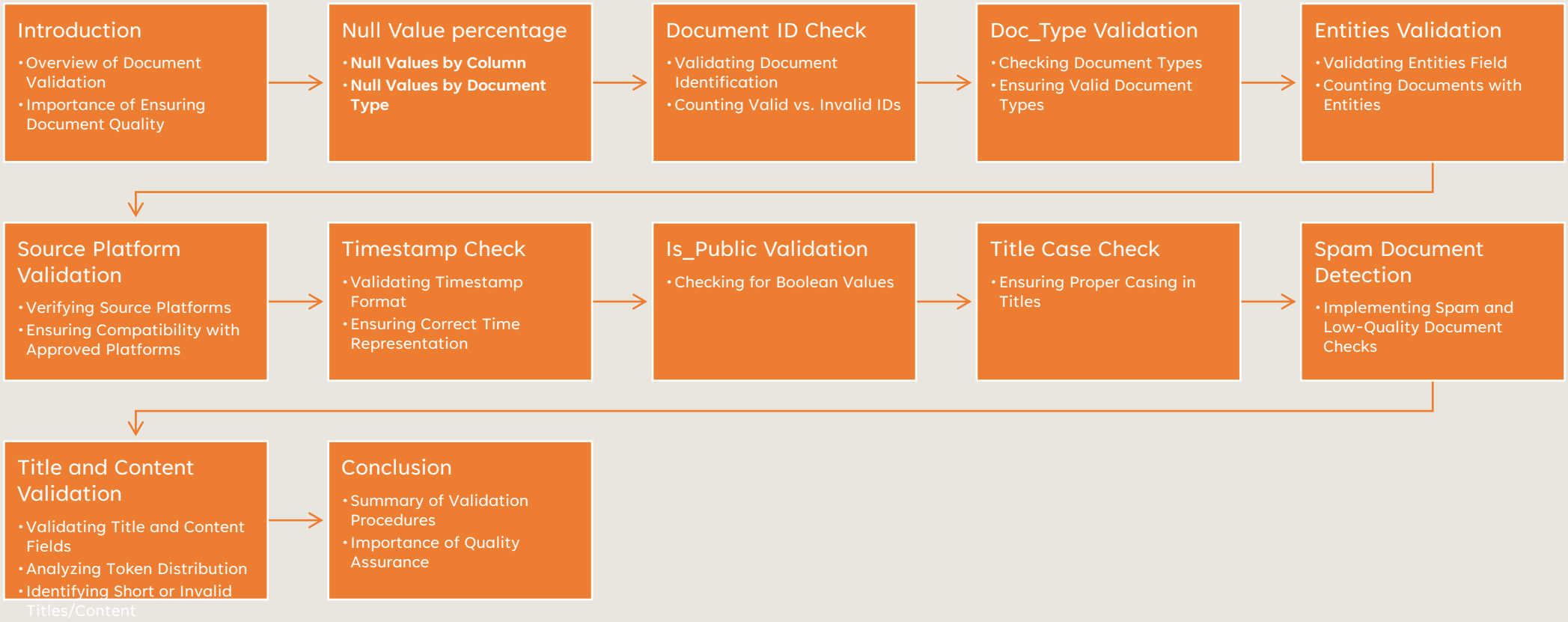
DOCUMENT QUALITY ASSURANCE AND VALIDATION PROCEDURES

Shashank Shukla

AGENDA


Introduction
• Overview of Document Validation
• Importance of Ensuring Document Quality
Null Value percentage
• Null Values by Column
• Null Values by Document Type
Document ID Check
• Validating Document Identification
• Counting Valid vs. Invalid IDs
Doc_Type Validation
• Checking Document Types
• Ensuring Valid Document Types
Entities Validation
• Validating Entities Field
• Counting Documents with Entities
Source Platform Validation
• Verifying Source Platforms
• Ensuring Compatibility with Approved Platforms
Timestamp Check
• Validating Timestamp Format
• Ensuring Correct Time Representation
Is_Public Validation
• Checking for Boolean Values
Title Case Check
• Ensuring Proper Casing in Titles
Spam Document Detection
• Implementing Spam and Low-Quality Document Checks
Title and Content Validation
• Validating Title and Content Fields
• Analyzing Token Distribution
• Identifying Short or Invalid Titles/Content
Conclusion
• Summary of Validation Procedures
• Importance of Quality Assurance

INTRODUCTION

In this presentation, we will delve into the meticulous process of data validation and anomaly detection. With a dataset comprising approximately 2 million records, our objective is to ensure data accuracy and integrity before feeding it to the Vespa. We'll explore various validation checks and methodologies, shedding light on key insights and findings. Join us on this journey as we unravel the intricacies of data quality assurance and anomaly identification.

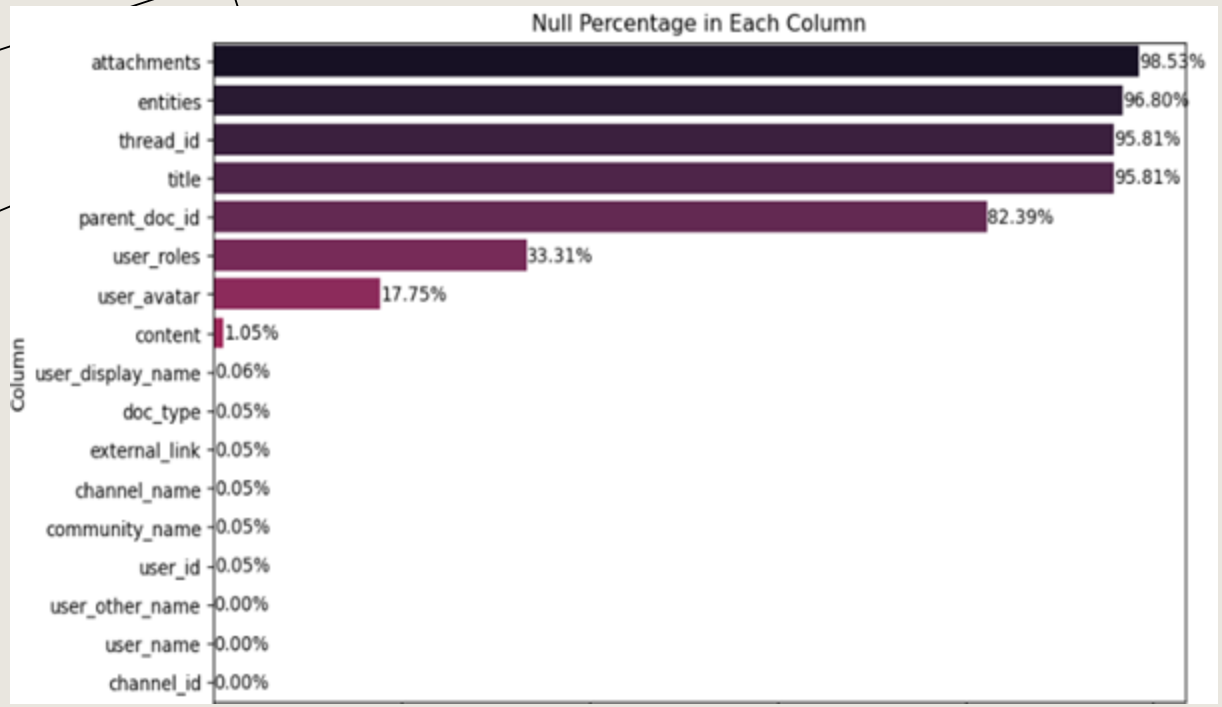
PERCENTAGE OF NULL VALUES FOR EACH COLUMN

Column	Null Value percentage (Discord)
attachments	98.53
entities	96.80
thread_id	95.80
title	95.80
parent_doc_id	82.39
user_roles	33.30
user_avatar	17.75
content	1.04
user_display_name	0.05
doc_type	0.05
external_link	0.05
channel_name	0.05
community_name	0.05
user_id	0.05
user_other_name	0.002
user_name	0.002
channel_id	0.000051

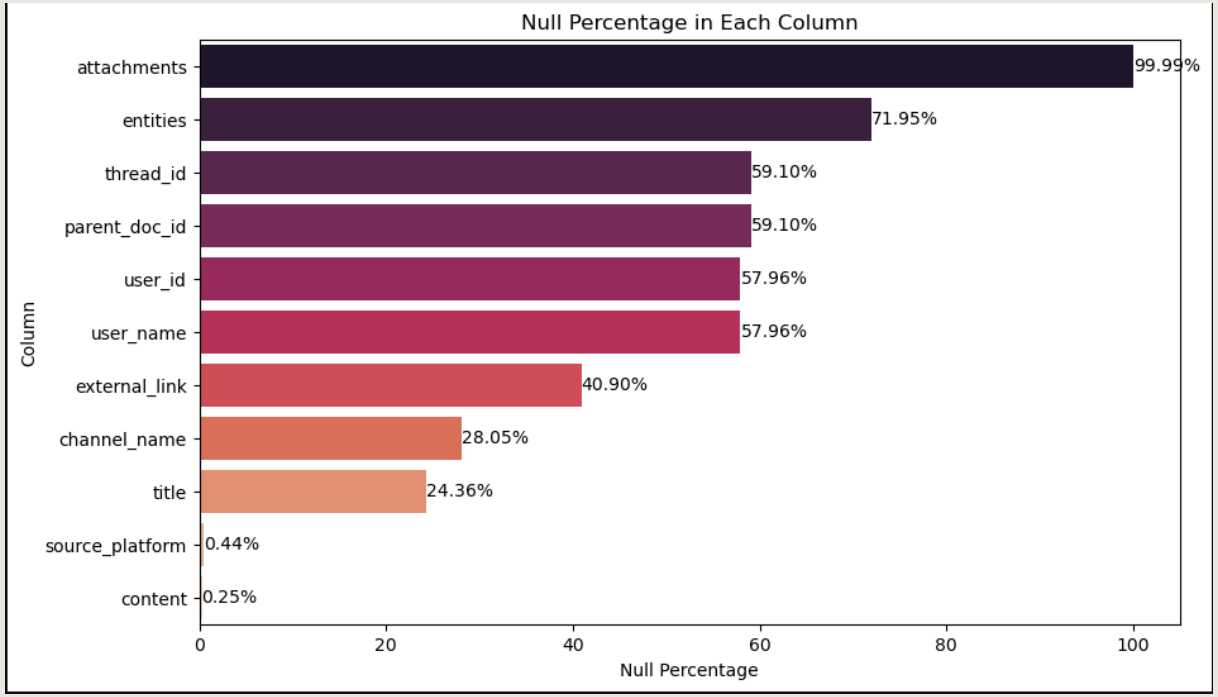
Column	Null Value percentage (Diablo)
attachments	99.99
entities	71.94
thread_id	95.80
title	24.35
parent_doc_id	59.10
content	0.25
external_link	40.89
channel_name	28.05
user_id	57.95
user_name	57.95
channel_id	0.000051

