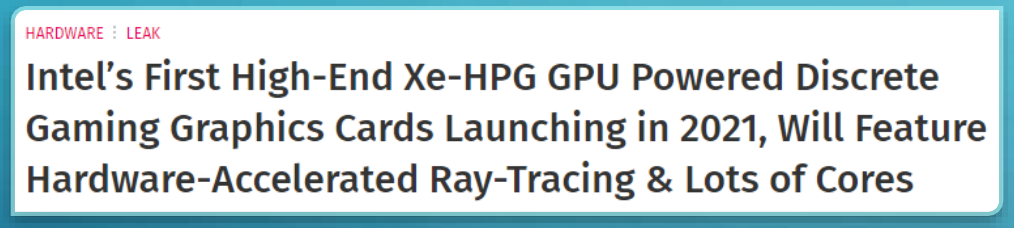
TOPIC MODELING ON AMD AND NVIDIA'S GPU LAUNCH COMMENTS

DONE BY: N KUMERESH, DSI 18

GENERAL ASSEMBLY

CONTENTS

INTRODUCTION: BACKGROUND

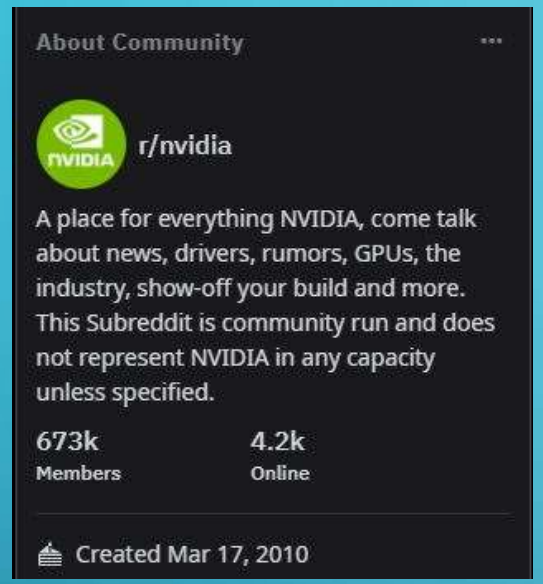


- Intel is planning to release their first gaming graphics card into the market.
- They're intending to compete against AMD's RDNA and Nvidia's Ampere series.

INTRODUCTION: PROBLEM STATEMENT

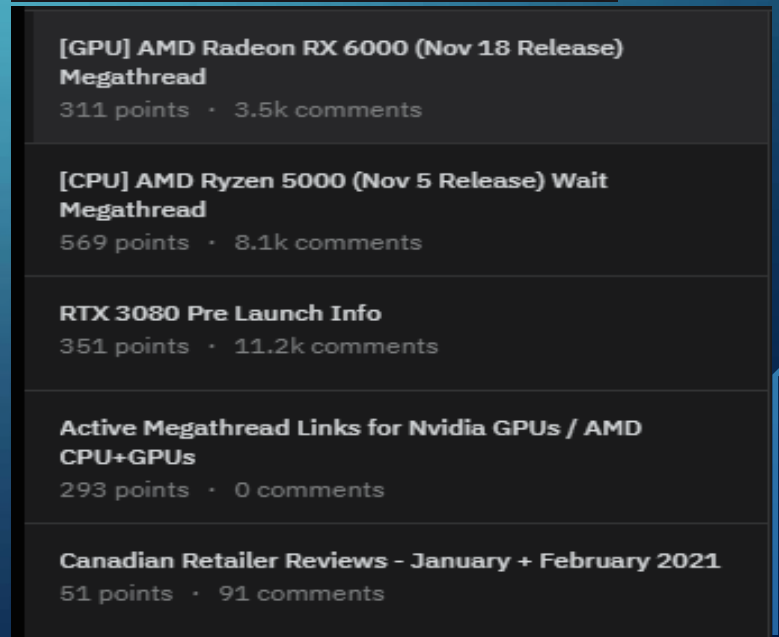
- As a data scientist working in intel, my team is tasked with understanding the consumer behavior surrounding the launch of AMD and Nvidia's GPUs.
- Based on our research, majority of the fanbase primarily resides in their respective subreddits.
- This problem is solved by conducting topic modeling on the comments understand what are the consumers looking out for when purchasing a GPU to ensure a smooth release of our own GPUs.

SCRAPPING THE DATA



Scrapped using PRAW's API

A total of 34,000 comments were scrapped

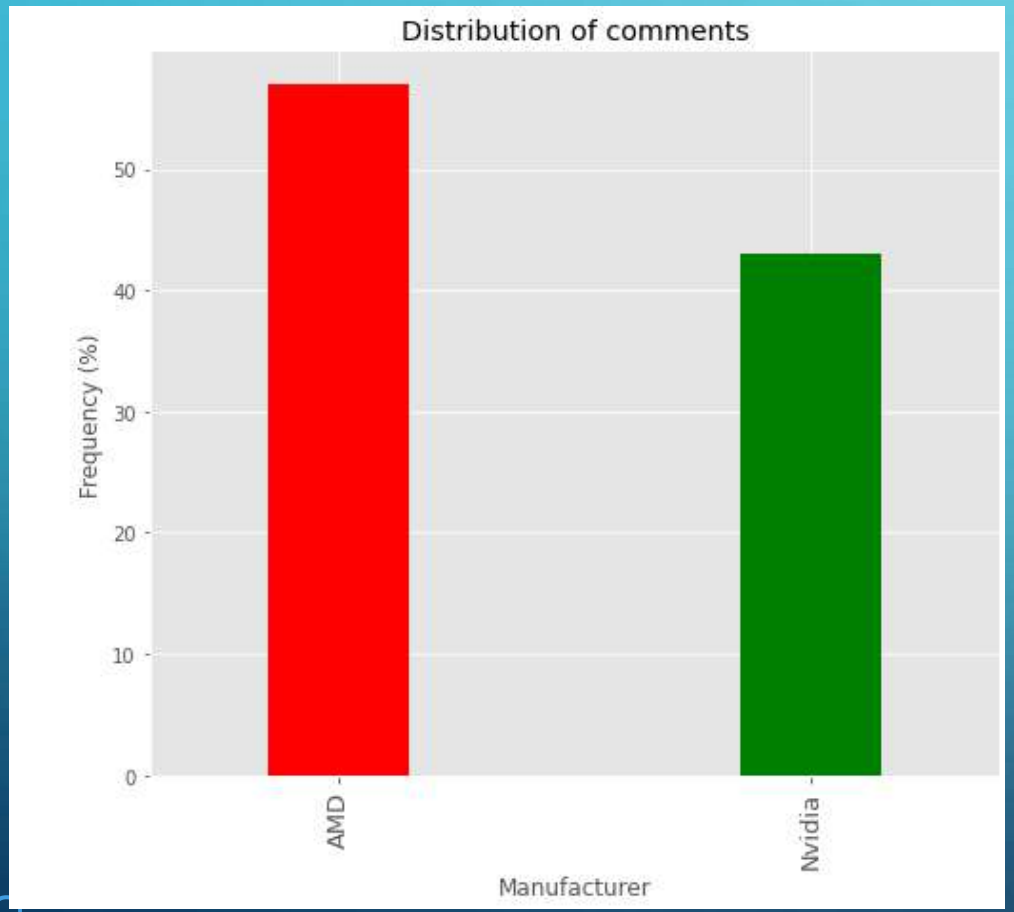


CLEANING THE DATA

Remove the nulls values → Drop comments that are empty → Remove duplicate values → Remove stop words, lemmatize the words

