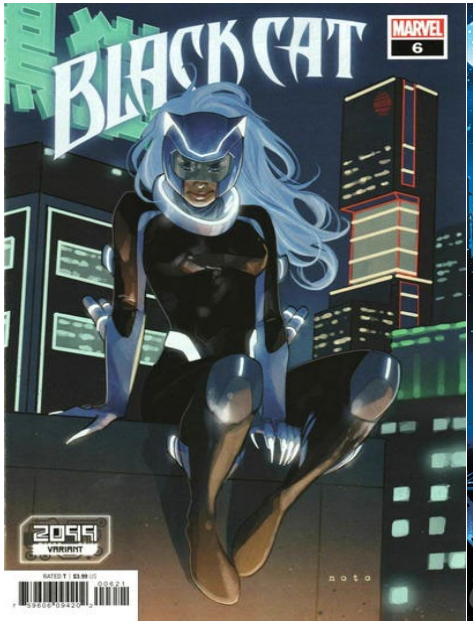
Identifying Marvel vs DC Comics based on Reddit Posts

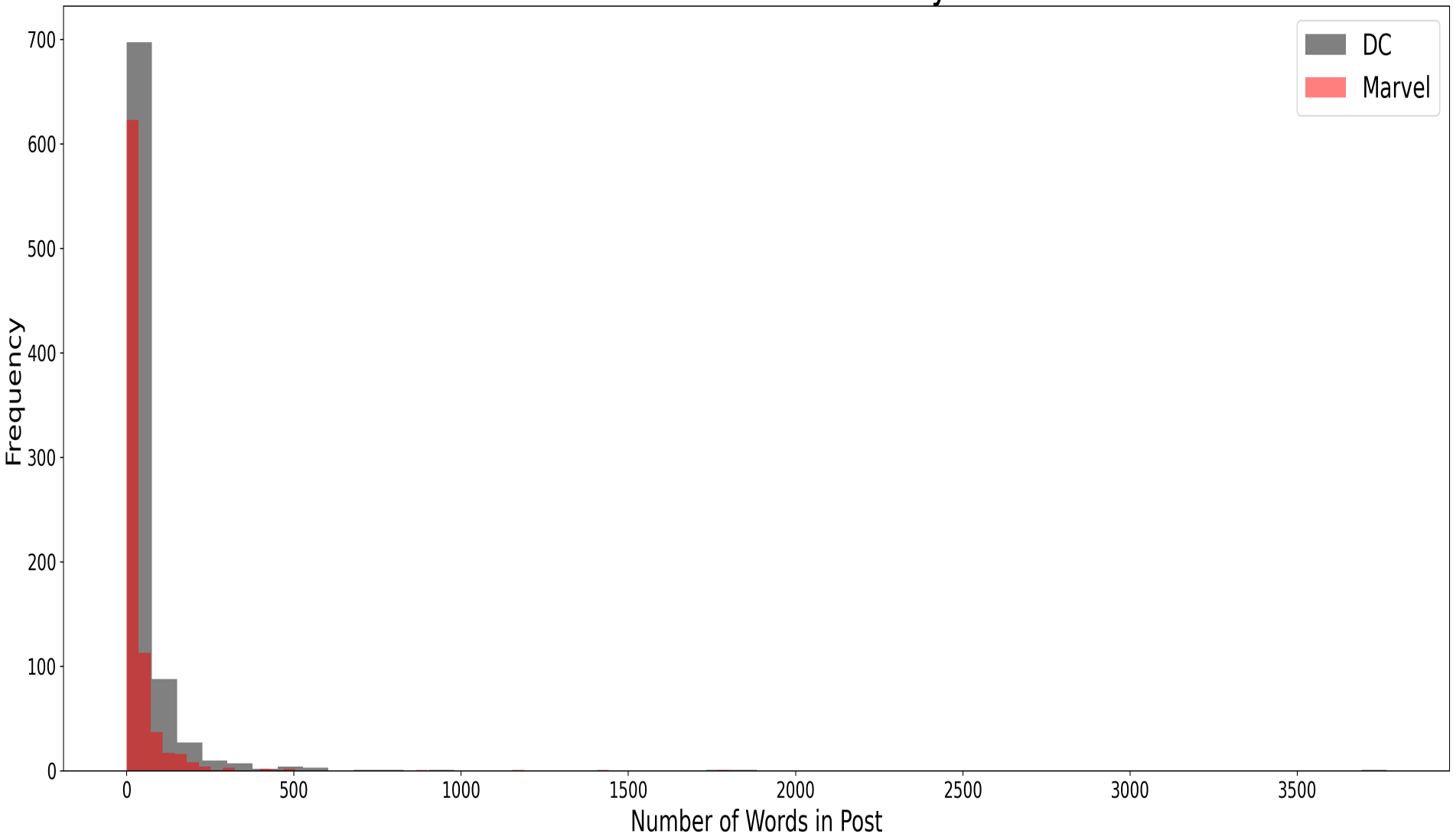
David Kanevsky, DSB 318

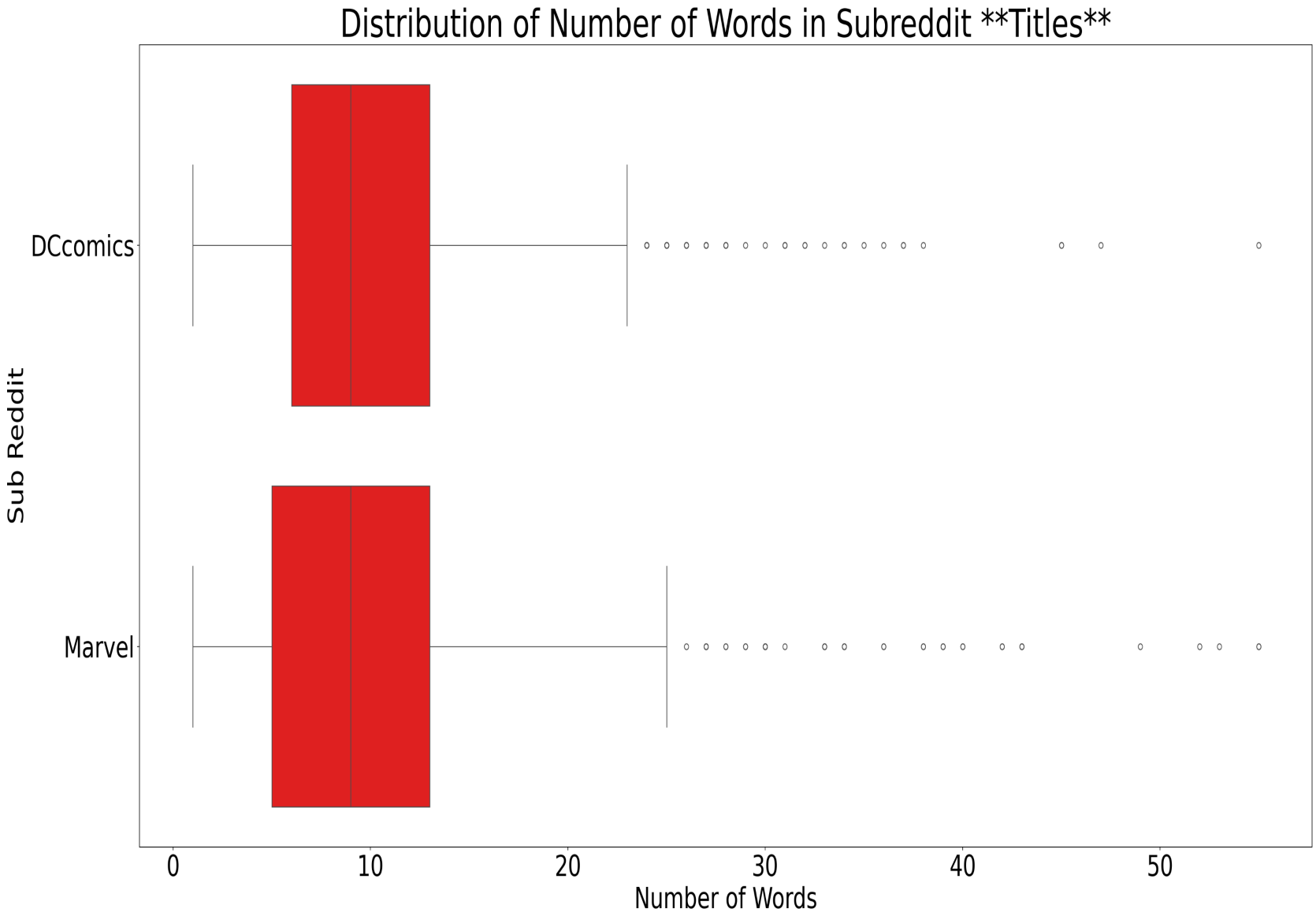