VISUAL REPRESENTATION OF NULL VALUES COLUMN WISE

Discord



Diablo



PERCENTAGE OF NULL VALUES BY DOCUMENT TYPE

DISCORD

doctype	username	content	user avatar	user roles	Parent Doc Id	Attachments	title	thread_id	entities
forum	0.02%	3.68%	51.48%	81.35%	8.19%	99.20%	81.57%	81.57%	82.05%
news	0%	0.07%	0.03%	21.51%	95.31%	90.16%	98.98%	98.98%	100%
public thread	0%	1.73%	100%	100%	0.99%	100.%	98.88%	98.88%	99.38%
stage voice	0%	1.11%	0%	10%	100%	100%	100%	100%	100%
text	0%	0%	10.44%	22.84%	98.50%	98.41%	98.89%	98.89%	100%
voice	0%	0.40%	0%	4.53%	100%	98.80%	100%	100%	100%

PERCENTAGE OF NULL VALUES BY DOCUMENT TYPE

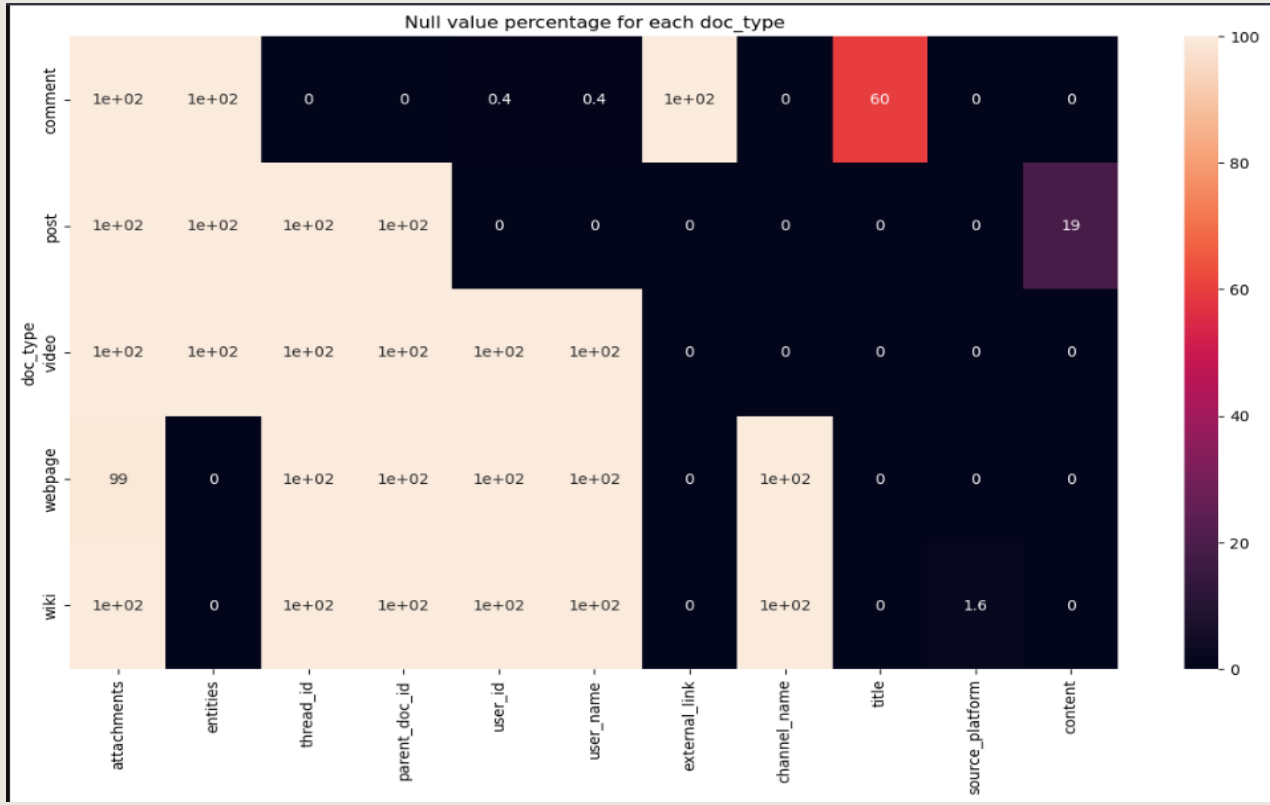
DIABLO

Doc Type	Attachments	Entities	Thread id	Parent doc id	Use Id	Username	External link	Channel name	title	Source platform	Content
comment	100%	100%	0%	0%	0.4%	0.4%	100%	0%	59%	0%	0%
post	100%	100%	100%	100%	0%	0%	0%	0%	0%	0%	19.11%
video	100%	100%	100%	100%	100%	100%	0%	0%	0%	0%	0%
webpage	99%	0%	100%	100%	100%	100%	0%	100%	0%	0%	0%
wiki	100%	0%	100%	100%	100%	100%	0%	100%	0%	1.63%	0%

VISUAL REPRESENTATION OF NULL VALUES

Discord

Diablo



DOCUMENT ID VALIDATION

Discord

Total Documents	Documents with Valid ids	Documents with invalid ids
19,67,974	19,67,974	0

Diablo

Total Documents	Documents with Valid ids	Documents with invalid ids
2,38,000	0	2,38,000

SAMPLE OF ID'S

Discord

```
id:uds:uds::10977554884368876 42
id:uds:uds::10792221350017106 62
id:uds:uds::10939558261730959 90
id:uds:uds::11027864170521026 66
id:uds:uds::10929561779330212 84
                ...
id:uds:uds::10918982576276767 62
id:uds:uds::11020029007812485 64
id:uds:uds::10964566661086699 92
id:uds:uds::10941999873924997 93
id:uds:uds::11013121078253445 22
```

Diablo

```
a4e4f17c8db45c9d50b53e80efbd77f0236f31995866f53936af503dddbd0b8a
72e0b30245f230156a3a98d9c57206e6e5521689c155520beb0de8a62d3f34aa
5fb5b906cf7527704267b95f175277195a6c9cd36d524144c2e45f57ea5ad4d0
264ed75c4870502b2ab92720a0115bf80450c17badc97a1e4573669daee90659
t1_jm7ry6e
efbbeff2a84b8b6814a2222101acaf8ea9b32edd15f7e78b54b3796141c3f240
t1_jnfzad3
t1_jm8dsmp
t1_jmxtifh
t1_jl68xo5
```

DOCUMENT TYPE VALIDATION

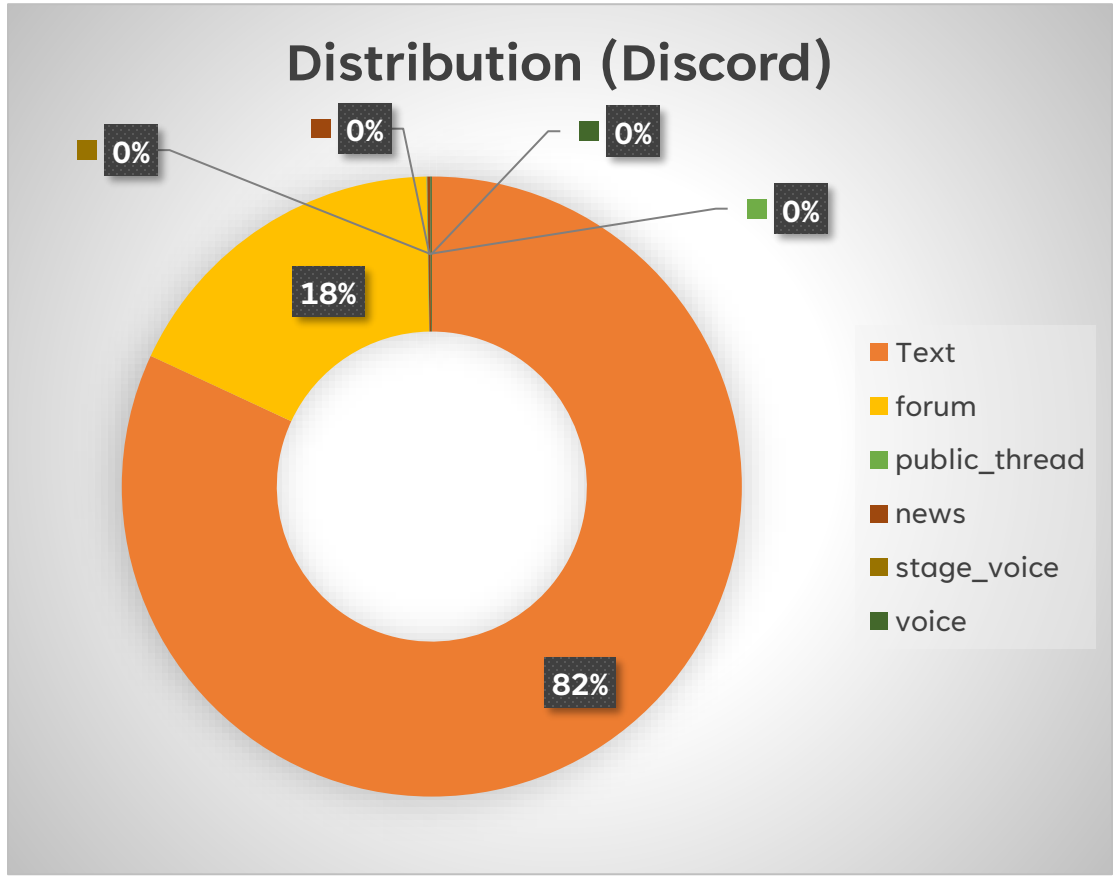
Set Of Valid Document Types (Discord)