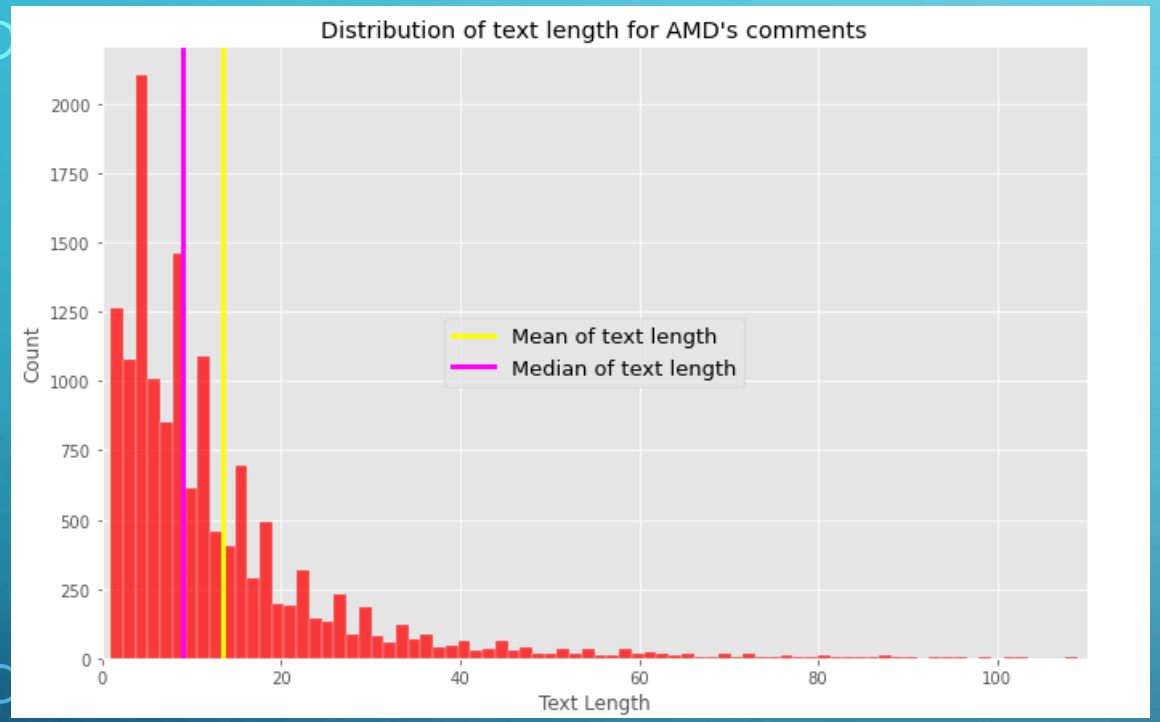
	Reddit comments	tag
0	pre order time releasing 17th	nvidia
1	going hard grab 3080 17th	nvidia
2	uk price 3090 1399 3080 649 3070 469 scan aib ...	nvidia
3	talking spatula jensen pot	nvidia
4	10k core completely insane	nvidia
...	...	...
25511	motherboard say pcie 0 compatible	nvidia
25512	thanks understanding know course 3000s dvi por...	nvidia
25513	buy 1400 gpu	nvidia
25514	know ill wait 3070 3060	nvidia
25515	generational compatibility compatible size pci...	nvidia

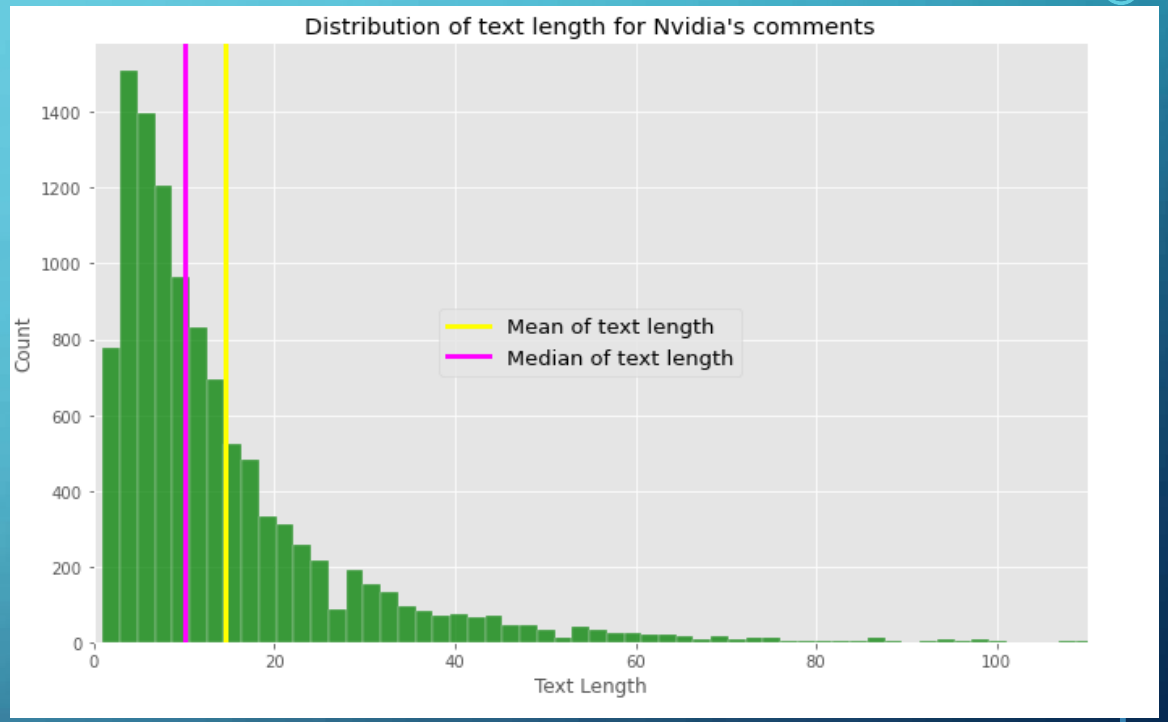
EDA: DISTRIBUTION OF AMD AND NVIDIA'S COMMENTS