2024-05-03

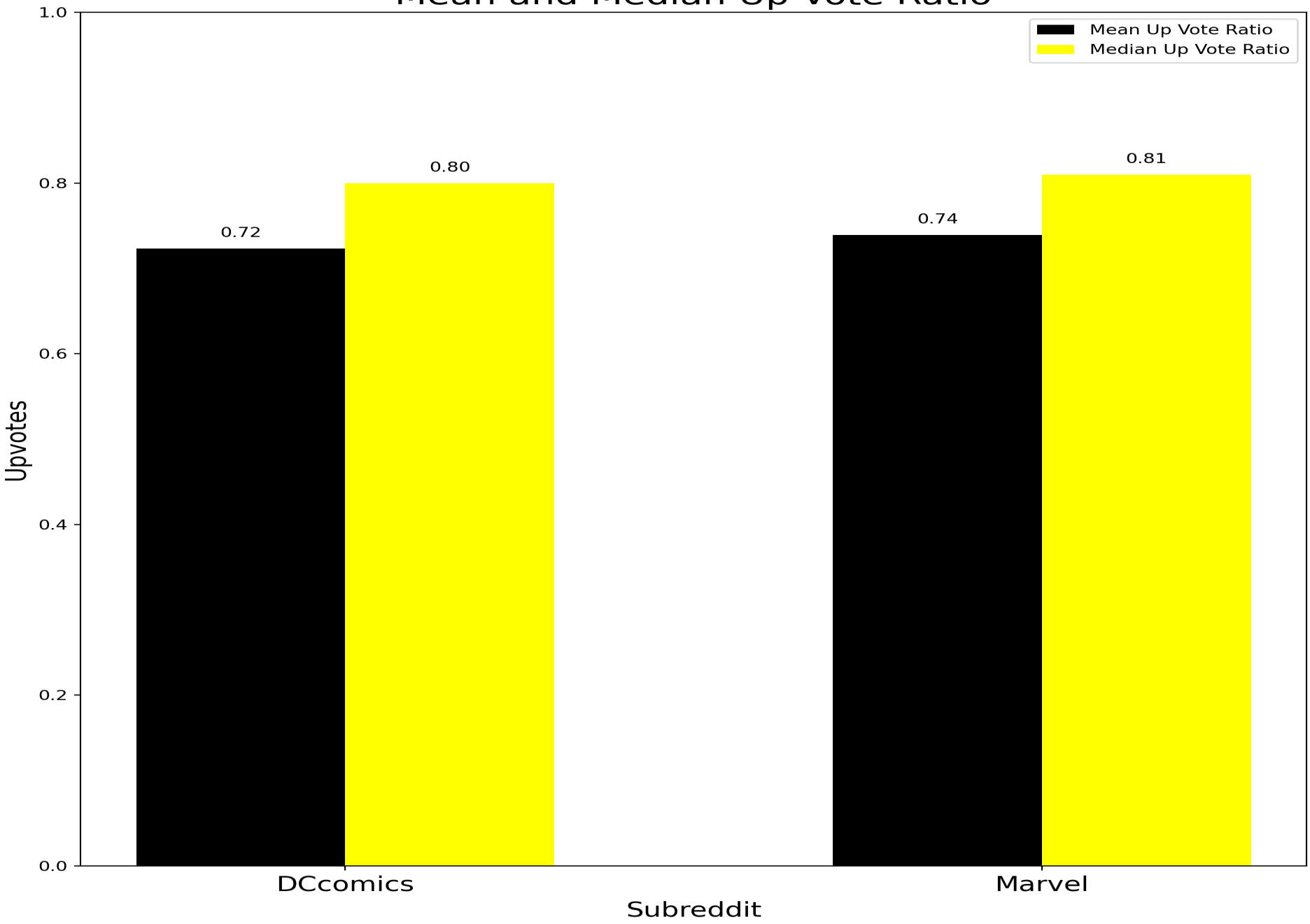
Many Marvel & DC Characters are similar to each other