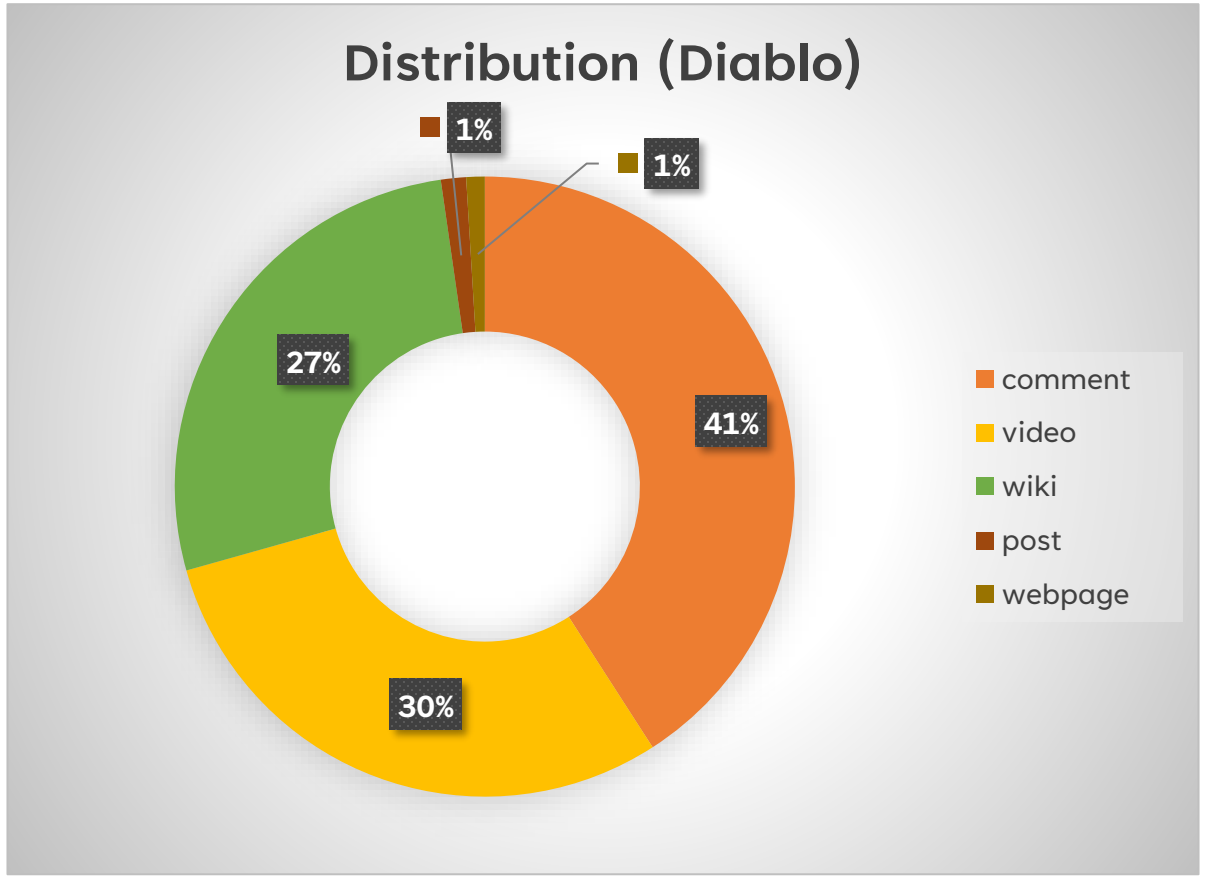
- News
- text
- public thread
- stage voice
- voice
- Forum

Set Of Valid Document Types (Diablo)

- Comment
- Video
- Wiki
- Post
- Webpage

DOC TYPE PERCENTAGE DISTRIBUTION





ENTITIES VALIDATION

Discord

In this column 96.80% of values are Null

The total Invalid Entities is 97.87%

Total Valid Entities are 3.2%

Diablo

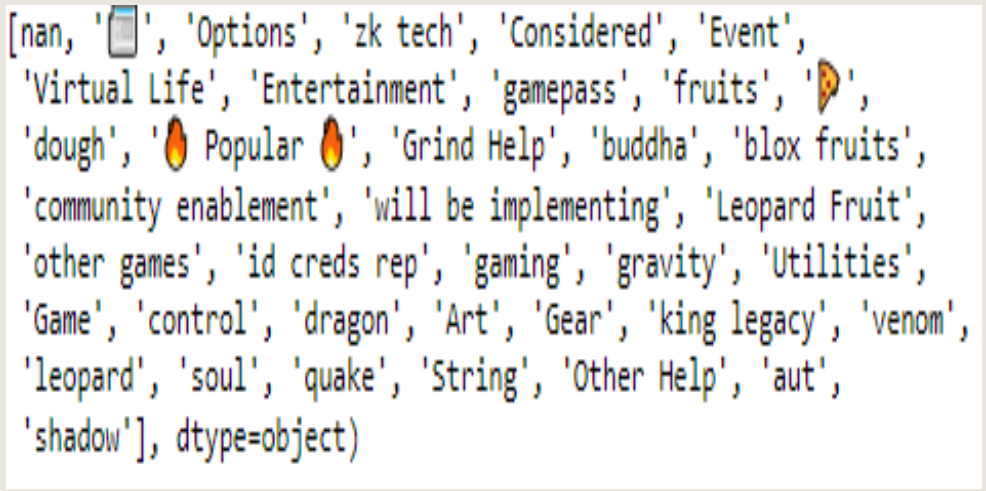
In this column 71.95% of values are Null

The total Invalid Entities is 72.41%

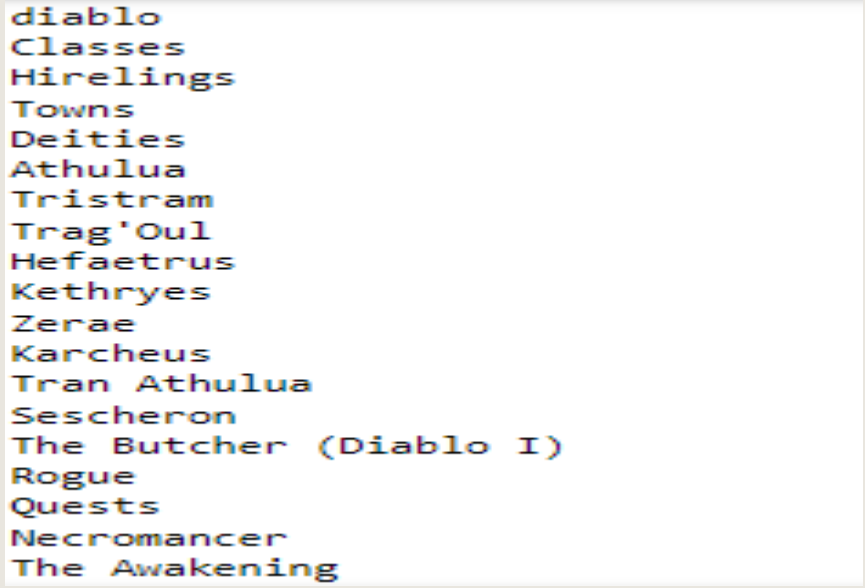
Total Valid Entities are 28.05%

SAMPLE OF ENTITIES

Discord

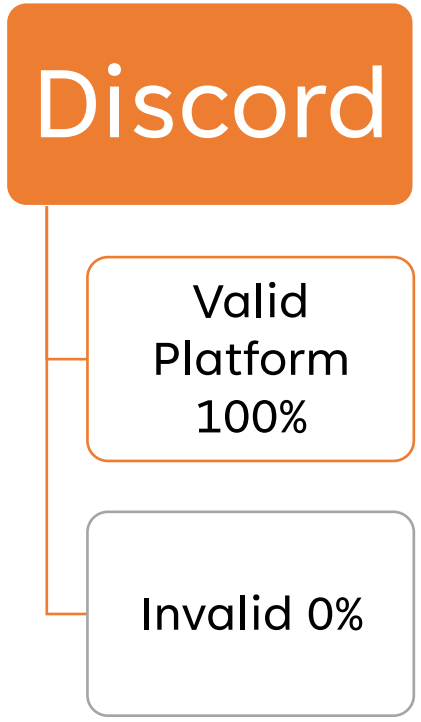


Diablo

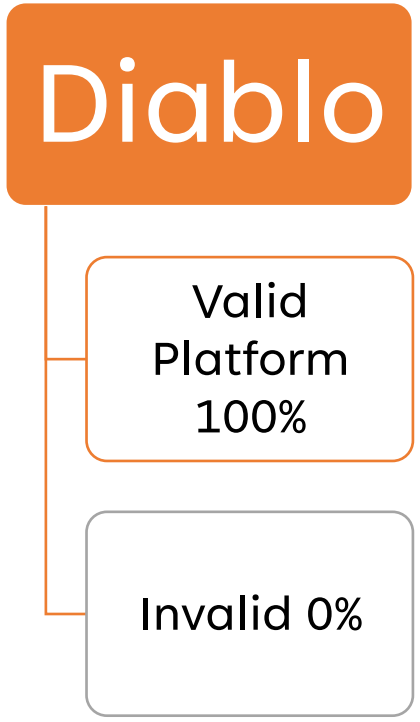


SOURCE PLATFORM VALIDATION

There is only 1 Platform Type Discord



Platform Type Diablo
fandom, official, max roll, Reddit, YouTube



TIMESTAMP VALIDATION

Discord

01

In this specific field, there are no null values.

02

In this validation, we have ensured that the dates in both the "created_at" and "created_at_str"

fields match and that the dates are not greater than today's date.

03

The total number of documents is 1,967,974, and all of them are considered valid.

Diablo

01

In this specific field, there are no null values.

02

In this validation, we have ensured that the dates in both the "created_at" and "created_at_str"

fields match and that the dates are not greater than today's date.

03

The total number of documents is 2,38,111.

- All the dates are earlier than today's date
- Among these documents, 191,882 of them exhibit matching dates in the field labeled "created_at_st,"
- while the remaining 46,229 documents do not align with the date provided in the "created_at_st" field.

