 'message': 'Possible spelling mistake found.', 'replacements': ['Cross posting'], 'offsetInCont
         ext': 0, 'context': 'Crossposting from r/AskAcademia \\- thought I could g...', 'offset': 0, 'errorLength': 12, 'category': 'TYPOS', 'rule
         IssueType': 'misspelling', 'sentence': 'Crossposting from r/AskAcademia \\- thought I could get some useful knowledge from a different gro
         up of folks.'}),
```

Good 😃



```
In [44]: # Showing an error

         misspellingErrors[2203]

Out[44]: Match({'ruleId': 'MORFOLOGIK_RULE_EN_US', 'message': 'Possible spelling mistake found.', 'replacements': ['one', 'done', 'gone', 'none',
         'zone', 'bone', 'tone', 'cone', 'ion', 'ions', 'lone', 'hone', 'pone', 'ICNE', 'INE', 'ONE', 'i one', 'ION'], 'offsetInContext': 26, 'cont
         ext': 'I (43f) have been married ione time about 20ish years ago.  We were yo...', 'offset': 26, 'errorLength': 4, 'category': 'TYPOS', 'r
         uleIssueType': 'misspelling', 'sentence': 'I (43f) have been married ione time about 20ish years ago.'})
```

In this instance, however, the language tool has detected a spelling error, however it is not a true spelling error. The tool has an issue with the word "Embiid", when this is the last name of someone(Joel Embiid, NBA player). This isn't an actual error, but the tool believes that it is. The same thing happens in the error below it, but this time with the name of an application, 'CashApp'.

```
In [28]: # Showing an error

         misspellingErrors[1226]

Out[28]: Match({'ruleId': 'MORFOLOGIK_RULE_EN_US', 'message': 'Possible spelling mistake found.', 'replacements': ['Embed'], 'offsetInContext': 43,
         'context': '...ng. Both have won one game a piece, but Embiid has outplayed Jokic H2H and visually lo...', 'offset': 175, 'errorLength':
         6, 'category': 'TYPOS', 'ruleIssueType': 'misspelling', 'sentence': 'Both have won one game a piece, but Embiid has outplayed Jokic H2H an
         d visually looks like the better player with more dominant scoring and defense when they play each other.'})

In [74]: # Showing an error

         misspellingErrors[2311]

Out[74]: Match({'ruleId': 'MORFOLOGIK_RULE_EN_US', 'message': 'Possible spelling mistake found.', 'replacements': ['Shape', 'Cash', 'Sharp', 'Casua
         l', 'Camshaft', 'Cascade', 'Sasha', 'Cashed', 'Cashier', 'Cassava', 'Champ', 'Mishap', 'Mishaps', 'ASAP', 'Caspar', 'Cathay', 'Cashes', 'C
         ashew', 'Cashing', 'Chap', 'Geisha', 'Pasha', 'Reshape', 'Casals', 'Tasha', 'Chaps', 'Carnap', 'Cassatt', 'Keisha', 'Cashews', 'Pashas',
         'Carhop', 'Casuals', 'Cardsharp', 'Carhops', 'Cascara', 'Sashay', 'Sashays', 'Casaba', 'Catnaps', 'CASA', 'CCSPP', 'CESAP', 'CIAPP', 'SHAR
         P', 'Asap', 'Canape', 'Cashback', 'Catnap', 'Chappy', 'Mashup', 'Mashups', 'ASCAP', 'Asha', 'Bashar', 'C-shaped', 'CASHU', 'Cassady', 'Cas
         taño', 'Chapo', 'Dasha', 'Rashad', 'Rashawn', 'Canapé', 'Cashable', 'Cashout', 'Unsharp'], 'offsetInContext': 43, 'context': '...ent to pa
         y a friend in Pennsylvania via CashApp for shipping my gear back to me here in...', 'offset': 43, 'errorLength': 7, 'category': 'TYPOS',
         'ruleIssueType': 'misspelling', 'sentence': 'I went to pay a friend in Pennsylvania via CashApp for shipping my gear back to me here in Fl
         orida.'})
```

In this instance, the language tool detects a spelling error, but its validity is debatable. The person who wrote the post decided to shorten "calculus" to "calc", a valid way of saying calculus online. Therefore, in the context of this post, this isn't really a grammatical error that tells us much.