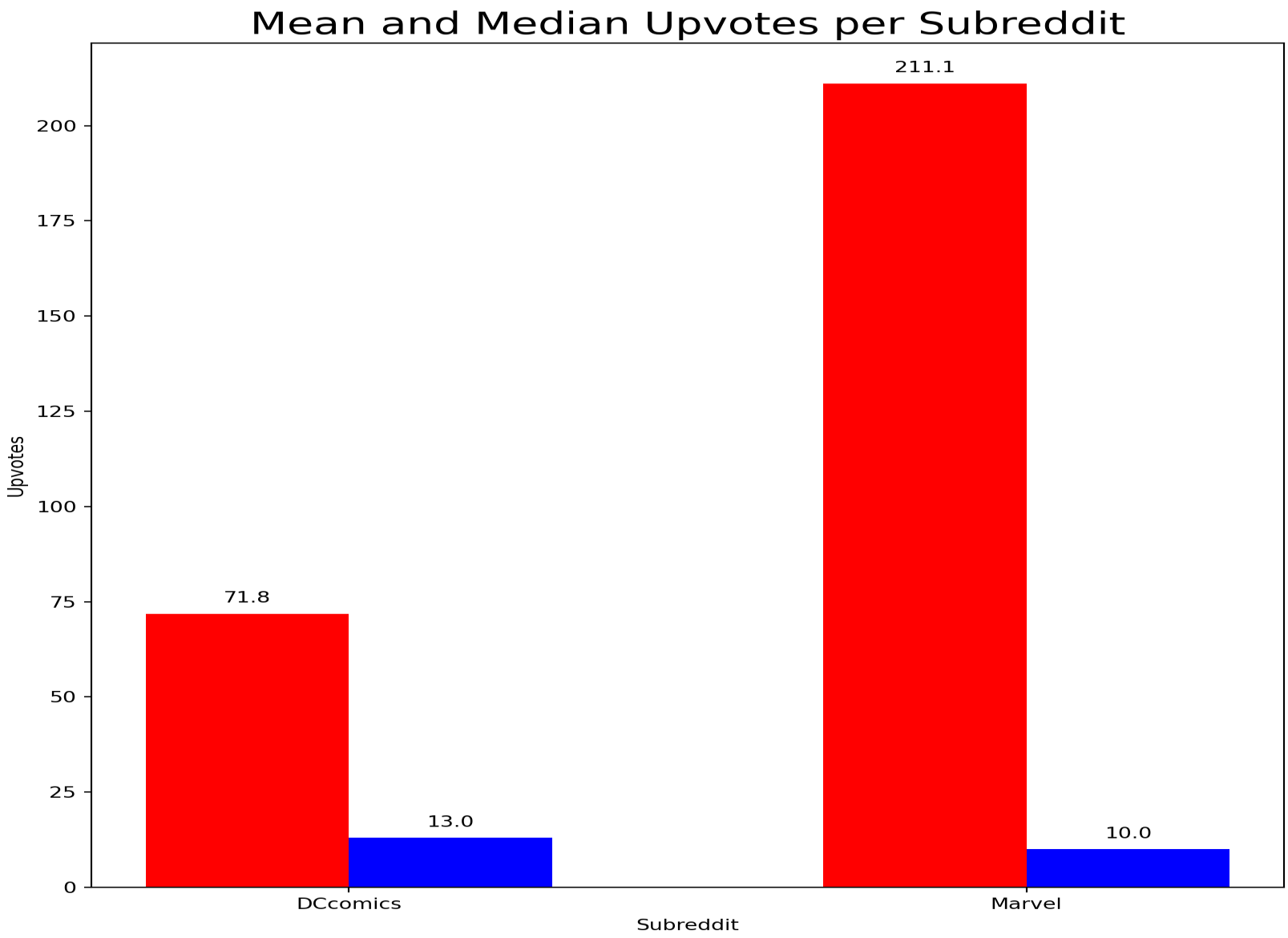
Most DC/Marvel reddit posts are images/video - making NLP more challenging