SAMPLE OF TIMESTAMP

Discord

	created_at	created_at_str
0	2023-04-18 05:28:23	2023-04-18T05:28:23.754000+00:00
1	2023-02-26 02:03:27	2023-02-26T02:03:27.962000+00:00
2	2023-04-07 17:49:53	2023-04-07T17:49:53.676000+00:00
3	2023-05-02 02:39:30	2023-05-02 02:39:30.606000+00:00
4	2023-04-04 23:37:38	2023-04-04T23:37:38.963000+00:00
...	...	...
1967969	2023-04-02 01:33:51	2023-04-02T01:33:51.107000+00:00
1967970	2023-04-29 22:46:05	2023-04-29 22:46:05.770000+00:00
1967971	2023-04-14 15:27:20	2023-04-14 15:27:20.380000+00:00
1967972	2023-04-08 10:00:06	2023-04-08T10:00:06.247000+00:00
1967973	2023-04-28 01:01:10	2023-04-28 01:01:10.695101+00:00

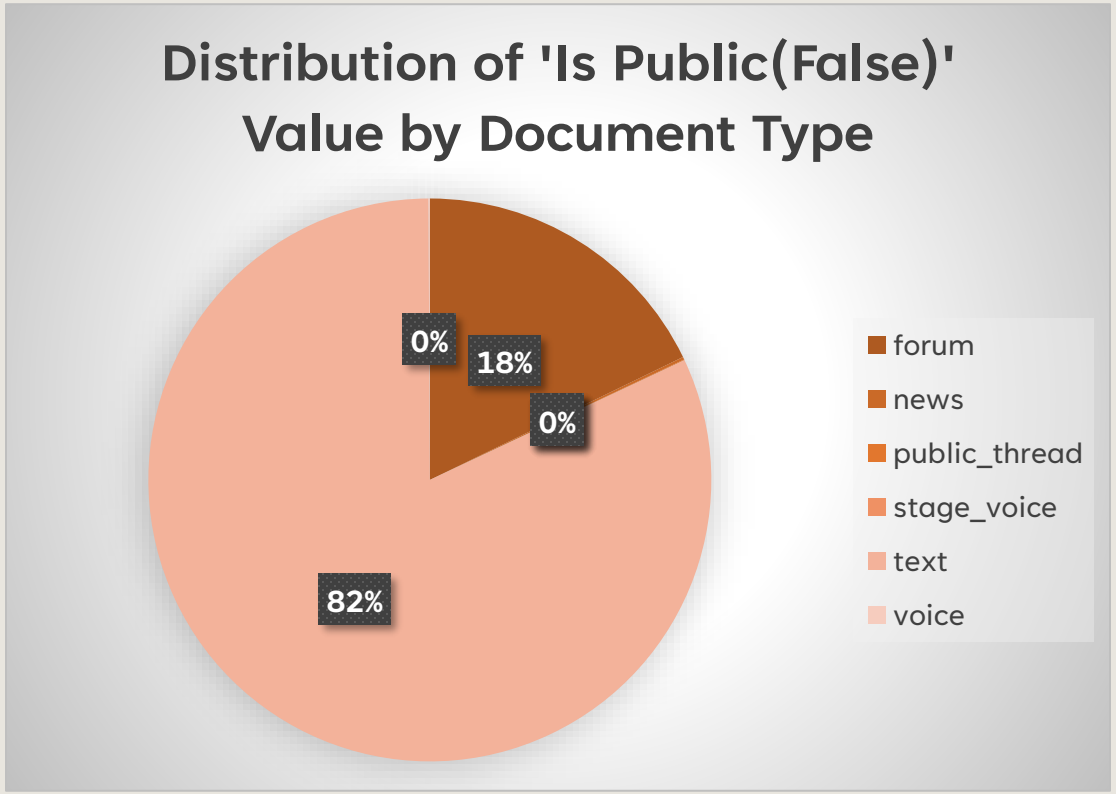
Diablo

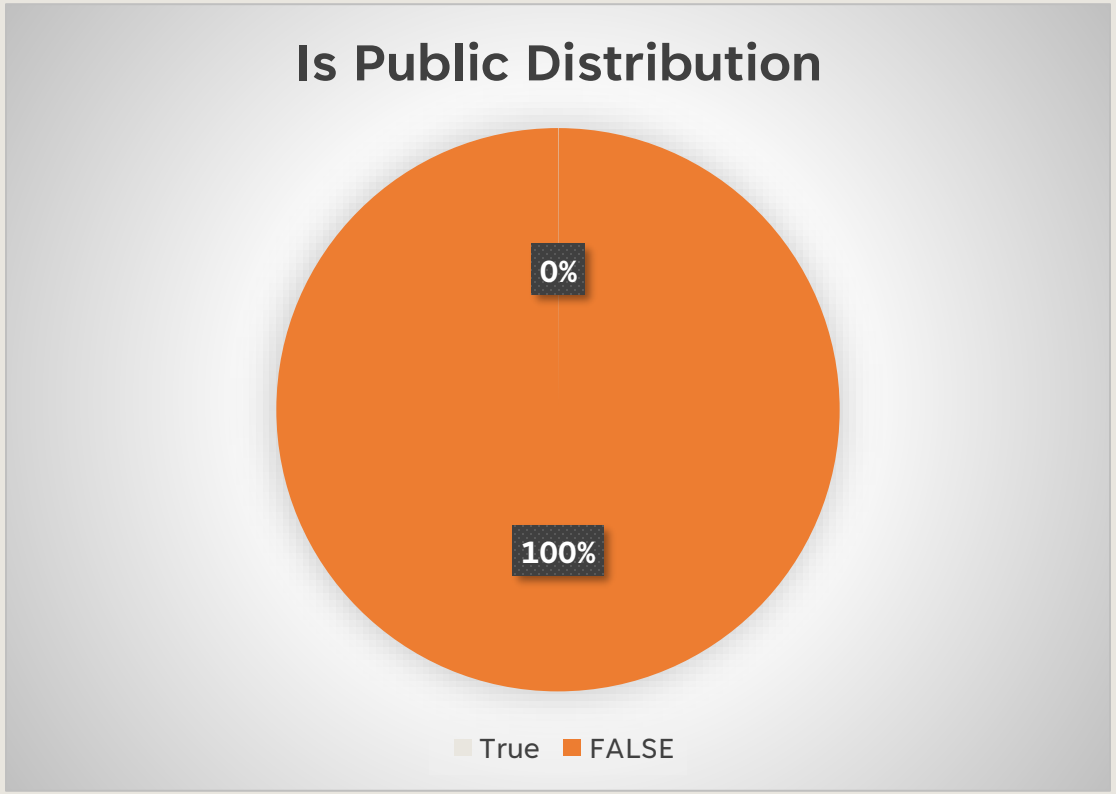
	created_at	created_at_str
65675	2023-05-15 18:45:14	2023-05-16 00:15:14
65680	2023-05-15 18:30:24	2023-05-16 00:00:24
65687	2023-05-15 18:47:45	2023-05-16 00:17:45
65688	2023-05-15 18:47:45	2023-05-16 00:17:45
65693	2023-05-15 22:13:31	2023-05-16 03:43:31
...	...	...
236362	2023-06-08 19:40:07	2023-06-09T01:10:07Z
236363	2023-06-08 19:40:07	2023-06-09T01:10:07Z
236364	2023-06-08 19:40:07	2023-06-09T01:10:07Z
236365	2023-06-08 19:40:07	2023-06-09T01:10:07Z
236366	2023-06-08 19:40:07	2023-06-09T01:10:07Z

IS PUBLIC VALIDATION

Total document	•19,67,974
True	•626
False	•19,67,348

THE DISTRIBUTION OF THE "IS PUBLIC" ATTRIBUTE BASED ON DOCUMENT TYPE?





TITLE CASE VALIDATION

Discord

Total Document	82,470 (Non-null)
Valid Title	59.19%
Invalid Title	40.81%

Diablo

Total Document	1,80,115 (Non-null)
Valid Title	93.51%
Invalid Title	6.49%

SAMPLE OF TITLE CASE VALIDATION

Discord

```
'will give love for dough king',
'venom,dough,paw and spider',
'budha portal rumble and love',
'trading rumble and venom',
'leopard for trade',
'trading leopard and spirit',
'thuat forum',
'helping trials or finding mirage island',
'who has 2x mastery',
'looking for stat reset',
'anyone willing to help me with dough king for v4
'i have shadow and gravity trading it for control
```

Diablo

```
' In-game',
' Quotes',
'"early release" 6pm CDT.',
'#flatSanctuary gang where you at?',
'+12 skill boots',
'10 COSAS QUE DEBES SABER ANTES DE JUGAR DIABLO IV',
'10 Tips You NEED Before Playing Diablo 4',
'100% Damage Reduction Challenging Shout?',
'11 Tipps die ich gerne früher gewusst hätte - Diablo 4 Tipps
'11 years later…',
"13 CRUCIAL Tips For Starting Diablo 4, Don't Waste Hours",
'15 Things To Know BEFORE YOU BUY Diablo 4',
'15 Tips EVERY Diablo 4 Player Should Know',
```

SPAM DOCUMENT DETECTION CONTENT

Discord

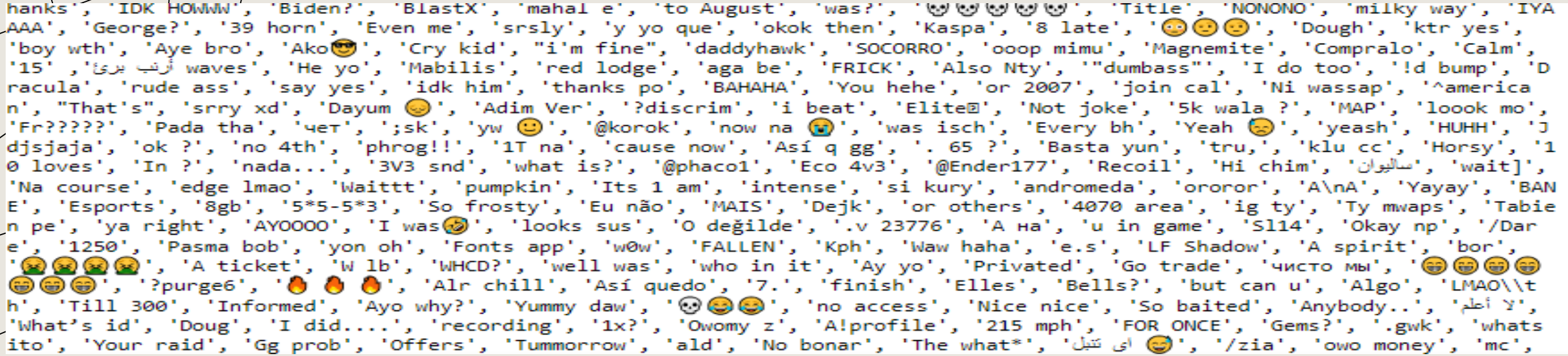
- Total Document
  - 19,67,974
- Valid Title count
  - 13,09,995 (67%)
- Invalid Title count
  - 6,57,979 (33%)

Diablo

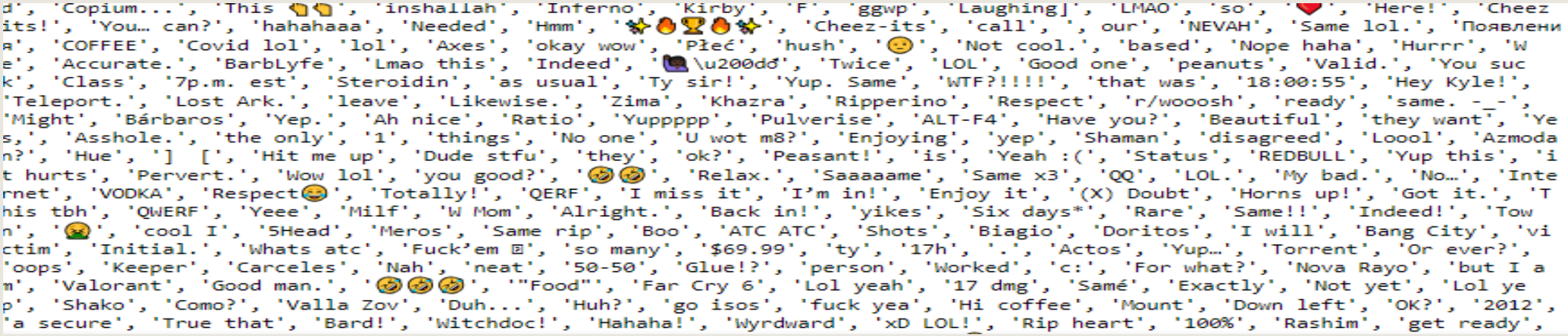
- Total Documents
  - 2,28,111
- Valid Title count
  - 1,63,971 (68%)
- Invalid Title count
  - 74,140 (31%)