```
In [291…  # Showing an error

          misspellingErrors[6450]

Out[291… Match({'ruleId': 'MORFOLOGIK_RULE_EN_US', 'message': 'Possible spelling mistake found.', 'replacements': ['call', 'calm', 'calf', 'CAC',
         'cal', 'Cali', 'talc', 'ALC', 'CAAC', 'CAL', 'CALR', 'CALT', 'CLC', 'Cal', 'TALC', 'calk', 'Calo'], 'offsetInContext': 43, 'context':
         '.... (I think they take both functions and calc for science based courses.) So I need h...', 'offset': 672, 'errorLength': 4, 'category':
         'TYPOS', 'ruleIssueType': 'misspelling', 'sentence': '(I think they take both functions and calc for science based courses.)'})
```

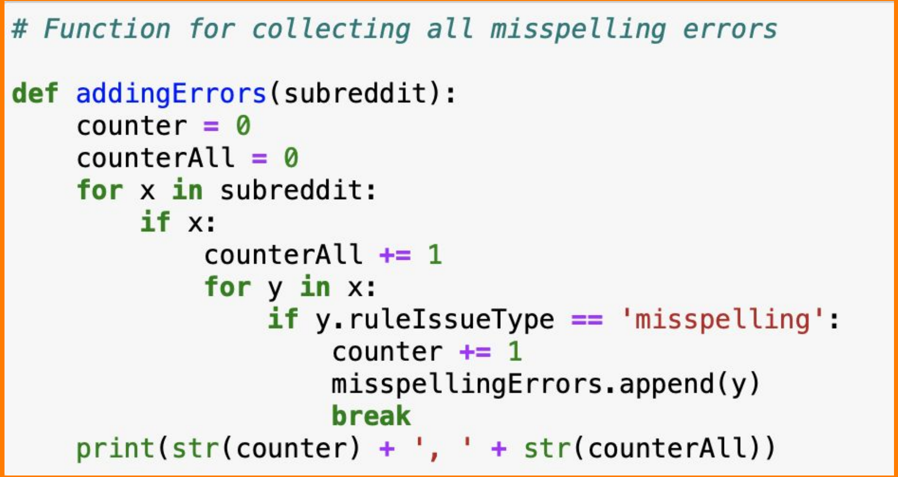
In this instance, the language too detects a spelling error, but its validity is debatable. The tool does not like the word 'yinzer', which is a word that is used, and is seen as valid, in a particular region. Therefore, this isn't a grammatical error of much substance.

```
In [140…  # Showing an error

          misspellingErrors[8249]