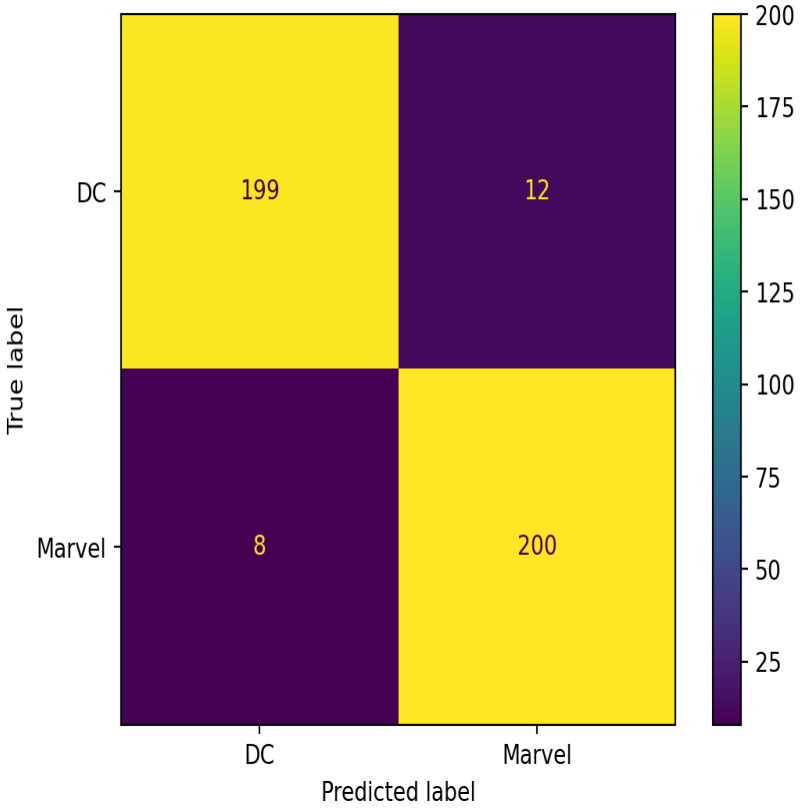

Distribution of Number of Words in Subreddit **Titles**
DCcomics
Marvel
Sub Reddit
0
10
20
30
40
50
Number of Words

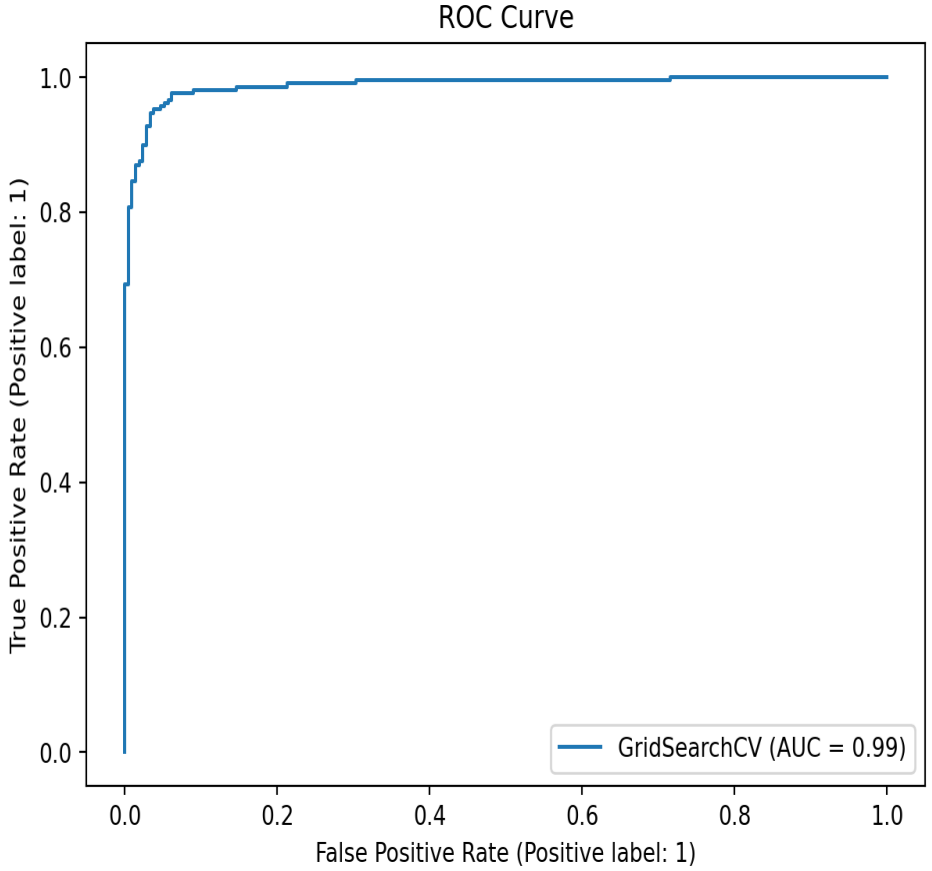

Mean and Median Up Vote Ratio
Mean Up Vote Ratio
Median Up Vote Ratio
1.0
0.8
0.6
0.4
0.2
0.0
Upvotes
0.72
0.80
0.74
0.81
DCcomics
Marvel
Subreddit