SAMPLE OF SPAM DOCUMENT DETECTION CONTENT

Discord



Diablo



SPAM DOCUMENT DETECTION TITLE

Discord			Diablo		
Total Document	Valid Title count	Invalid Title count	Total Documents	Valid Title count	Invalid Title count
19,67,974	67,365 (3%)	19,00,609 (97%)	2,28,111	1,74,887 (73%)	63,224 (27%)

SAMPLE OF SPAM DOCUMENT DETECTION TITLE

Discord

F Quake', 'My shop', 'LF magma', 'Raids', 'Offers??', 'phoenix', 'Vouches', 'Title', 'Have Leo', 'inv', 'priv sv', 'Uniramp', 'shut up', 'thuat pro', 'Perm Bomb', 'butterfly', 'lf venom', 'LF Huge', 'e', 'dmmm', 'Lf yoru', 'lf dragon', 'Tutorial', 'Dough', 'trding', 'Rumble', 'knife', 'Awesome', '!!', 'Aut', '#verify', 'tupperbox', 'LF rumble', 'LF:Yoru', 'anyone ?', 'Latin', 'offers???', 'LF: LEO', 'looking', 'BUDDHA', 'LF OFFER', 'LF Dough', 'Noobs', 'LF Buddah', 'trades !', 'V4', 'Plato', 'Notifier', 'Physics', 'lf lepord', '🐆', 'LEOPARD', 'Ck.reward', 'Trading!', 'Athread', 'help', 'Tradibg', 'a', 'TRADE!!!!', 'Trade up', 'Control', '333', 'Pick', 'Ahhh', 'OFFER!!', 'helping', 'TRAIDING', 'LF Light', 'LF SPIRIT', 'Catalyst', 'gravity', 'view 2', 'dm', 'ttt 1', 'for what?', 'lf spirit', 'Netflix', 'یادگست', 'trading.', 'Tarding', 'OFFERS', 'Lf spirit', 'ChainML', 'LF psx', 'bloxfruit', 'Lf buddha', 'LF yuro', 'offerrr', 'chqma', 'join me', 'ada', 'pls', 'cool', '🦁', 'gp', 'LF dought', 'Ababa 1', 'LF:DOUGH', 'lf indra', 'gamepass', 'Tassets', 'FFR 2', 'pizza', 'Join', 'OMG OMG', 'memes', 'FARM', 'trading #', '#hi', 'Lf light', 'LF Portal', 'UPDATE #2', 'LF venom', 'Lf perms', 'Offer', 'need help', 'Portal', 'Lf', 'venom', '1%er', 'pop', 'WWTBAOSC', '🛏️📚👓', ':/', 'huhu 111', 'trade 😈⚠', 'Japanese', 'Gameplay💻', 'LF:OFFERS', 'Ñ', 'LF PERMS', '#Trade', 'Afrikaans', 'new', '#', 'LK dough', '📕-чат-3', 'view 1', 'i n gp +1', 'guild', '☠', 'Offer cuz', 'Serotonin', 'Spanish', 'nnnnnn', 'LF: Dough', 'Biology', 'Forum 3', 'bbn', 'ye', 'ok', 'Eleos Lab', 'test', 'TRADE LEO', 'news', 'CONTROL', 'Bread', 'trade pls', 'lf shadow', 'perm ice', 'of dough', 'need perm', 'pls dough', 'lf dough', 'lf rumble', 'Rehehehe', 'tthread 2', 'k', 'Temu hel

Diablo

'Finished', 'War Fork', 'Inferno', 'Mundunugu', 'Blaze', 'Harstead', 'Razor Bow', 'Galactic', '塔·拉曼', 'Крылья', 'Atravesar', 'Jonas', 'Earthbind', 'Caligio', 'Frostbite', 'Arco hoja', 'Cold Snap', 'Codex', 'Valla', 'Pole Axe', 'Porcupine', 'Foulwing', 'Cragworm', 'Razortine', 'Golems', 'Freeze', 'Zov', 'Axes', 'Arcstone', 'Arthur', 'Quietus', 'Shumbeel', 'Scourge', 'Hoplon', 'Strong', 'Esadora', 'Leaper', 'Harbinger', 'Alchemist', 'Peth', 'Adrahau', 'Haches', 'Харроrат', 'Vampire', 'Shroud', 'Slaughter', 'Marteks', 'Fire Hit', '暗黑破坏神', 'Cobalt', 'Lich Lich', 'Teffeney', 'Sucker', 'Tomahawk', 'Maw Axe', 'Riesgo', 'Twin Seas', 'Honor', 'Old Diary', 'Staalgard', 'Mydas', 'Viridian', 'Rani Oran', 'Slash', 'Lysander', 'Scrimshaw', 'Sammash', 'Khazra', 'Merendi', 'Hamit', 'Edyrem', 'Gull', 'Гризвольд', 'Razorclaw', 'Yumi', 'PP', 'Shingo', 'Might', 'Cutlass', 'Delsere', 'Słownik L', 'Bárbaros', 'Great Axe', "Rhau'Kye", 'Infoboksy', 'Собор', 'Mojo', 'Skorn', 'Meditator', 'Zaboul', 'Yoon', 'Baliste', 'Terrene', 'Стихии', 'Gheed', 'Grim Ward', 'Torr', 'Animals', 'Liv Moore', 'Воин', 'Empaleur', 'Andariel', 'Hyadures', 'Assur', 'Shaman', 'Trejiak', 'Ponzi', 'Standoff', 'Arcos', 'Rethald', 'Damius', 'Low Hills', 'Kerykeion', 'Reiter', 'Ekthul', 'Hakkara', 'Caduceus', 'Thrum', 'Korbal', 'Sampha', 'Elora', "Gro'Mag", 'Basilard', 'Yew Wand', 'Psychorb', 'Tin', 'Zhou Lore', 'Askari', 'Kingsport', 'Лилит', 'Shaftstop', 'Boosenian', 'Lava Lord', 'Jackal', 'Woh', 'Pet', 'Twin Axe',

TITLE AND CONTENT VALIDATION(DISCORD)

Missing or Empty Values

- Title Missing or Empty: 1,885,504(95.81%) documents
- Content Missing or Empty: 20,602(1.05%) documents

Title Less Than 3 Tokens

- Title non-null values: 82,470
  - 1 token: 14,148 (17%) documents
  - 2 tokens: 11,793(14%) documents
- Content
  - 1 token: 555,182(28%) documents
  - 2 tokens: 255,502(13%) documents

Has Only Numbers

- Title Has Only Numbers: 5 documents
- Content Has Only Numbers: 108 documents

Has Only Special Characters

- Title Has Only Special Characters: 158 documents
- Content Has Only Special Characters: 992 documents

SAMPLE OF TITLE AND CONTENT VALIDATION(DISCORD)

Less then 3 Token(Title)

	title	count
0	hehe	5397
1	BBois	3225
2	trading	903
3	trading dough	533
4	ThetaTom's BBBBQ	447
...	...	...
2226	TRADING THIS:→	1
2227	😡DRAGON😡	1
2228	TRADING POSTS	1
2229	😡DOUGH😡	1
2230	test 345	1

Less than 3 Token(Content)

	content	count
0	$wa	10866
1	.bake	10174
2	.harvest	9855
3	bal	9767
4	.stealcookie	9086
...	...	...
213949	Only 2?	1
213950	https://tenor.com/view/over-and-over-again-cre...	1
213951	rolled	1
213952	join back	1
213953	https://tenor.com/view/omori-gif-25514678	1

213954 rows × 2 columns

Have special character and number only(Title)

	title	count
0	سوالات	51
1	.	33
2	تداعیهای شما چیست؟	10
3	..	5
4	تصویرگری	4
5	...	3
6	Trading	3
7	익헨 리스트	3
8	:/	2
9	Ñ	2
10	스레드	2

Have special character and number only(content)

	content	count
0	?	4360
1	💀	3038
2	.	2833
3	😭	1458
4	...	839
...	...	...
19744	عجب جمله ای گفتی	1
19745	мрр~	1
19746	سال کبری ای بود	1
19747	اینجا هم معامله بوده با ایرون جان	1
19748	...به نظرم با فضای زفوس آپلویی نمیشه همدلی. خوبی	1

19749 rows × 2 columns

TITLE AND CONTENT VALIDATION(DIABLO)

Missing or Empty Values

- Title Missing or Empty: 57,996(95.81%) documents
- Content Missing or Empty: 598(1.05%) documents

Title Less Than 3 Tokens

- Title non-null values: 1,80,115
  - 1 token: 5,665(3%) documents
  - 2 tokens: 16,344(9%) documents
- Content non-null values: 2,37,513
  - 1 token: 5,361(2%) documents
  - 2 tokens: 3,823(13%) documents

Has Only Numbers

- Title Has Only Numbers: 0 documents
- Content Has Only Numbers: 19 documents

Has Only Special Characters

- Title Has Only Special Characters: 1784 documents
- Content Has Only Special Characters: 1892 documents

SAMPLE OF TITLE AND CONTENT VALIDATION(DIABLO)

Less then 3 Token(Title)

	title	count
0	Real disappointed...	561
1	Seasonal Renown	404
2	Login issues?	350
3	Minecraft Dungeons	136
4	500k~ exp/2mins	123
...	...	...
14106	Chantodo's Resolve	1
14107	Chantodo	1
14108	Archmage	1
14109	Chantodo's Force	1
14110	Beyatt	1

14111 rows × 2 columns

Less than 3 Token(Content)

	content	count
0	foreign	532
1	[deleted]	258
2	thank you	251
3	Same	211
4	[	95
...	...	...
5265	Lmao.	1
5266	😂😂😂😂	1
5267	0$	1
5268	Easy pass	1
5269	Hey, it's	1