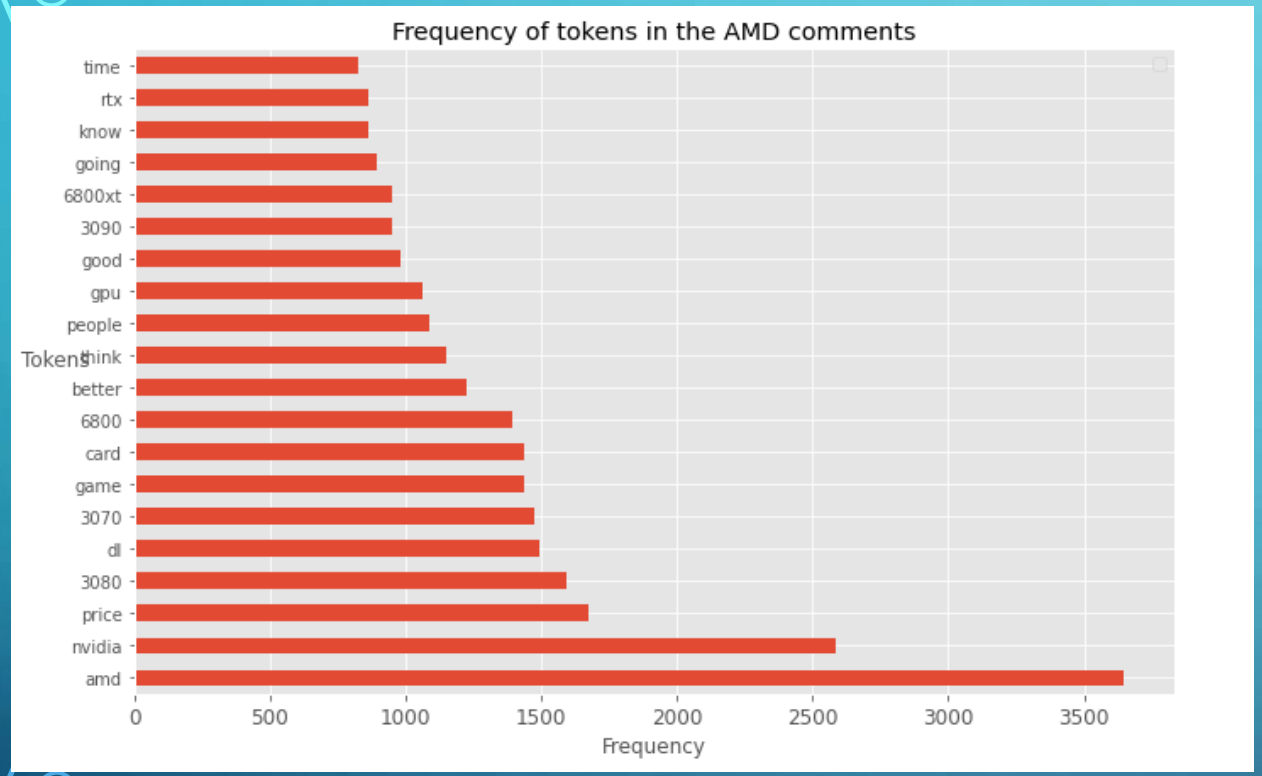
Around 56% of the comments were from AMD and 44% of the comments were from Nvidia