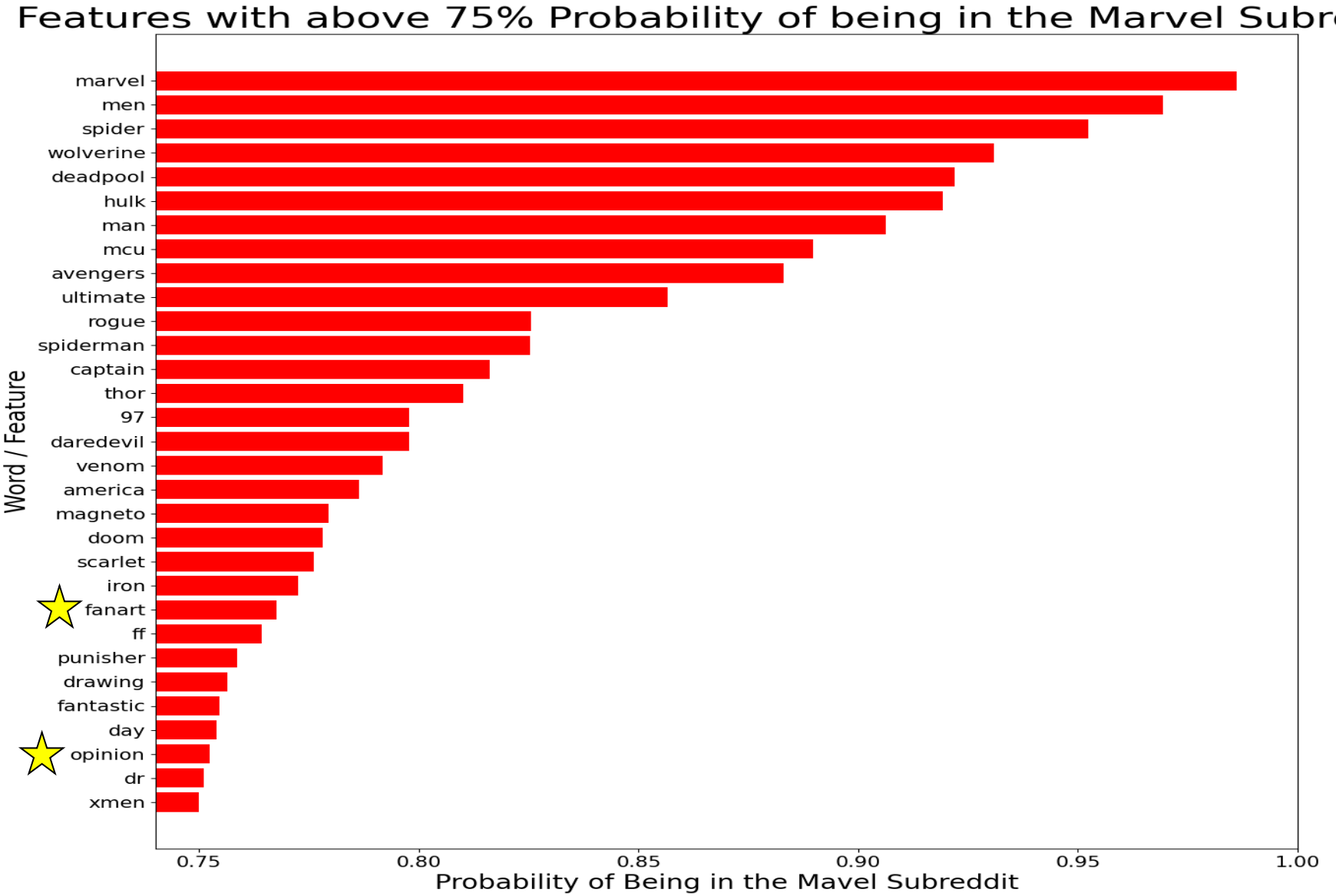
SLIDE 5

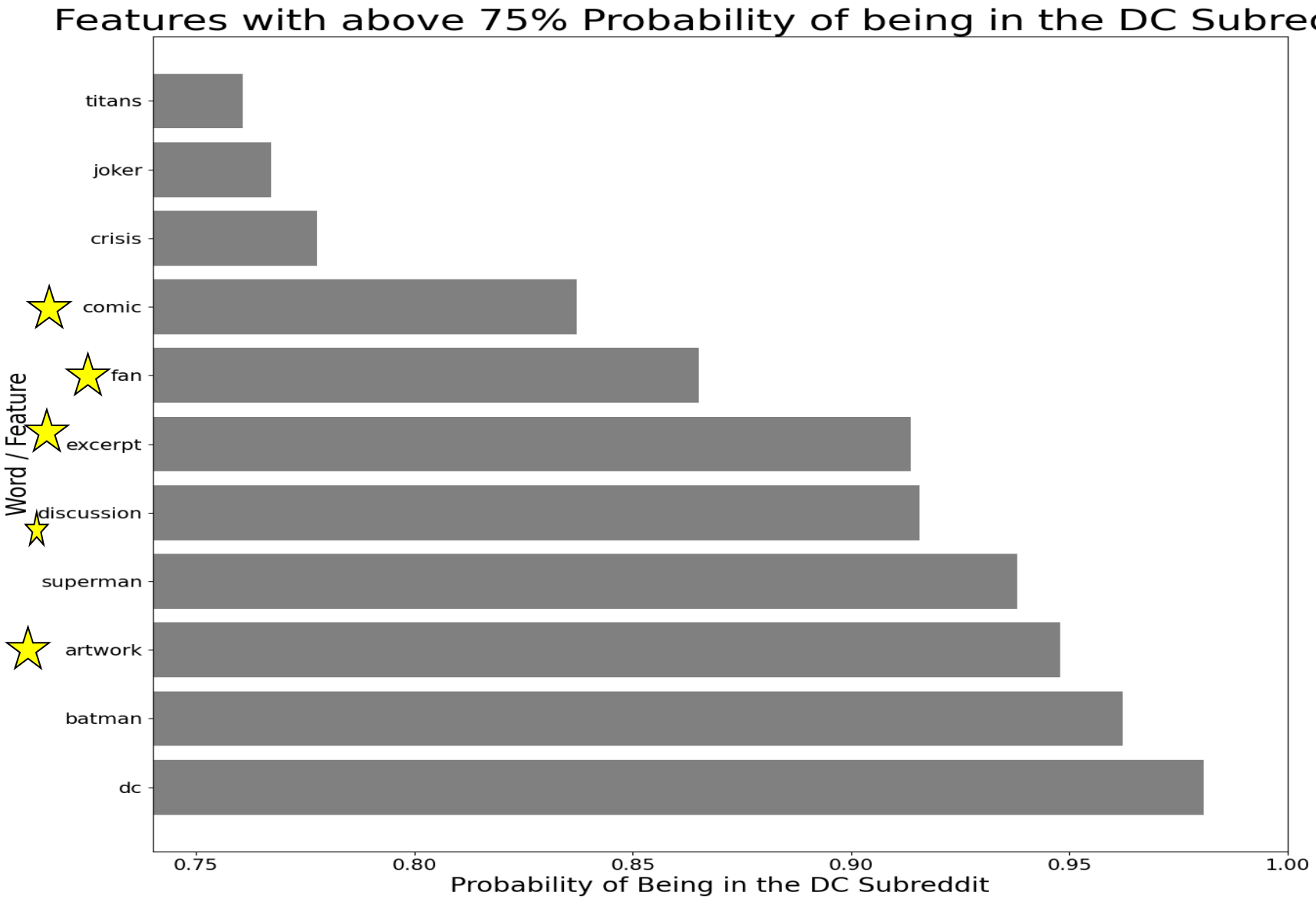

Mean and Median Upvotes per Subreddit
Upvotes
0
25
50
75
100
125
150
175
200
71.8
13.0
211.1
10.0
DCcomics
Marvel
Subreddit

Model Evaluation






Features with above 75% Probability of being in the Marvel Subreddit
marvel
men
spider
wolverine
deadpool
hulk
man
mcu
avengers
ultimate
rogue
spiderman
captain
thor
97
daredevil
venom
america
magneto
doom
scarlet
iron
fanart
ff
punisher
drawing
fantastic
day
opinion
dr
xmen
Word / Feature
0.75
0.80
0.85
0.90
0.95
1.00
Probability of Being in the Mavel Subreddit


Features with above 75% Probability of being in the DC Subreddit
titans
joker
crisis
comic
fan
excerpt
discussion
superman
artwork
batman
dc
Word / Feature
0.75
0.80
0.85
0.90
0.95
1.00
Probability of Being in the DC Subreddit