Out[140… Match({'ruleId': 'MORFOLOGIK_RULE_EN_US', 'message': 'Possible spelling mistake found.', 'replacements': ['singer', 'inner', 'winner', 'wi
         nter', 'dinner', 'inter', 'finger', 'liner', 'winger', 'Singer', 'diner', 'finer', 'ginger', 'hinder', 'infer', 'miner', 'Ginger', 'binde
         r', 'cinder', 'finder', 'Pinter', 'kinder', 'linger', 'linker', 'pincer', 'ringer', 'sinner', 'sinker', 'tinder', 'tinker', 'yonder', 'min
         der', 'winder', 'hinter', 'pinker', 'sizer', 'tinier', 'mincer', 'minter', 'winker', 'zinger', 'dinker', 'pinier', 'winier', 'Finder', 'Ti
```

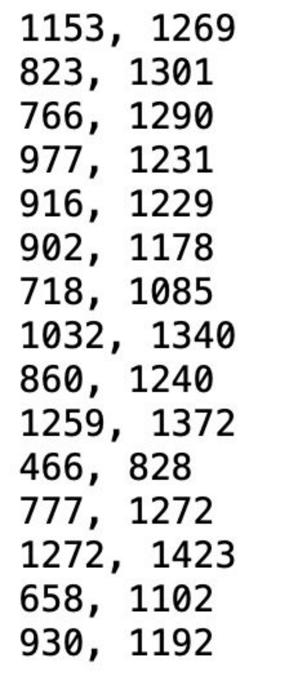
Bad 😢

MORFOLOGIK_RULE_EN_US

- **Problem:** Incorrectly marks certain words as spelling errors
  - Expected behavior with a grammar tool
  - Takes up 73.61% of **ALL** errors
  - Typos don’t reveal as much as other errors do
- **Solution:** Filter them out!



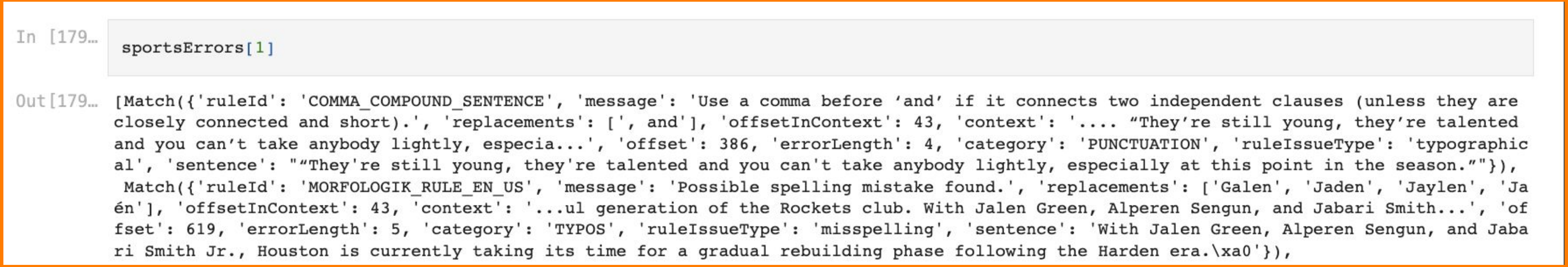
```python
# Function for collecting all misspelling errors

def addingErrors(subreddit):
    counter = 0
    counterAll = 0
    for x in subreddit:
        if x:
            counterAll += 1
            for y in x:
                if y.ruleIssueType == 'misspelling':
                    counter += 1
                    misspellingErrors.append(y)
                    break
    print(str(counter) + ', ' + str(counterAll))
```