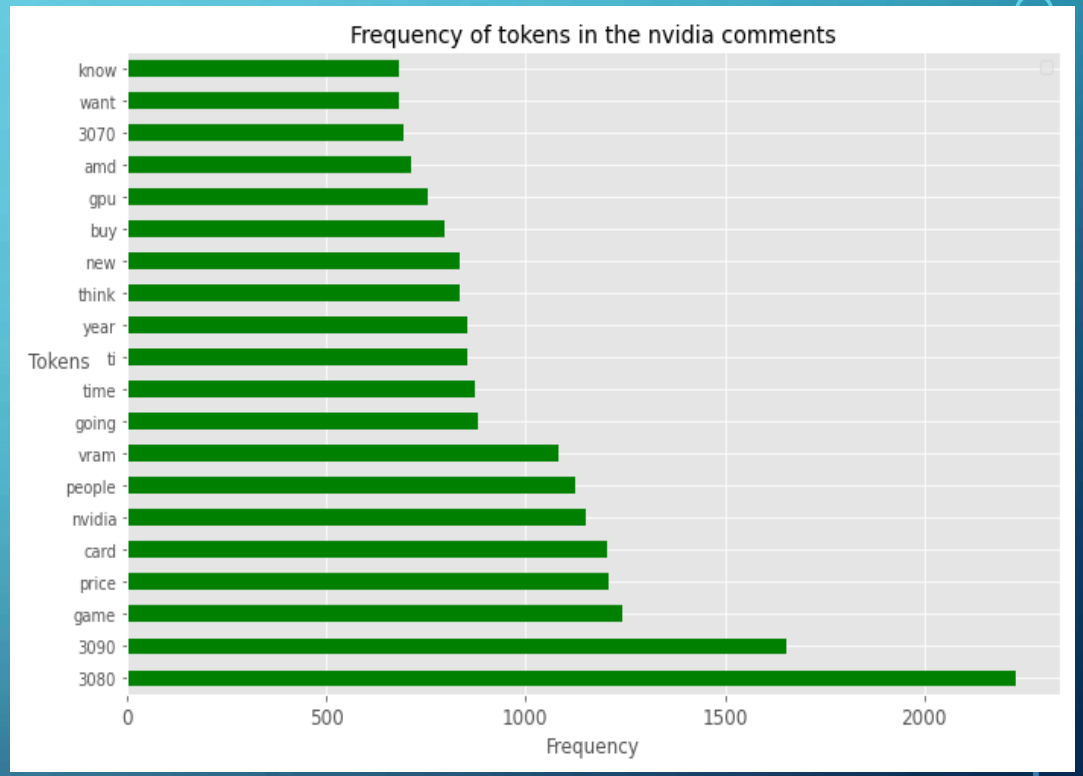
EDA: DISTRIBUTION OF TEXT LENGTH





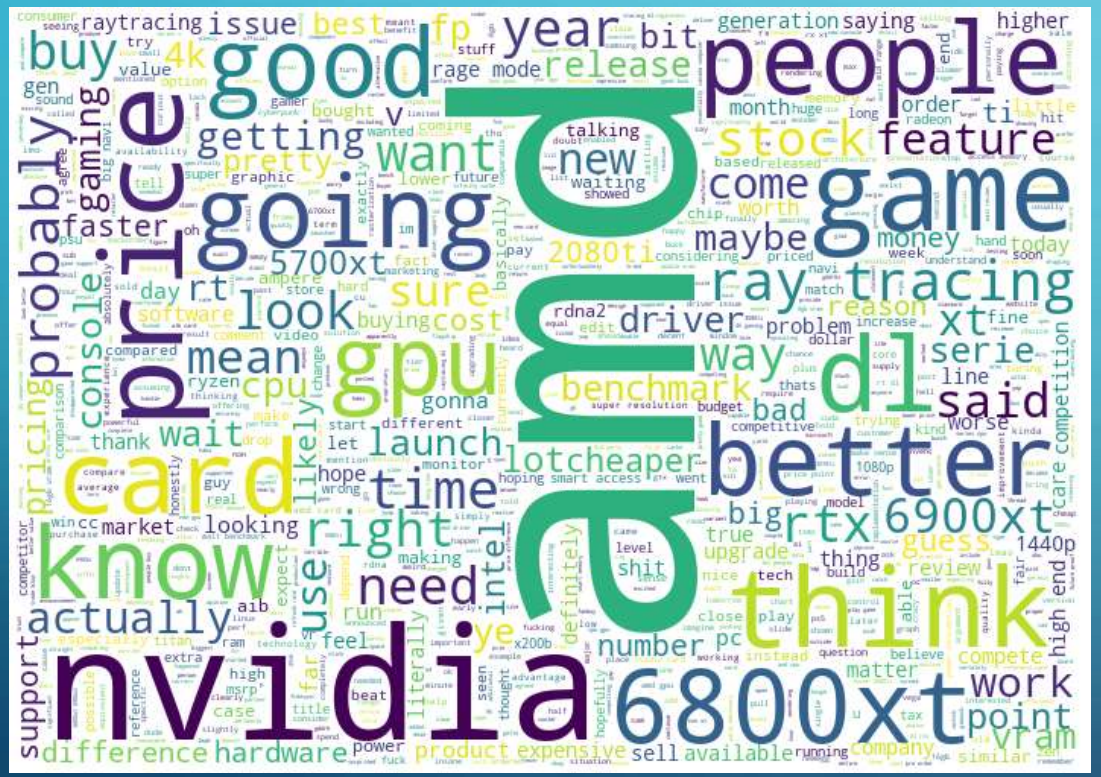
EDA: FREQUENCY OF TOKENS



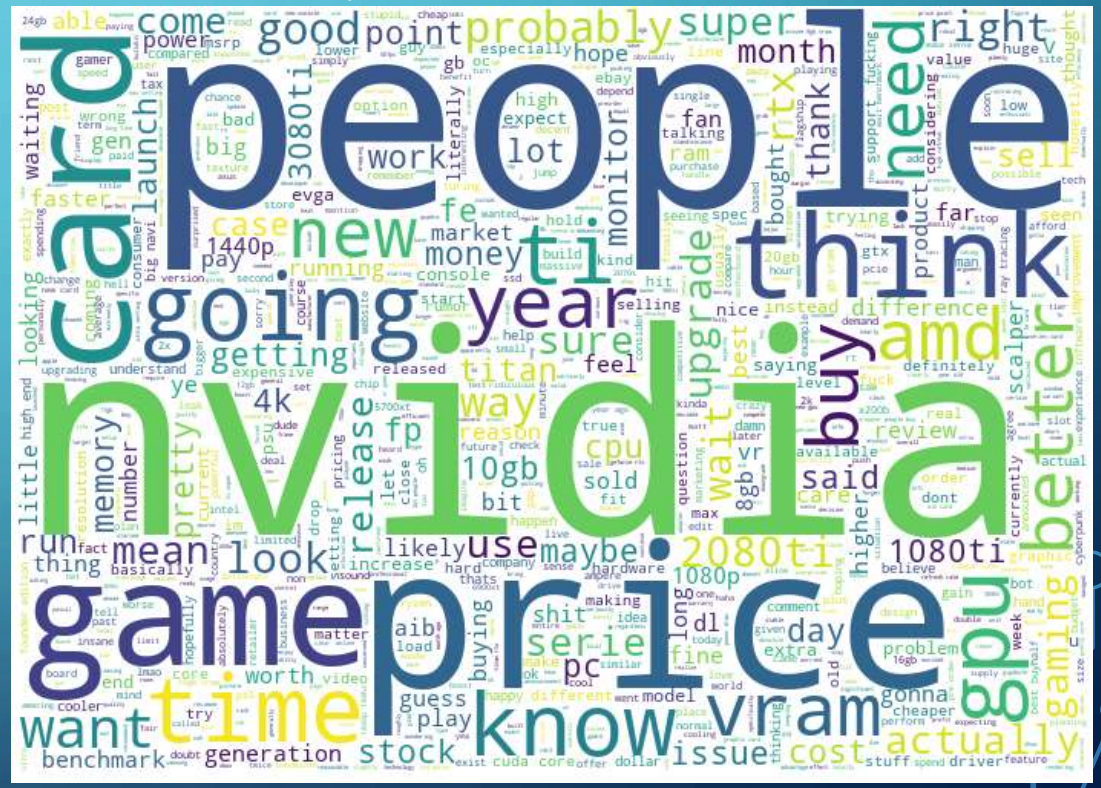


EDA: WORD CLOUDS ON NVIDIA AND AMD'S COMMENTS

r/AMD's word cloud



r/Nvidia's word cloud



PREPARE THE DATA FOR MODELING

Create the bag of words

```
# Creates the bag of words for each document
data = [token.split() for token in amd_df['Reddit comments'].tolist()]
```

executed in 31ms, finished 16:55:22 2021-01-26

Using Genism's bigram and trigram Phraser

```
# Build the bigram and trigram models
bigram = Phrases(data, min_count=5, threshold=100) # higher threshold fewer phrases.
trigram = Phrases(bigram[data], threshold=100)

# Faster way to get a sentence clubbed as a trigram/bigram, this reduces memory, making the model smaller and faster
bigram_mod = Phraser(bigram)
trigram_mod = Phraser(trigram)
```

executed in 3.10s, finished 16:55:25 2021-01-26

Create the bigrams and trigrams

```
def make_bigrams(texts):
    return [bigram_mod[doc] for doc in texts]

def make_trigrams(texts):
    return [trigram_mod[bigram_mod[doc]] for doc in texts]

# Form Bigrams
data_bigrams = make_bigrams(data)

# Form Trigrams
data_trigrams = make_trigrams(data)
```

executed in 1.29s, finished 16:55:27 2021-01-26

AMD COMPARING LDA MODEL AND LDA MALLET RESULTS

LDA

	Perplexity score	Coherence score	no of topics intrepretable
unigram	-7.678092	0.529254	3
bigrams	-7.722717	0.585108	3
trigrams	-7.759911	0.590325	4

LDA Mallet

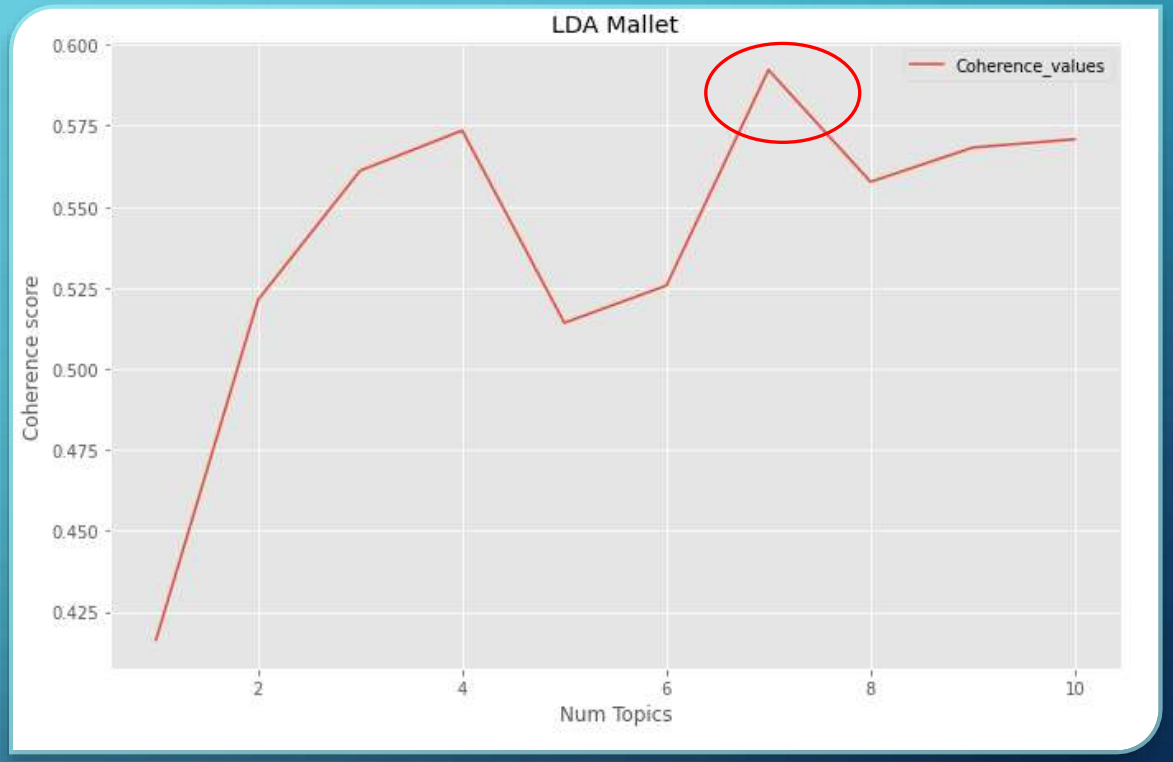
coherence_values	num_topics
0.416386	1
0.521286	2
0.561158	3
0.573551	4
0.514202	5
0.525694	6
0.592246	7