5270 rows × 2 columns

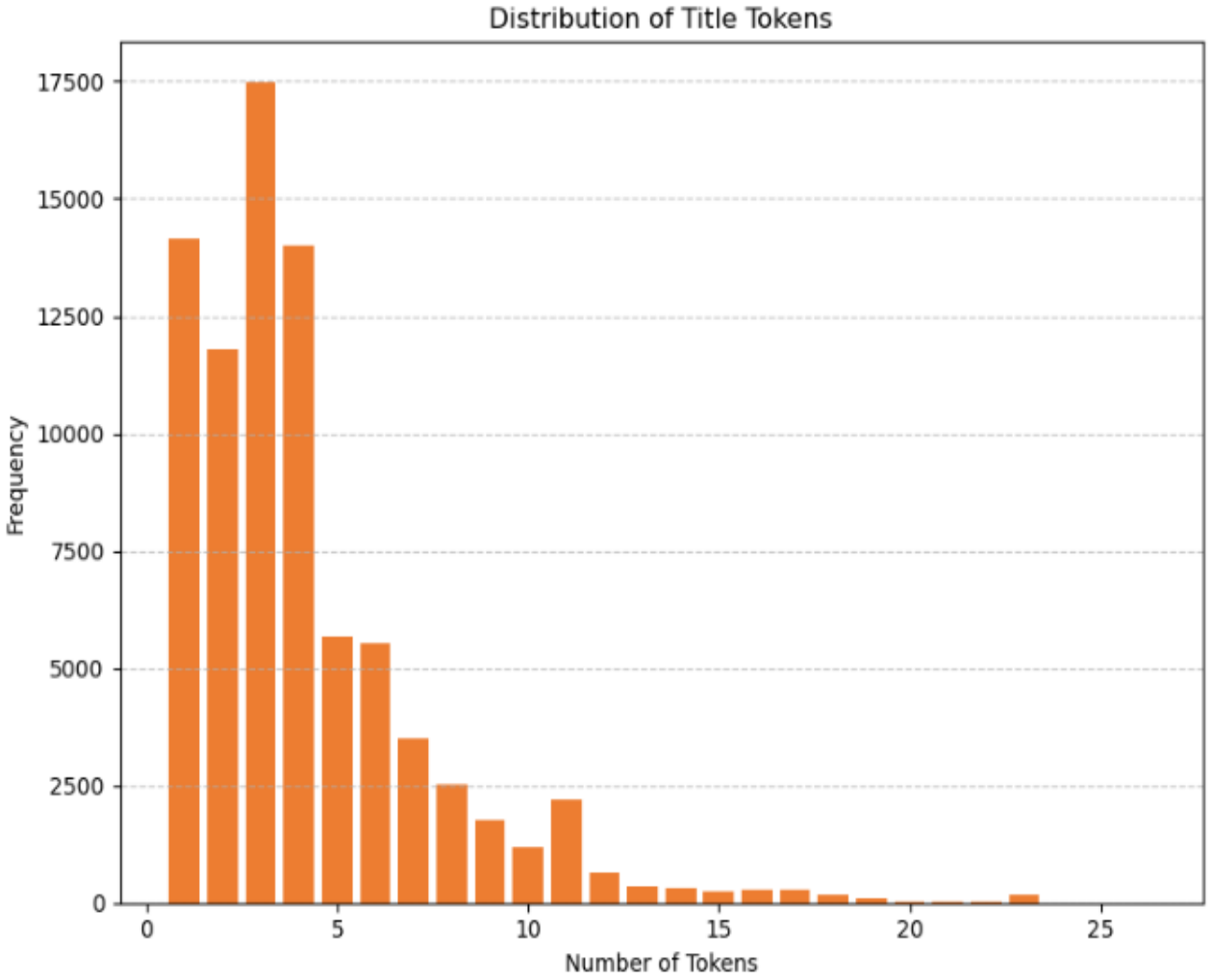
Have special character and number only(Title)

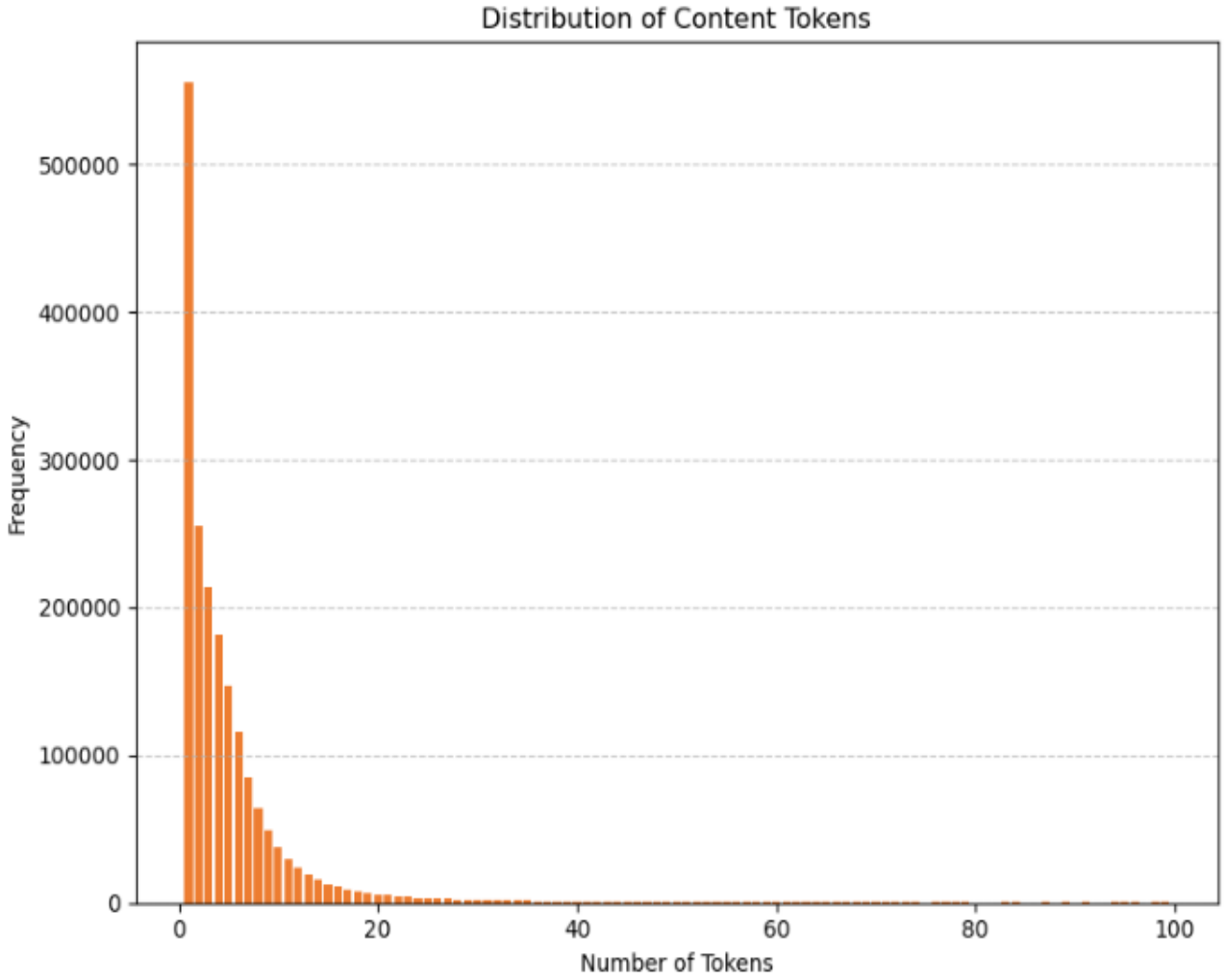
	title	count
0	Охотник на Демонов: Диалоги Кровь и Песок[]	47
1	Охотник на Демонов: Диалоги Гиблые поля[]	39
2	Хорадрический Куб Для любого арсенала[]	29
3	Кабраксис Эпизод с открытием Врат[]	13
4	Эйрина Описание[]	12
...	...	...
1041	Событие: Кристальная тюрьма	1
1042	Вывеска Огдана Пепин[]	1
1043	Событие: Души демонов Этапы события:[]	1
1044	Экзарх (дневник)	1
1045	🔥🔥🔥🔥	1

Have special character and number only(content)

	content	count
0	[	95
1	]	23
2	Появления	10
3	🤣	8
4	?	7
...	...	...
1771	моснтры Демоны-солдаты Путь ведет в Заоблачные...	1
1772	Перед смертью Леорик проклял всех, кто был ряд...	1
1773	лишь склонили головы в молчаливом согласии, от...	1
1774	Лахданан Титул\n\nЧемпион Закарума Пол\n\nМужской ...	1
1775	но без открытой войны, поскольку это могло бы ...	1

TOKEN DISTRIBUTION(DISCORD)





SAMPLE OF TOKEN DISTRIBUTION (DISCORD)

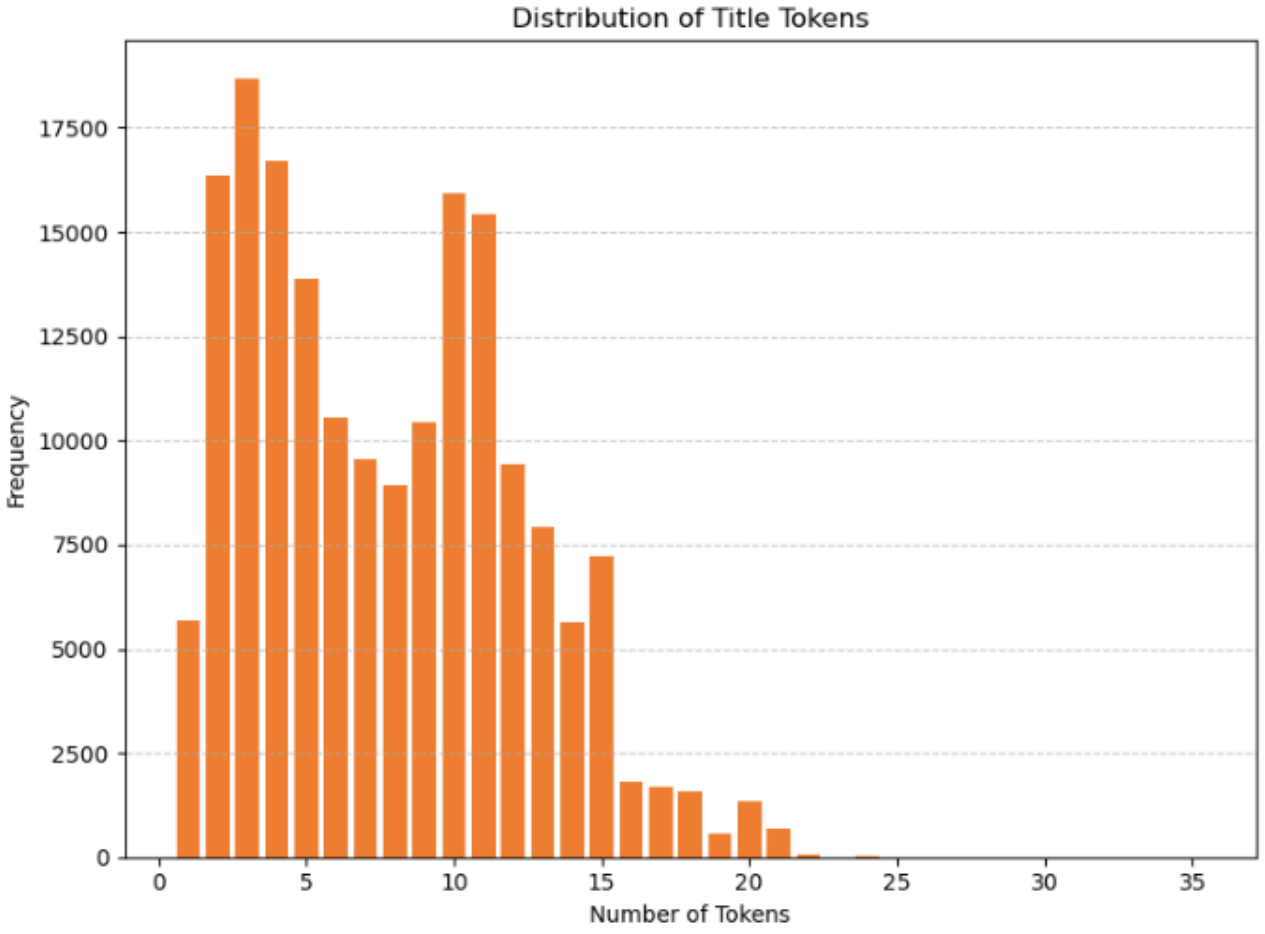
Title(Token==3)