```
1153, 1269
823, 1301
766, 1290
977, 1231
916, 1229
902, 1178
718, 1085
1032, 1340
860, 1240
1259, 1372
466, 828
777, 1272
1272, 1423
658, 1102
930, 1192
```

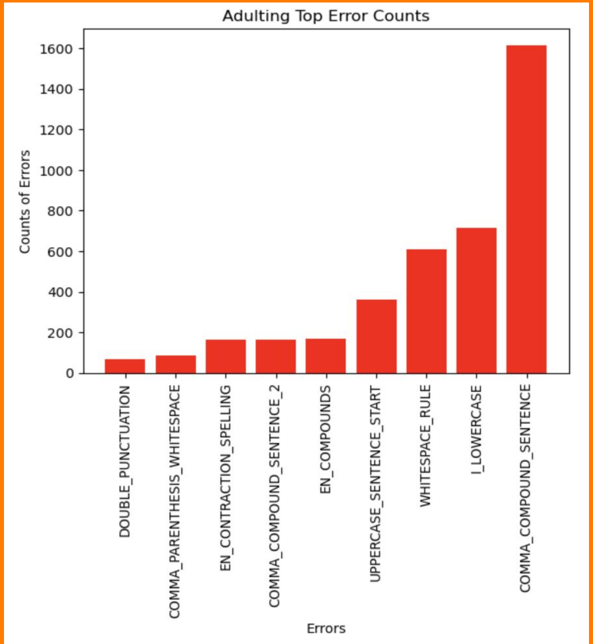
Top Error - Revised

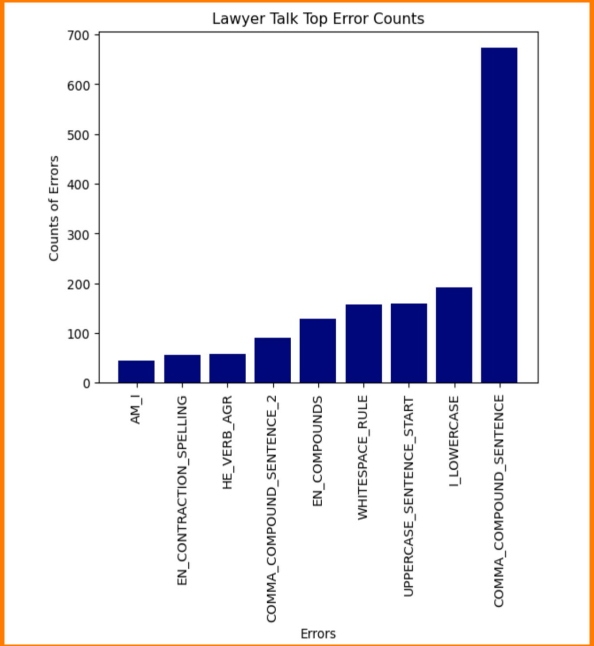
- **COMMA_COMPOUND_SENTENCE** in 11/15 Subreddits
  - 2nd top error in 4/15 subreddits
- What does this error mean?
  - Compound sentence missing a comma
- Most prevalent error by a large margin

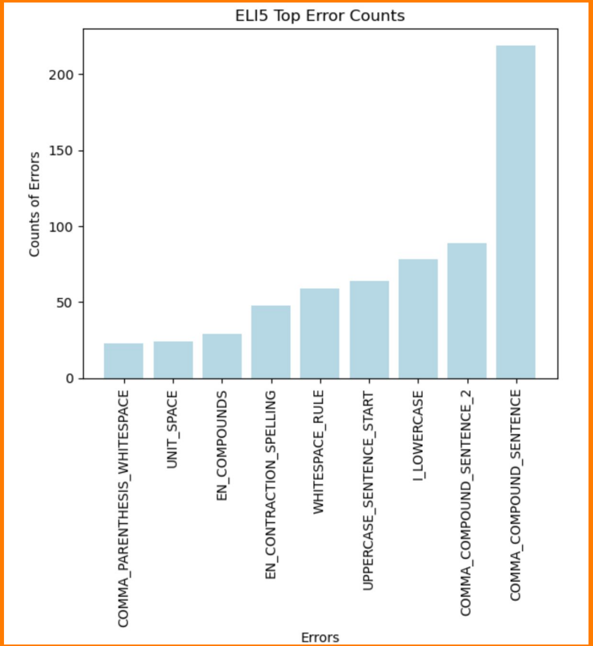