LDA Topics

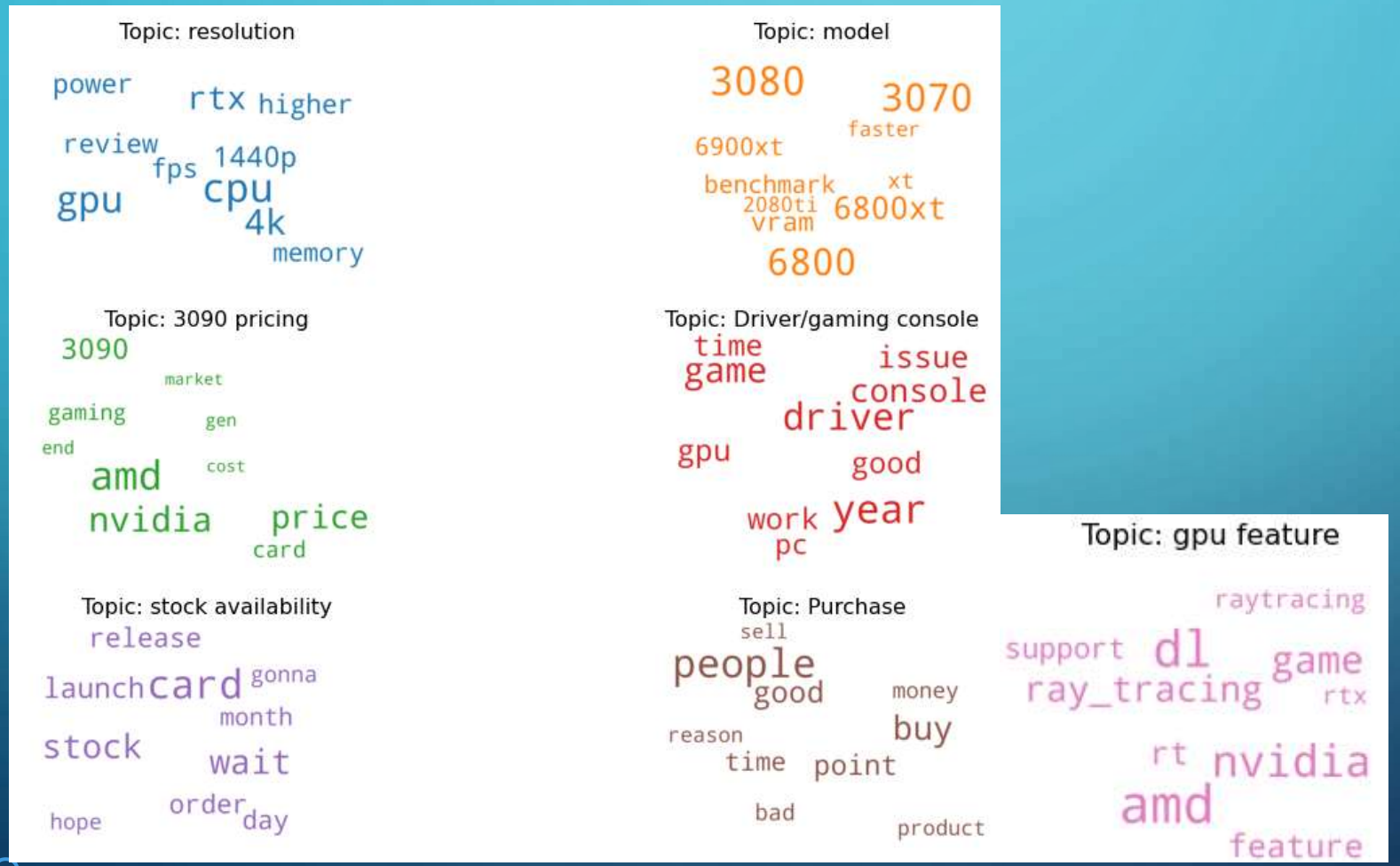
```
*********Displaying 5 topics for trigrams*****************
[(0,
  '0.013*"card" + 0.011*"amd" + 0.011*"time" + 0.010*"people" + 0.010*"stock" '
  '+ 0.009*"buy" + 0.008*"launch" + 0.008*"day" + 0.007*"month" + '
  '0.007*"year"'),
 (1,
  '0.018*"game" + 0.016*"4k" + 0.013*"gpu" + 0.012*"work" + 0.012*"monitor" + '
  '0.012*"cpu" + 0.008*"gaming" + 0.008*"fps" + 0.008*"1080p" + 0.007*"use"'),
 (2,
  '0.038*"amd" + 0.031*"nvidia" + 0.027*"dl" + 0.017*"game" + '
  '0.014*"ray_tracing" + 0.009*"driver" + 0.009*"rt" + 0.009*"feature" + '
  '0.008*"card" + 0.008*"better"'),
 (3,
  '0.010*"power" + 0.009*"psu" + 0.008*"comment" + 0.008*"thanks" + '
  '0.007*"efficiency" + 0.007*"yes" + 0.006*"think" + 0.006*"read" + '
  '0.006*"load" + 0.005*"company"'),
 (4,
  '0.028*"price" + 0.026*"3080" + 0.025*"3070" + 0.024*"6800" + 0.016*"3090" + '
  '0.016*"6800xt" + 0.012*"vram" + 0.011*"better" + 0.010*"6900xt" + '
  '0.009*"nvidia"')]
```