Understanding Incorrect Predictions

Clean Text	Subreddit	Marvel Probability
what comic is this	Marvel	0.162889
who 's autograph is this found a french black panther comic \n does anyone know the signature is on the cover	Marvel	0.316420
took a stab at making my own action figure comic book cover	Marvel	0.366051
g.o.d.s 7 preview	Marvel	0.402259
is what modern spider- man run is worth reading i was wanting to read nick spencer 's run of sm but heard he messes up his story or it 's really bad etc and just wondering if there 's a better recent current ongoing sm comic that i should read \n	Marvel	0.445635
who is your favorite team and why	Marvel	0.495747
anyone know which version of galactus this is curious cause i 've never seen this version before anyone know anything about him or know his name	Marvel	0.498882
is there any difference between livewires trades with blue or red spines i 've been interested in picking up the livewires trade but i always see two a blue and a red spine \n\n i just wanted to know if there 's a difference and what it is i wanna make sure i 'm not gon na end up getting a black and white reprint or a smaller copy or something	Marvel	0.499925
i prefer to read clark and bruce as supporting or supporting main rather then the focus i enjoy them both as a supporting characters they 're fun to read but i do as much enjoy picking up a comic when it is their story only about them \n\n granted i 've read very few of them as main so i 'm happy for some recs to change my mind	DC	0.500924
help with dc characters i was always more of a marvel guy when it came to comics i watched dc movies but have read any dc comics other then the batman prey storyline i some recommendations for characters you think i would like some of my favorite characters from marvel are wolverine punisher ghost rider and spider man another one of my favorite super hero is not from marvel or dc he is the crow if you know you know any characters recommendations based on some of my favorites from other things	DC	0.514172
the flash series i realize why i stop watching it before and why i stop watching again now haha so i remember when the flash series was on it 's season 2 i was so into it then i stopped watching after that season then now 2024 i saw it again on netflix so i decided to start watching it again from the beginning and i was enjoying it i kinda ask myself why did i stop watching this awesome series what made me not continue watching it then on the last part of season 2 when barry go save her moth...	DC	0.516560
worlds first fully rexine built dr fate cosplay my first attempt at making dr fate cosplay using rexine i was literally in an oven	DC	0.518235
how many lobo compendium will there be first one will appear in june but what about other volumes	DC	0.524633
what are your thoughts on the start of house of brainiac so far i thought it was an amazing start for the event and i like the twist with ll-01 what are your thoughts and theories what do you think brainiac wants \n\n is lena gon na be the brainiac queen \n\n\n rafa sandoval 's art is amazing i first saw his art in hal and glc book and man i hope we get another series with him again \n\n\n williamson does have a best track record for events and i really hope he is able to stick the landing t...	DC	0.549556
james tynion iv returns to nice house on the lake for sequel	DC	0.569148
merchandise static shock figure collection i 've collected every officially licensed static shock figure released so far i 'm bummed this all but it feels good to have everything	DC	0.593261
form and function is there an in canon explanation why so many alien races are generally shaped like earthlings that is why bifurcated	DC	0.597453
what is high father the god of so with the new god 's they 're all supposed to represent platonic ideas granny goodness is the god of child abuse darkseid is the god of tyranny and evil and so on so what does that make high father the god of is he the god of hope and life i know he 's also the prophet of the source that 's why i say life and hope	DC	0.609467
\n josh keaton 's web warriors \n \n\n	DC	0.614206
other marvel taking a jab at dc comics for identity crisis marvel knights 4 23 how would a man with stretchy powers protect their wife named sue	DC	0.660884

Conclusions and Recommendations

- Despite most reddit posts having short titles without much text in the body of the post, a Logistic Regression with Term Frequency-Inverse Document Frequency Vectorizer produces a very good model with over 95% accuracy in predicting whether a title/post is from a Marvel or DC subreddit by identifying the name of the comic book company and types of character names and storylines in each company.

Next Steps

- Adding DC and Marvel (and related like MCU, DCEU) to the stop words to see how well the model works if it can't select on the dependent variable.

- Look at getting comments. Especially for posts with images and short titles, comments may help identify whether a post is for DC or Marvel comics.

- Look at using up vote counts as another feature. Since Marvel posts get more up votes, in cases where the model is uncertain, it could use the up votes to help identify which subreddit it came from.

- Explore computer vision. Since many of the posts are images/video, computer vision may be able to identify whether a character is from Marvel or DC based on previous images it has seen. Would that be better/worse than text (since a character may have different looks based on an artist's style or costume changes)

- Look at using the words (and number of words) in a title/post to predict the number of up votes and up vote ratio with a linear regression. Would be helpful to figure out if I wanted to make a post go viral, what are the elements that would help. Ex : do I draw a picture of Superman or Batman? Do I link to a video clip about Wolverine or Deadpool?

- Now that we understand what words/phrases predict whether someone is talking about Marvel or DC comics, use that to predict whether a post on the general comicbooks reddit thread is about Marvel or DC (or neither company)