```
In [179…   sportsErrors[1]
```

```
Out[179…  [Match({'ruleId': 'COMMA_COMPOUND_SENTENCE', 'message': 'Use a comma before ‘and’ if it connects two independent clauses (unless they are
          closely connected and short).', 'replacements': [', and'], 'offsetInContext': 43, 'context': '.... “They’re still young, they’re talented
          and you can’t take anybody lightly, especia...', 'offset': 386, 'errorLength': 4, 'category': 'PUNCTUATION', 'ruleIssueType': 'typographic
          al', 'sentence': "“They're still young, they're talented and you can't take anybody lightly, especially at this point in the season.”"}),
           Match({'ruleId': 'MORFOLOGIK_RULE_EN_US', 'message': 'Possible spelling mistake found.', 'replacements': ['Galen', 'Jaden', 'Jaylen', 'Ja
          én'], 'offsetInContext': 43, 'context': '...ul generation of the Rockets club. With Jalen Green, Alperen Sengun, and Jabari Smith...', 'of
          fset': 619, 'errorLength': 5, 'category': 'TYPOS', 'ruleIssueType': 'misspelling', 'sentence': 'With Jalen Green, Alperen Sengun, and Jaba
          ri Smith Jr., Houston is currently taking its time for a gradual rebuilding phase following the Harden era.\xa0'}),
```

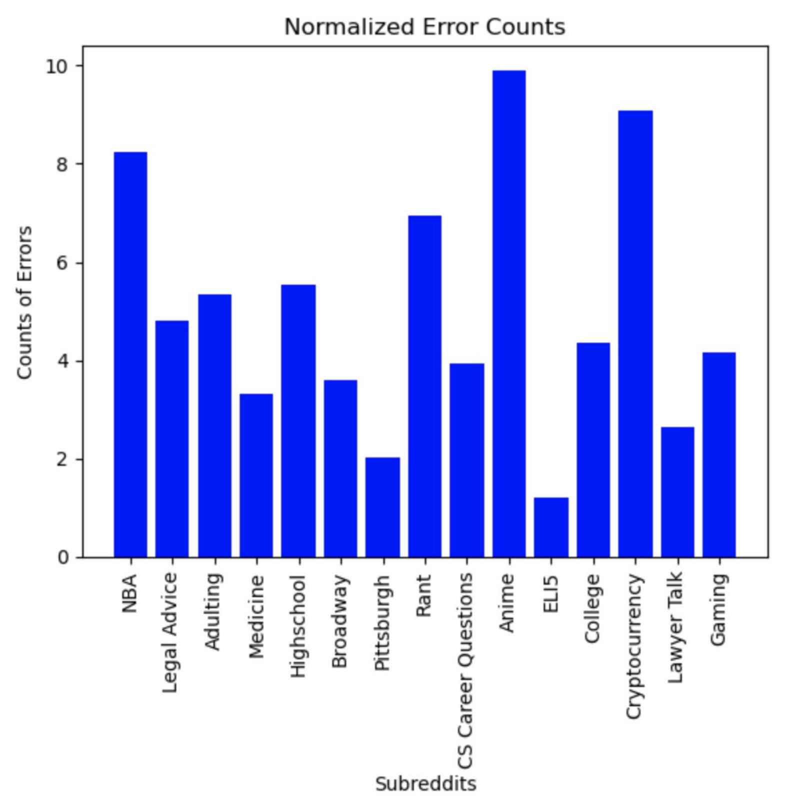

Adulting Top Error Counts
Counts of Errors
0
200
400
600
800
1000
1200
1400
1600
DOUBLE_PUNCTUATION
COMMA_PARENTHESIS_WHITESPACE
EN_CONTRACTION_SPELLING
COMMA_COMPOUND_SENTENCE_2
EN_COMPOUNDS
UPPERCASE_SENTENCE_START
WHITESPACE_RULE
I_LOWERCASE
COMMA_COMPOUND_SENTENCE
Errors


Lawyer Talk Top Error Counts
Counts of Errors
0
100
200
300
400
500
600
700
AM_I
EN_CONTRACTION_SPELLING
HE_VERB_AGR
COMMA_COMPOUND_SENTENCE_2
EN_COMPOUNDS
WHITESPACE_RULE
UPPERCASE_SENTENCE_START
I_LOWERCASE
COMMA_COMPOUND_SENTENCE
Errors


ELI5 Top Error Counts
Counts of Errors
0
50
100
150
200
COMMA_PARENTHESIS_WHITESPACE
UNIT_SPACE
EN_COMPOUNDS