AMD LDA MALLET: CHOOSING THE BEST K TOPIC

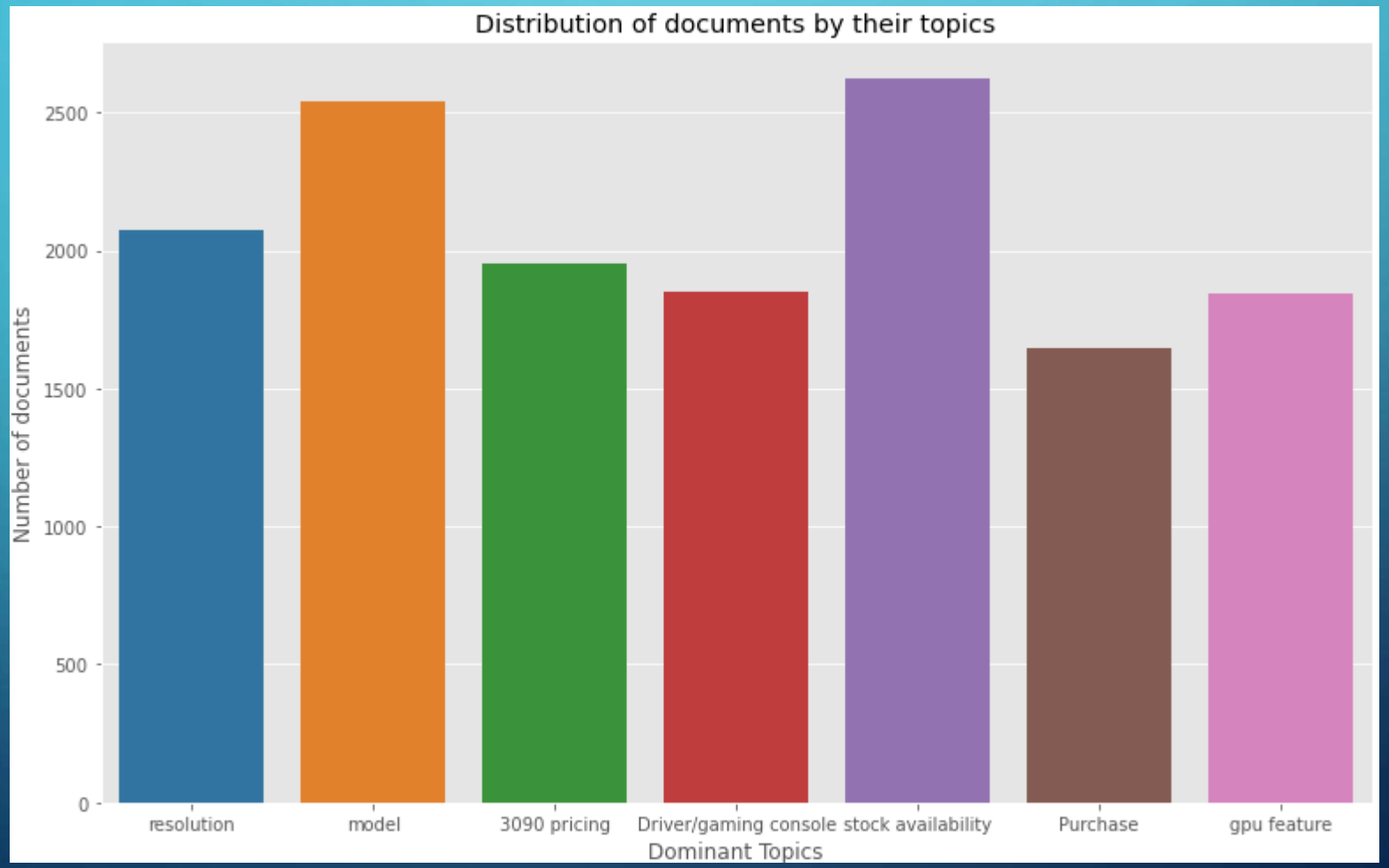
- The trigrams data was trained on the LDA mallet model
- 7 topics was chosen as it had the highest coherence score



AMD LDA MALLET: WORD CLOUD OF 7 TOPICS



AMD: DISTRIBUTION OF DOCUMENTS BY THEIR TOPICS



AMD: MOST REPRESENTATIVE DOCUMENTS FOR TOPIC 4 AND 6

*******Topic 4********

Dominant keywords: card, wait, stock, launch, release, order, day, month, gonna, hope

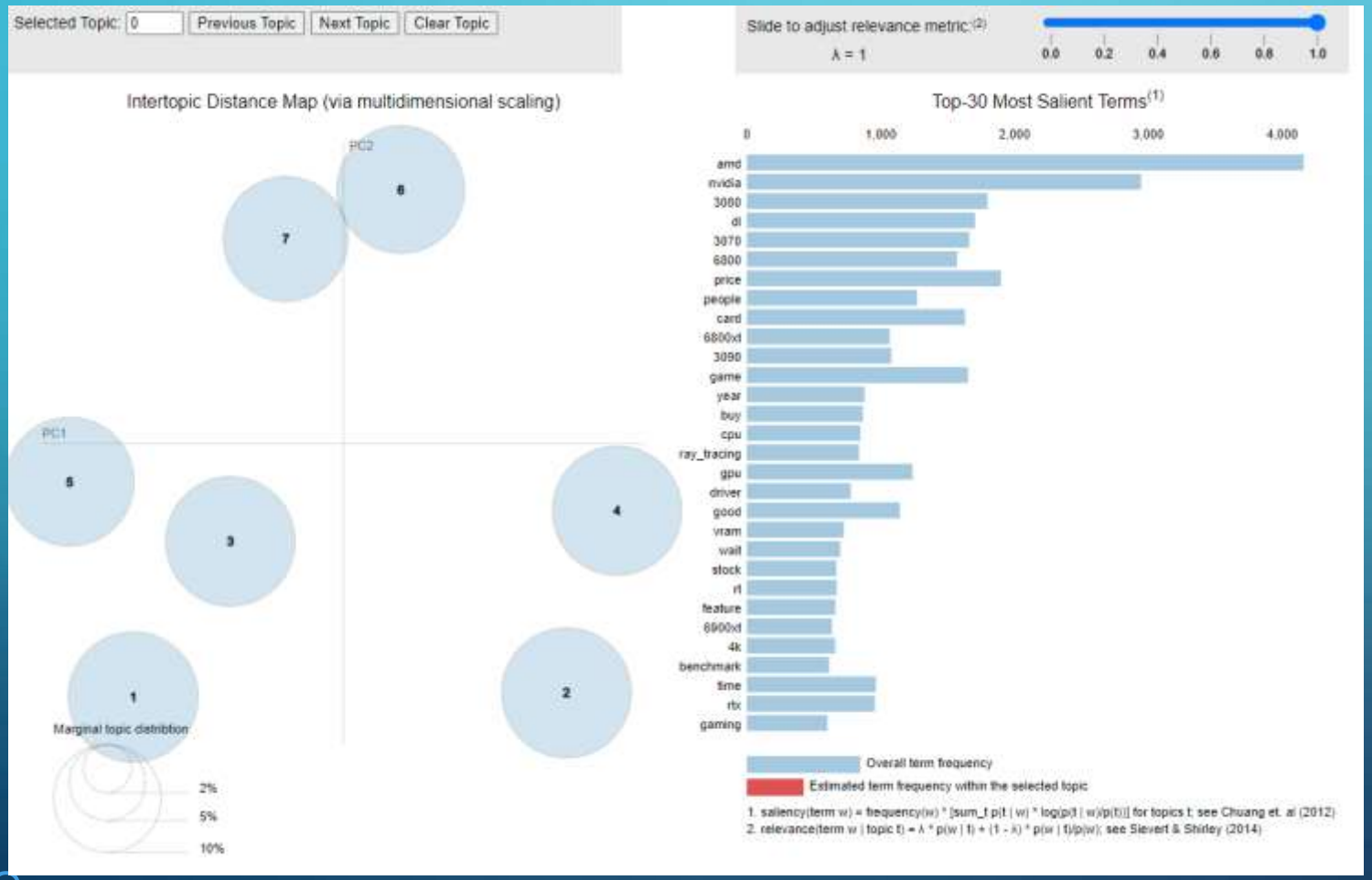
Document:alright called couple cc location 4038 hwy 7 unionville personal fav 6600 yonge st manager unionville sure 6800xt card tmr allow pre_order partner card preorders reference version priority taking pre_order reference card apparently come head office rep spoke manager centerpoint know tmr unlikely taking pre_order mixed message existing pre_order priority partner pre_order location belief reference card existing order need fulfill aib card absolutely word model going coming