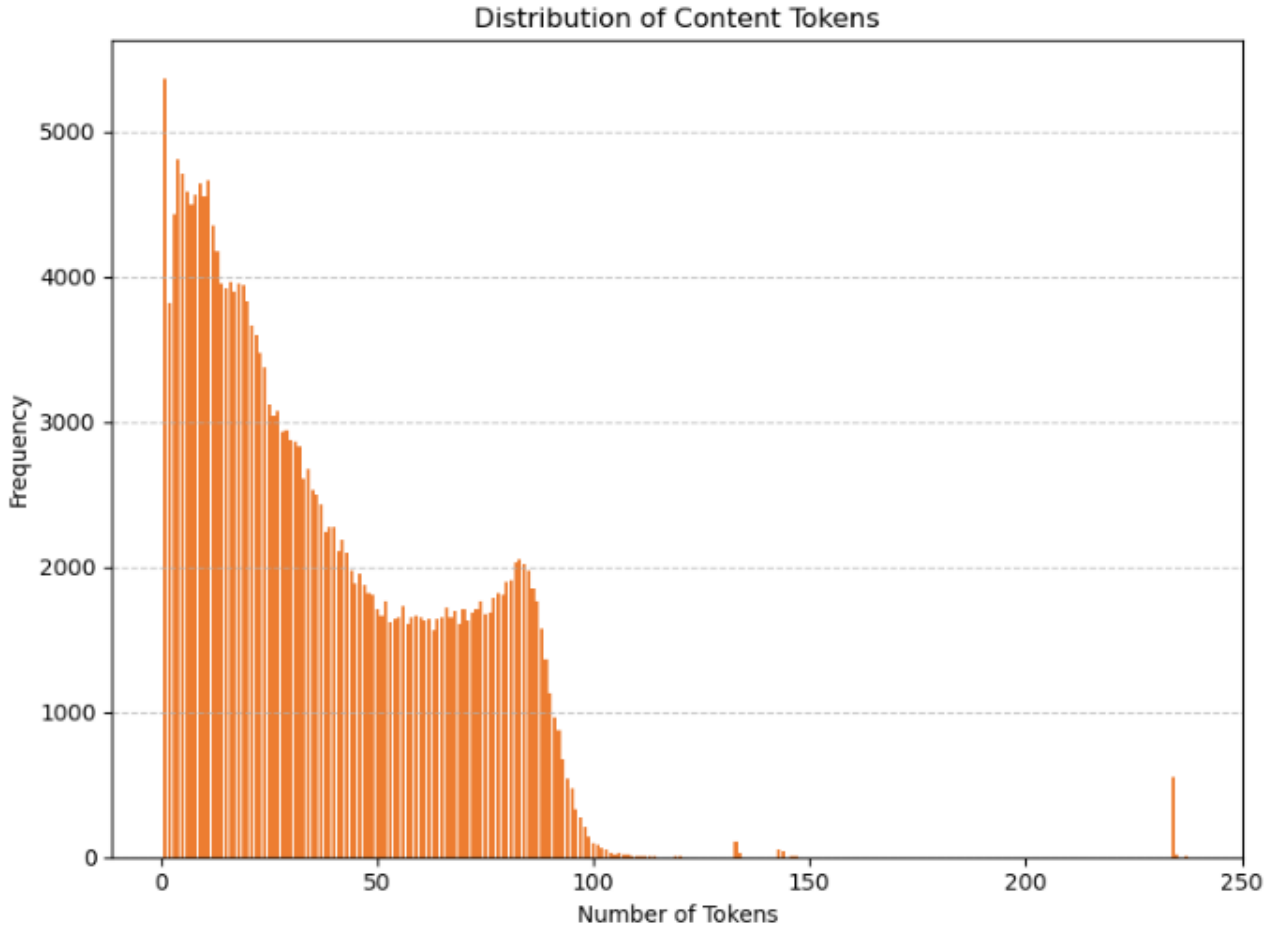
	title	count
0	Theta Wins Only	3347
1	Theta Galactic Empire	2533
2	Kolo's Future Korner	309
3	Balthius' Better Bazaar	291
4	trading my inv	163
...	...	...
2564	anyone trade budha	1
2565	Daily fact #89	1
2566	Trading two dough	1
2567	Trading Dough Offers?	1
2568	LF dough perm	1

Title(Token==1)

	content	count
0	$wa	10866
1	.bake	10174
2	.harvest	9855
3	bal	9767
4	.stealcookie	9086
...	...	...
93401	ñ	1
93402	Sing*	1
93403	HAHAHAHAHAAA	1
93404	really?!	1
93405	https://tenor.com/view/omori-gif-25514678	1

TOKEN DISTRIBUTION(DIABLO)





SAMPLE OF TOKEN DISTRIBUTION （DIABLO)

Title(Token==3)

	title	count
0	【Diablo 4】最強ローグでいく、ティア4トーメント攻略＆レベル上げやレジェンダリー稼ぎ...	608
1	Tips for leveling	438
2	Server's finally down	225
3	11 years later...	114
4	Server just crashed?	102
...	...	...
11643	Beast (Disambiguation) General	1
11644	Hawthorne Gable Trivia	1
11645	Horrus the Nightstalker	1
11646	Hrugowl the Defiant	1
11647	Demi Lich Stats	1

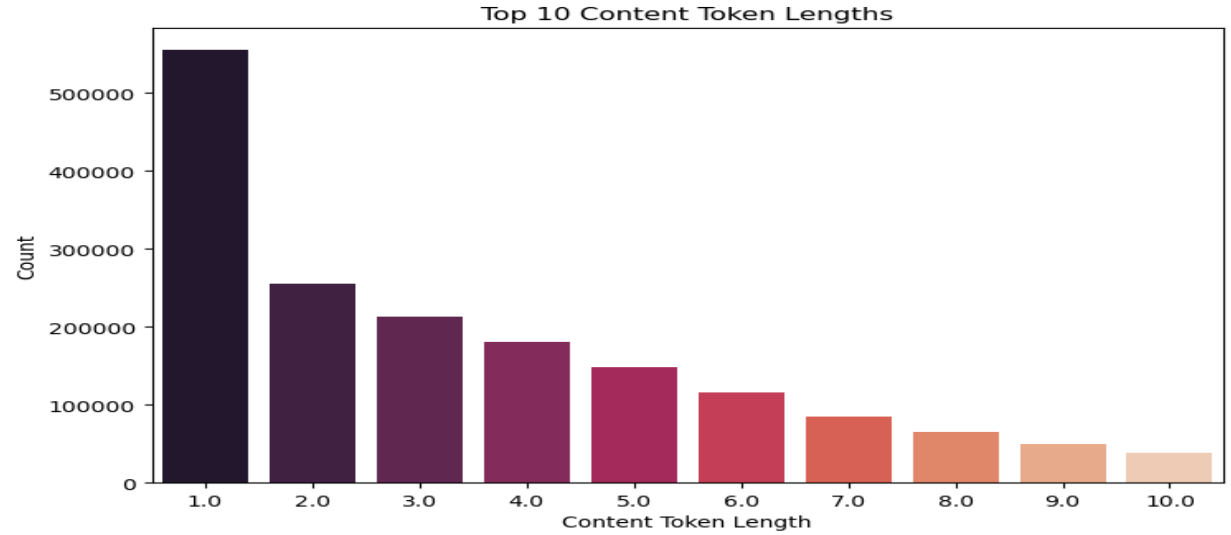
11648 rows × 2 columns

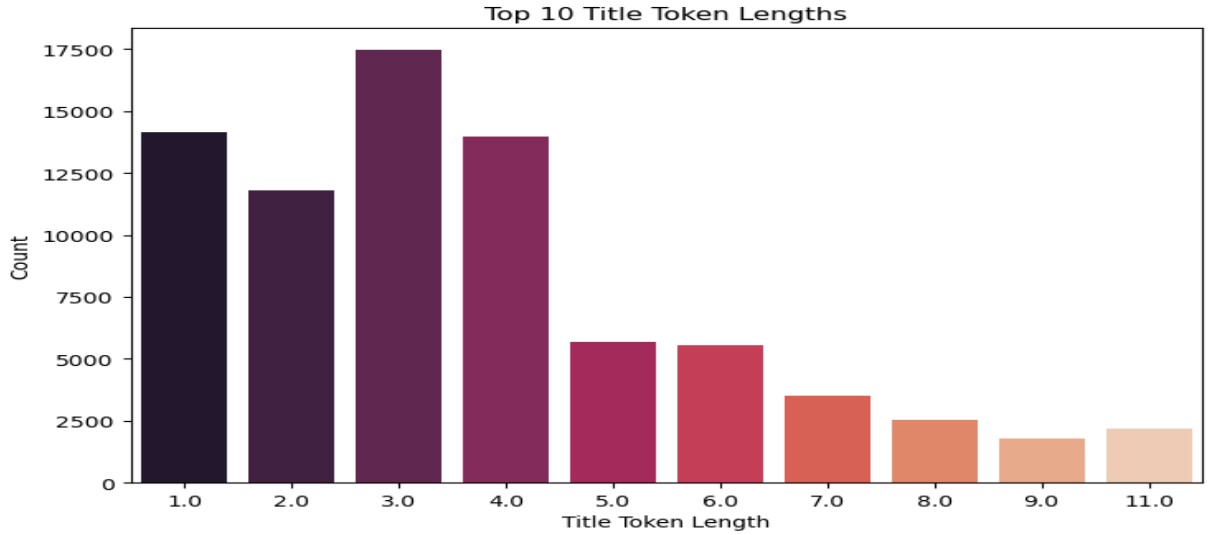
Title(Token==1)