EN_CONTRACTION_SPELLING
WHITESPACE_RULE
UPPERCASE_SENTENCE_START
I_LOWERCASE
COMMA_COMPOUND_SENTENCE_2
COMMA_COMPOUND_SENTENCE
Errors

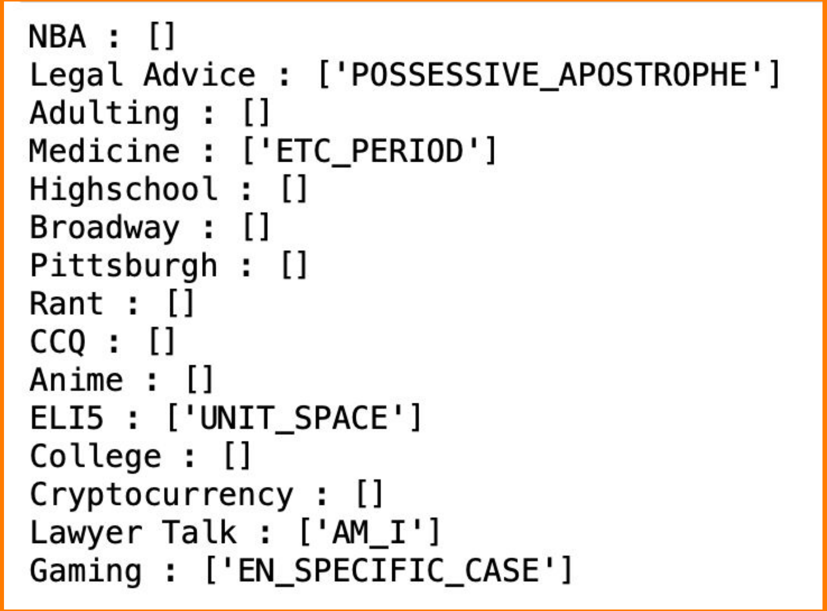
Comparing subreddits - Counts

Average of the number of errors per post in each subreddit



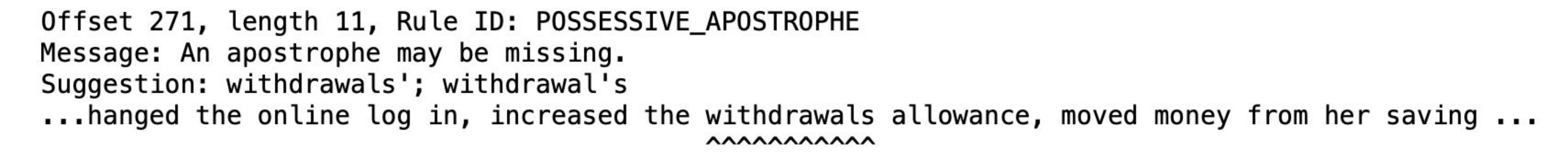
Comparing Subreddits - Errors

- Which top errors are unique to specific subreddits?
  - POSSESSIVE_APOSTROPHE
  - ETC_PERIOD
  - UNIT_SPACE
  - AM_I
  - EN_SPECIFIC CASE
- Demonstrates commonality of errors
  - Lots of overlap in top errors



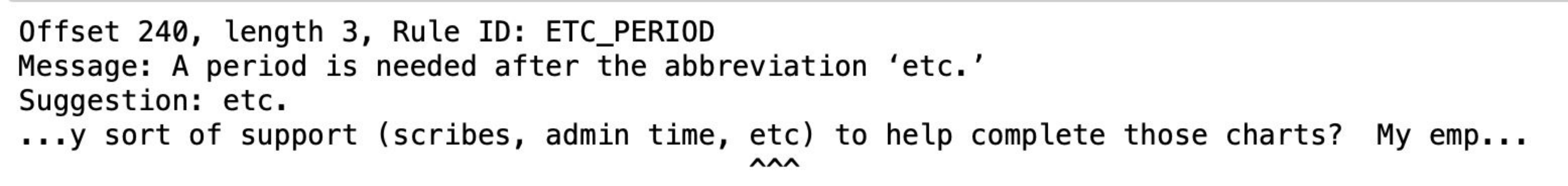
```
NBA : []
Legal Advice : ['POSSESSIVE_APOSTROPHE']
Adulting : []
Medicine : ['ETC_PERIOD']
Highschool : []
Broadway : []
Pittsburgh : []
Rant : []
CCQ : []
Anime : []
ELI5 : ['UNIT_SPACE']
College : []
Cryptocurrency : []
Lawyer Talk : ['AM_I']
Gaming : ['EN_SPECIFIC_CASE']
```

POSSESSIVE APOSTROPHE



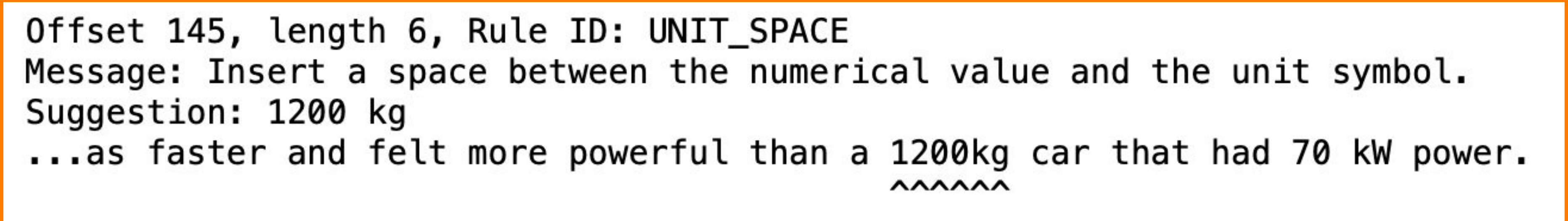
```
Offset 271, length 11, Rule ID: POSSESSIVE_APOSTROPHE
Message: An apostrophe may be missing.
Suggestion: withdrawals'; withdrawal's
...hanged the online log in, increased the withdrawals allowance, moved money from her saving ...
                                           ^^^^^^^^^^^
```

ETC_PERIOD



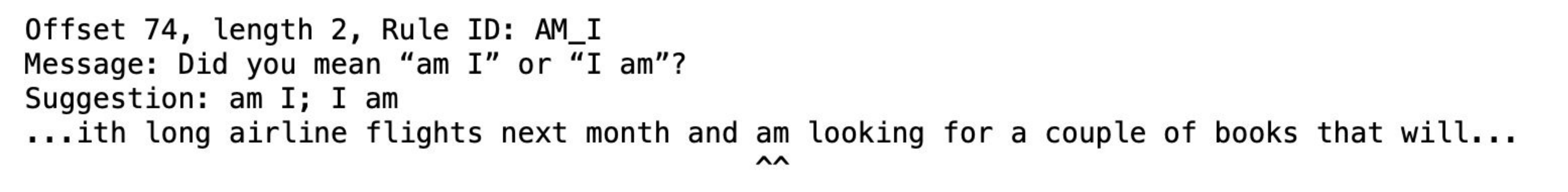
```
Offset 240, length 3, Rule ID: ETC_PERIOD
Message: A period is needed after the abbreviation ‘etc.’
Suggestion: etc.
...y sort of support (scribes, admin time, etc) to help complete those charts?  My emp...
                                           ^^^
```

UNIT_SPACE



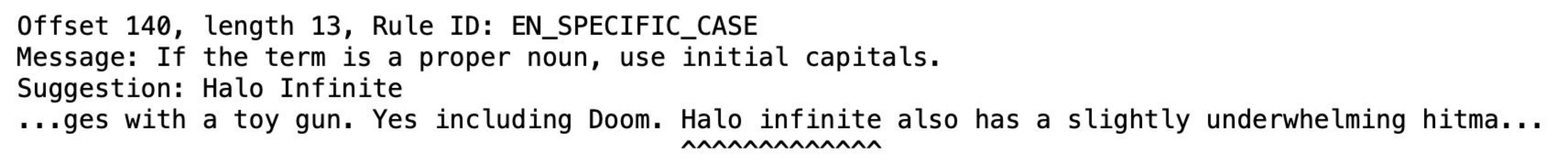
```
Offset 145, length 6, Rule ID: UNIT_SPACE
Message: Insert a space between the numerical value and the unit symbol.
Suggestion: 1200 kg
...as faster and felt more powerful than a 1200kg car that had 70 kW power.
                                           ^^^^^^
```

AM_I