*******Topic 6********

Dominant keywords: amd, dl, nvidia, game, ray_tracing, rt, feature, support, raytracing, rtx

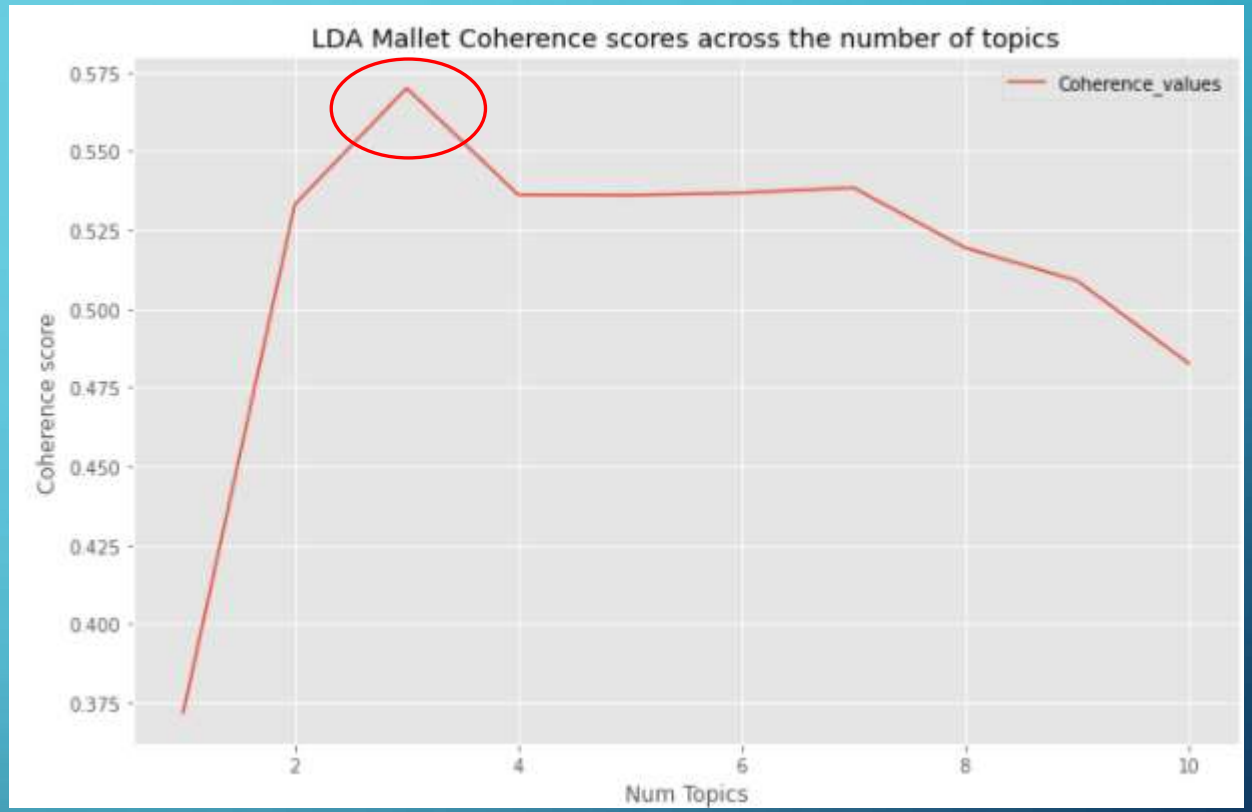
Document:example ray_tracing insanely computationally expensive push higher higher resolution jumping 1080p 1440p 4k adding big load ray_tracing exacerbates essentially enabling ray_tracing half huge hit think agree exactly dl important negates halving ray_tracing guarantee devs away completely fake baking lighting cube map use rtx know long way away unfortunately yes computationally expensive far easier devs implement physic based work box hundred hour cube map hit generally getting ray_tracing effect certain are a example ray_traced environmental lighting effect muzzle trade lighting option severe penalty yep absolute best implementation control far good cyberpunk implementation albeit bet extremely good given close partnership nvidia said implementation computational cost high best cast ray_pixel enormously far realistic rendering theoretically say ne

AMD: PYLDAVIS



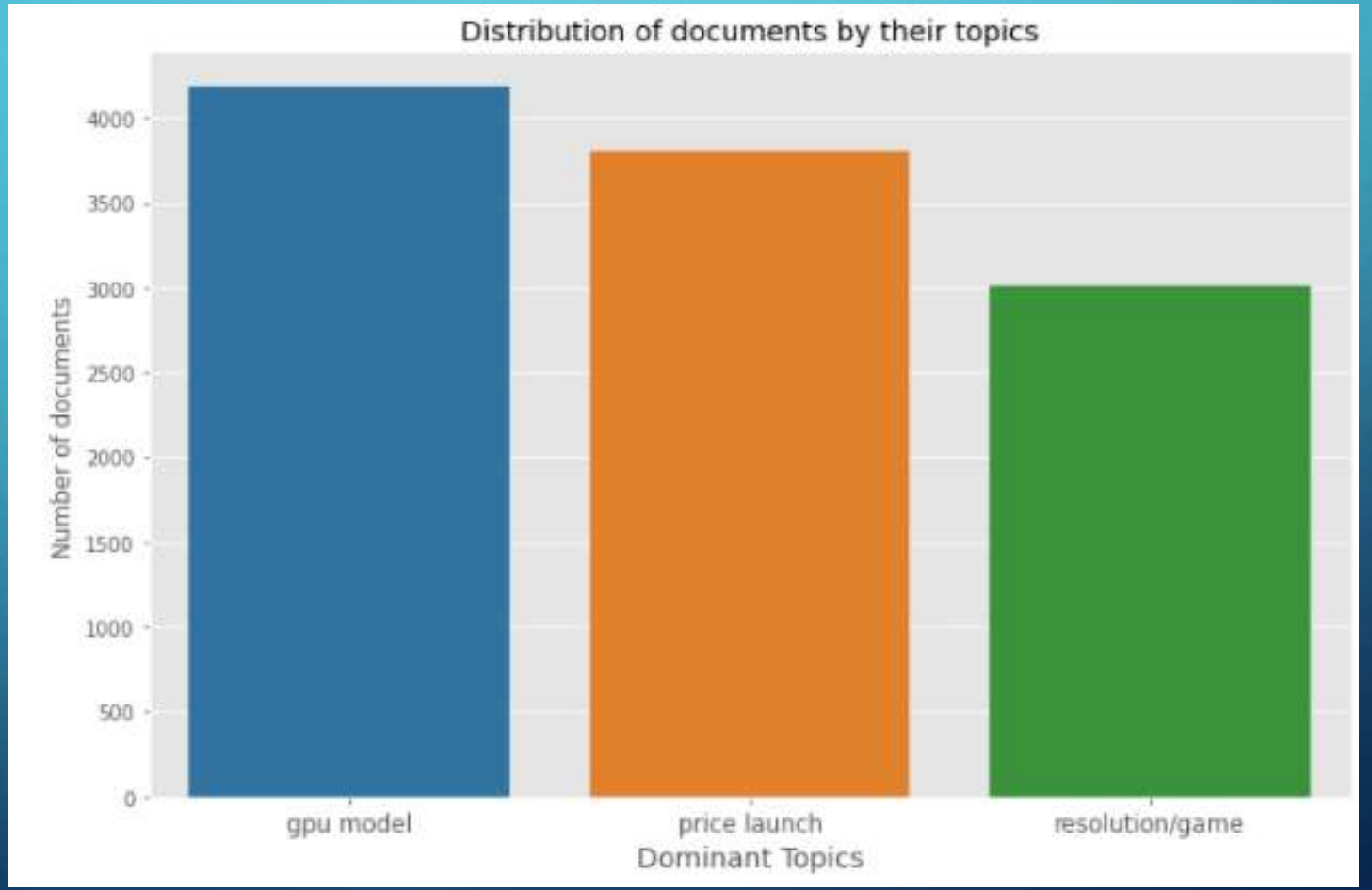
NVIDIA LDA MALLET: CHOOSING THE BEST K TOPIC