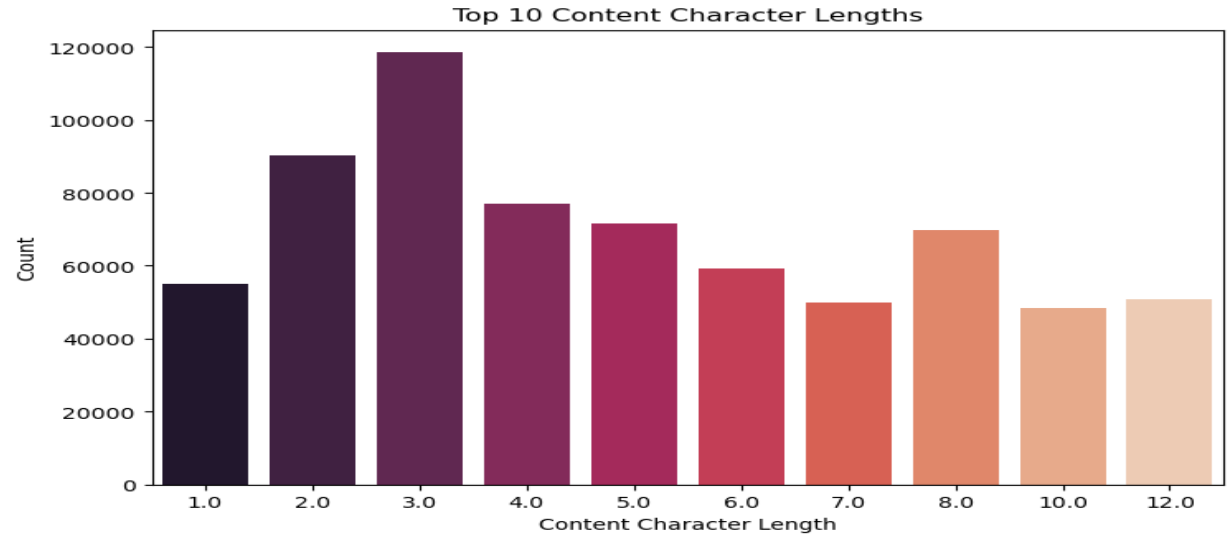
	content	count
0	foreign	532
1	[deleted]	258
2	Same	211
3	[	95
4	[removed]	91
...	...	...
2444	Yikes	1
2445	Facts!	1
2446	https://youtu.be/tAtalZD0Ebs	1
2447	Minging	1
2448	maybe	1

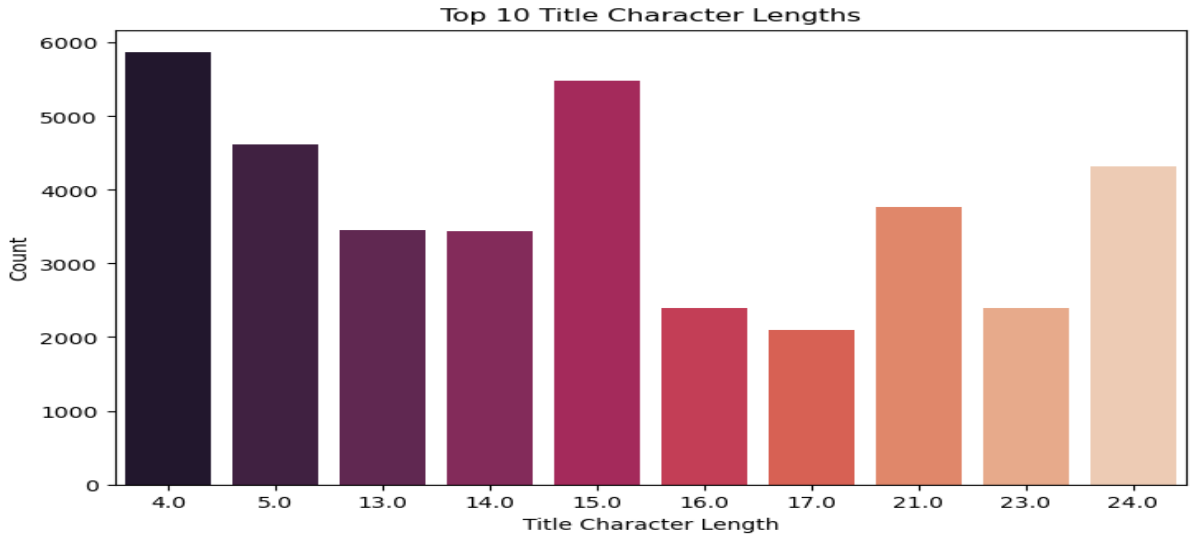
2449 rows × 2 columns

TOP 10 TOKENS AND CHARACTER LENGTH(DISCORD) (1 WORD=1TOKEN)

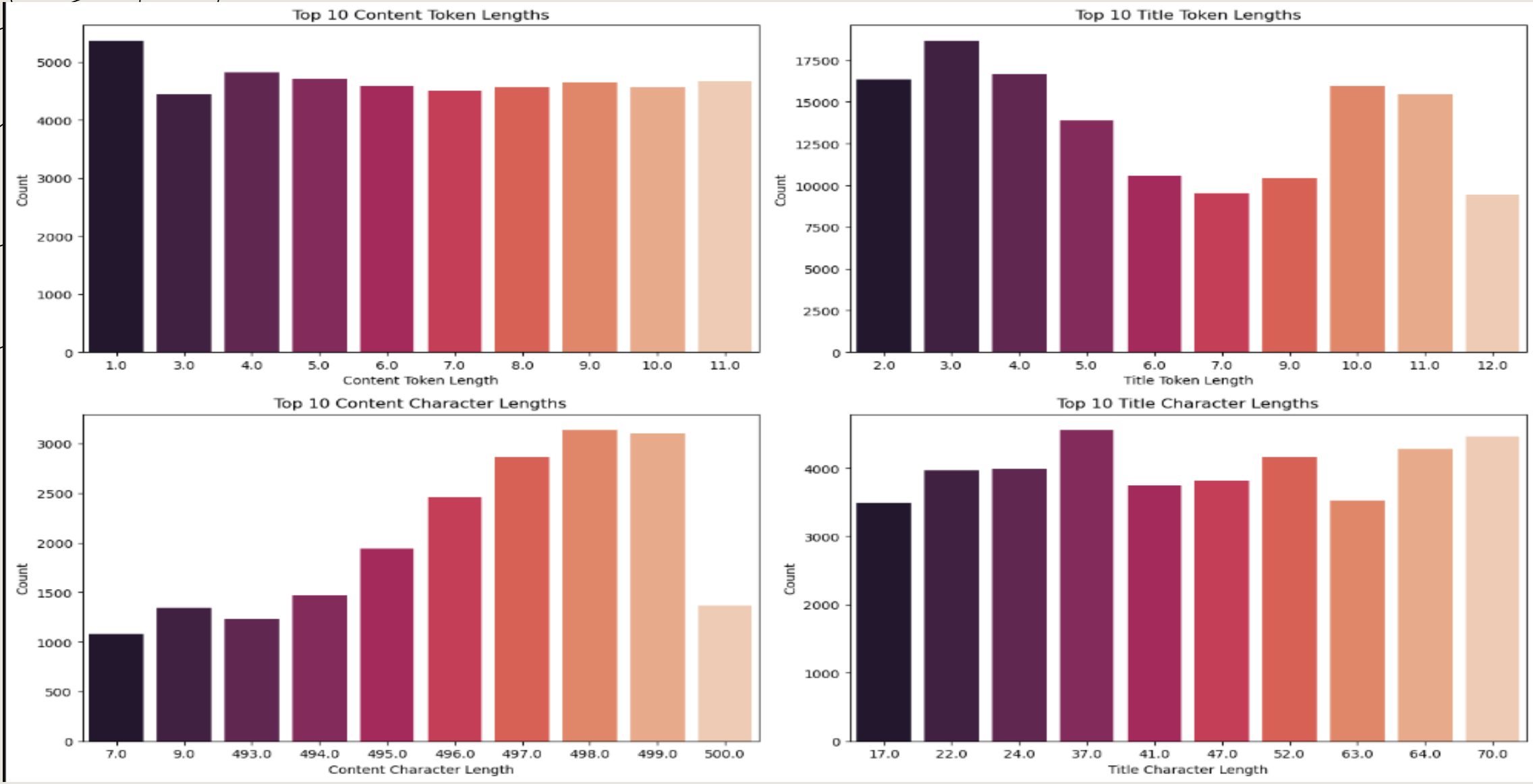








TOP 10 TOKENS AND CHARACTER LENGTH(DIABLO) (1 WORD=1TOKEN)



SUMMARY

In our comprehensive data analysis, we have identified and examined several critical aspects of the dataset, which is comprised of a total of 1,967,974 documents. Here are the key findings and insights:

Null Value Analysis

We observed varying degrees of null values across different columns, with some columns having a high null percentage. For instance, 'attachments,' 'entities,' and 'thread_id' columns have null percentages exceeding 95%.

Document Type Distribution

The dataset contains a variety of document types. Most prevalent are 'text' and 'forum' types, accounting for 81.92% and 17.80% of the dataset, respectively. There is also a small percentage of 'news,' 'voice,' 'public_thread,' and 'stage_voice' document types.

Document Validation

We validated documents based on their document type, identifying discrepancies in null value percentages among different types. Notably, 'forum' and 'news' types exhibit significantly different null value patterns.

Entity Analysis

A portion of the documents (3.2%) contains valid entities. Additionally, 2.13% of the documents have at least two entities, while the majority (97.87%) do not contain any entities.

Platform and Timestamp Validation

All documents have valid source platforms and timestamps, ensuring data integrity.

Is Public Attribute

The 'is_public' attribute has a true value for 626 documents, indicating public accessibility, while 1,967,348 documents have a false value, indicating non-public content.

Title and Content Quality

The dataset has 2.47% of documents with valid titles, while the majority (97.53%) have missing or inadequate titles. Additionally, some documents have content with fewer than three tokens, and a few contain titles or content with numbers or special characters.

Spam Detection

Among the documents, 657,979 are identified as potential spam.

THANK YOU

Shashank Shukla

shashanks.ven@splore.com