```
Offset 74, length 2, Rule ID: AM_I
Message: Did you mean “am I” or “I am”?
Suggestion: am I; I am
...ith long airline flights next month and am looking for a couple of books that will...
                                           ^^
```

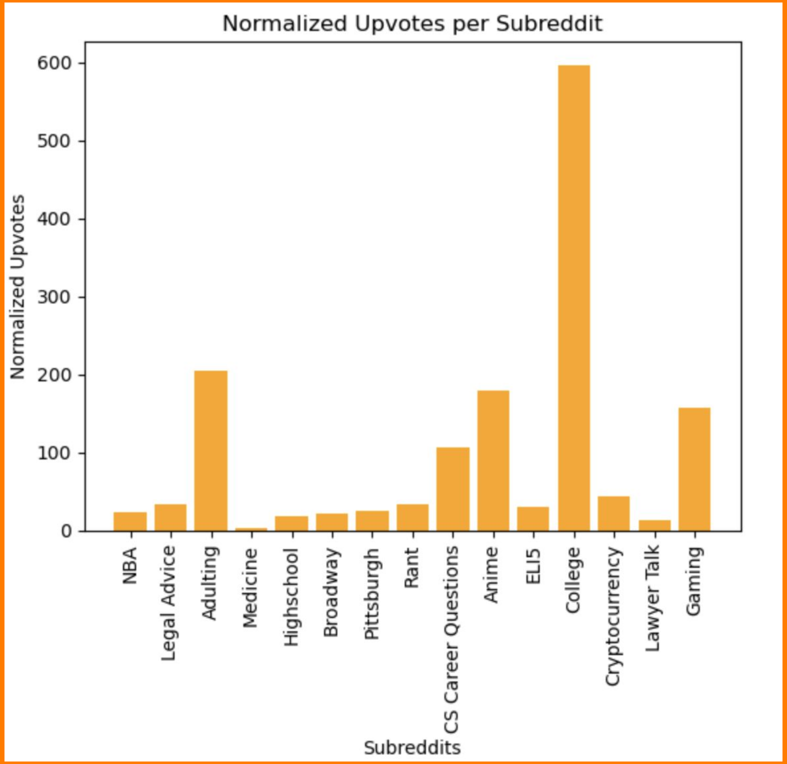
EN_SPECIFIC_CASE

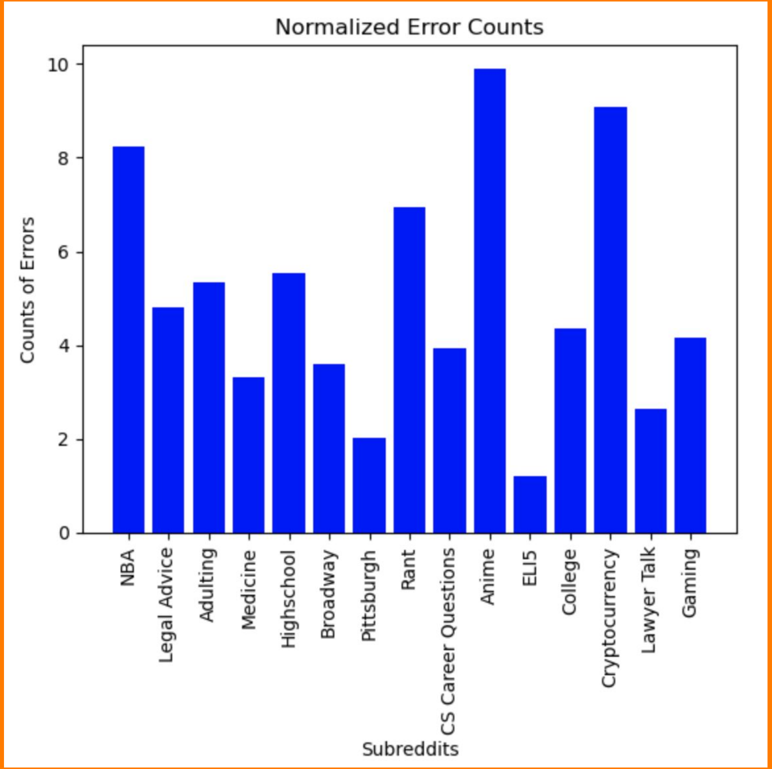