- The trigrams data was trained on the LDA mallet model
- 3 topics was chosen as it had the highest coherence score



AMD LDA MALLET: WORD CLOUD OF 7 TOPICS

NVIDIA: DISTRIBUTION OF DOCUMENTS BY THEIR TOPICS



NVIDIA: MOST REPRESENTATIVE DOCUMENTS FOR TOPIC 0 AND 1

*******Topic 0********

Dominant keywords: 3080, 3090, ti, amd, 3070, rtx, 2080, upgrade, 2080ti, wait

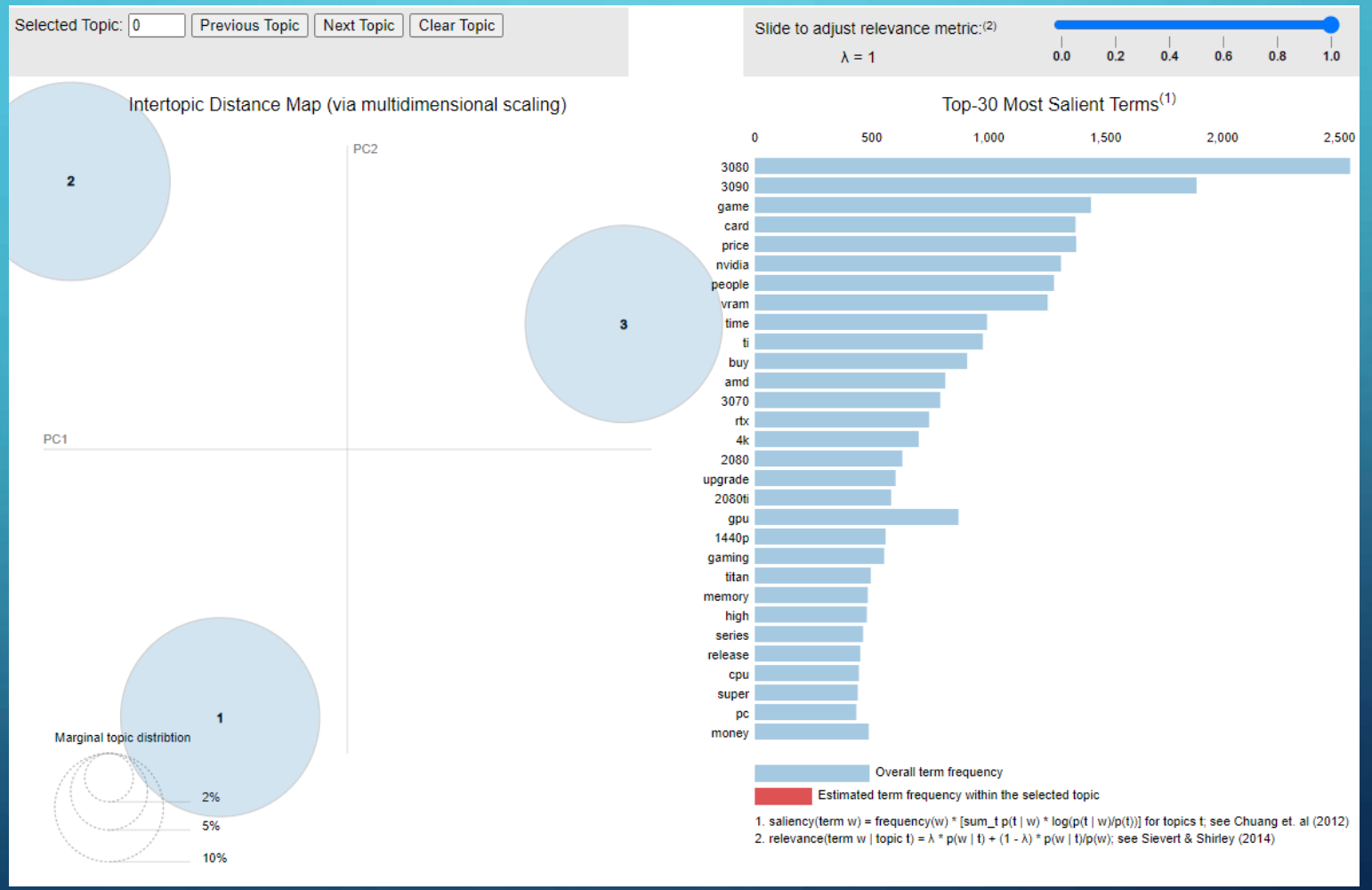
Document:3090 simply titan titan driver basically titan mind 2080 ti increase vram 1080 ti previous gen doubled respectively x70 class 8gb 1070 previously doubling gen prior 28gb 570 2gb 670 2gb 770 680 rebrand 4gb 970 8gb 1070 staying 8gb similarly x80 class doubled gen staying 8gb 1080 2080 receiving minor capacity upgrade 10gb 3080 supposed believe 10gb 3080 upgrade path 11gb 2080 ti compute 3090 titan driver actual titan

*******Topic 1********

Dominant keywords: price, card, nvidia, people, time, buy, money, series, release, day

Document:edit opinion x200b know teaser_rate car advertised insanely price dealership told sold cost little 699 3080 teaser_rate nvidia hyped insanely hint 3080 wanted insanely cheap price announced wanted problem nvidia stock 3080 place retailer stock place know run merchandise tracking system retailer nvidia hype knowing stock satisfy basic demand release day come nvidia counted bot ebay taking 3080 actually available blame bot nvidia overselling hype delivering card bot responsible retailer nvidia hype train knew sufficient stock satisfy basic demand retailer scam seriously believe retailer know stock send 14 99 deliver moon someday guarantee 1000 year offer valid entire universe send money backlash swift betting nvidia shat pant reading forum pissed consumer oh look pre_order delivery october november soon december year nvidia placate consumer come hey worse ussr ordered car year advance x200b nvidia destroyed marketing plan generate hype oh look mystery nda nvidia release video showing massive impro

NVIDIA: PYLDAVIS



PROJECT LIMITATIONS

- Data was scrapped from reddit only, which is US-based forum and might form a biased view in the features that consumers are looking out for hence, we can consider scrapping from other forums as well, such as local forums like hardwarezone.
- Consider other models like BertTopic and LSA to see if the topics are more interpretable and the topic coherence scores are higher.

CONCLUSION

- LDA's mallet coherence score is slightly higher compared to LDA's coherence score and the topics are more interpretable as well.
- 4k resolution has a higher topic weight compared 1440p resolution which suggests that these GPUs are targeted towards the enthusiast crowd.
- Nvidia's RTX 3070 and 3080 seems to be the most popular models, given their high word count and topic weightage.
- DLSS and Ray Tracing are coveted features among the users in the subreddit

End of presentation