```
Offset 140, length 13, Rule ID: EN_SPECIFIC_CASE
Message: If the term is a proper noun, use initial capitals.
Suggestion: Halo Infinite
...ges with a toy gun. Yes including Doom. Halo infinite also has a slightly underwhelming hitma...
                                           ^^^^^^^^^^^^^
```

Impact on interaction

- Have not fully explored yet
- **Hypothesis:** Grammaticality will not have any impact on interactions
    - **Generalization:** Most people only care about the content of post
        - Typos are often ignored
    - People upvote, comment, and downvote for a variety of reasons
        - Someone could upvote a post because they agree with a post
        - Another person could downvote that post because it is full of grammatical errors


Normalized Upvotes per Subreddit
Normalized Upvotes
0
100
200
300
400
500
600
NBA
Legal Advice
Adulting
Medicine
Highschool
Broadway
Pittsburgh
Rant
CS Career Questions
Anime
ELI5
College
Cryptocurrency
Lawyer Talk
Gaming
Subreddits


Normalized Error Counts
Counts of Errors
0
2
4
6
8
10
NBA
Legal Advice
Adulting
Medicine
Highschool
Broadway
Pittsburgh
Rant
CS Career Questions
Anime
ELI5
College
Cryptocurrency
Lawyer Talk
Gaming
Subreddits