In [1]:
```
import json
import pandas as pd
import matplotlib as mp
# for language detection
from langdetect import detect, DetectorFactory
from langdetect.lang_detect_exception import LangDetectException
from pathlib import Path
import re
import matplotlib as plt
from matplotlib import pyplot
from sklearn.feature_extraction.text import CountVectorizer
import numpy as np
from urllib.parse import urlparse
from collections import Counter
from nltk.corpus import stopwords
import nltk
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\micha\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[1]: True

In [2]:
```
def repetitiveness_score(text, n=3):
    """
    Split text to n-grams, count duplicates and divided by total n-gram count
    """

    words = text.split()
    if len(words) < n:
        return 0.0
    ngrams = [' '.join(words[i:i+n]) for i in range(len(words)-n+1)]
    counts = Counter(ngrams)
    total = len(ngrams)
    repeated = sum(v for v in counts.values() if v > 1)
    return repeated / total

def count_html_tags(text):
    """
    simple regex to get a general sense of the amt of html tags in text
    """
    TAG_REGEX = re.compile(r"<\s*/?\s*([a-zA-Z0-9]+)[^>]*>") # general to match
    return len(TAG_REGEX.findall(text))

def count_non_utf8_chars(text):
    """
    Return count of characters which cant be encoded to utf8
    """
    count = 0
    for c in text:
        try:
            c.encode('utf-8')
        except UnicodeEncodeError:
            count += 1
    return count

def detect_language(text):
    try:
```

```
        return detect(text)
    except LangDetectException:
        return "Unknown"
```

Data import

In [3]:
```
filepath = "../../data/raw/mainpipe_data_v1.jsonl"
data = []

with open(filepath, 'r', encoding='utf-8') as file:
    for line in file:
        try:
            data.append(json.loads(line))
        except json.JSONDecodeError:
            continue

df = pd.DataFrame(data)
print(df)
```

```
                                                     text  \
0       In the never ending battle to rid Alaska of it...
1       » Jackpot | Deutsche Online Casinos und Casino...
2       This really was an unexpected pleasure. When I...
3       def files(self):\n        """Files in torrent....
4       Patient engagement in the design and delivery ...
...                                                   ...
269373  Our 1 to 1 Karting lessons are ideal to give y...
269374  function read(model) {\n  var query = argument...
269375  In a land that is already fragile with earthqu...
269376                                  Simple, YES on 8!
269377  <p>How would I be able to get N results for se...

                                                      url
0                                                    None
1         http://www.casinodeutsch.net/stichwort/jackpot/
2       http://leekat.booklikes.com/post/608842/an-une...
3       https://github.com/idlesign/torrentool/blob/78...
4                     http://www.nhlc-cnls.ca/sessions/3/
...                                                   ...
269373  https://midlandkarting.co.uk/go-karting-events...
269374  https://github.com/endpoints/endpoints/blob/1e...
269375                                               None
269376                                               None
269377                                               None

[269378 rows x 2 columns]
```

Data exploration

Initial noisy text data exploration

Initially for this task I wanted to look into whats normally considered noisy data for LLM preprocessing.

- Null/Na records
- Html tags
- Special and non-utf8 characters

In [4]:
```
print(f"The number of null text records is {int(df['text'].isnull().sum())}")
print(f"The number of null url records is {int(df['url'].isnull().sum())}")

# Looking into html tags
df["element_count"] = df["text"].apply(count_html_tags)
print(f"The approximate number of records containing html elements is {int(len(d

# Looking into non-utf8 rows
df["nonutf8_count"] = df["text"].apply(count_non_utf8_chars)
print(f"The approximate number of records containing non-utf8 characters is {int
```

```
The number of null text records is 0
The number of null url records is 90519
The approximate number of records containing html elements is 65807
The approximate number of records containing non-utf8 characters is 0
```

Dataset content exploration

I wanted an idea of the kind of content the datset contained, looking into sources trying to classify content.

- Data sources
- Languages
- Content (code vs formal english vs informal)

Notably github made up a large portion of the sources, the rest of the top 20 sources were mostly news websites. I expected the data to be a strong mix of informal language, formal and code. There is a large number of items with no url, more on this later.

In [5]:
```
# Using the base of the url to tell the source
df['base_url'] = df['url'].apply(lambda x: f"{urlparse(x).scheme}://{urlparse(x)
df['base_url'].value_counts().head(20)
```

Out[5]:
```
base_url
b'':b''                          90519
https://github.com               71415
https://www.taiwannews.com.tw     2583
https://en.wikipedia.org          2326
https://placeholder.co            2110
https://sample-company.net        1997
https://example.com               1991
https://testsite.org              1964
https://demo-page.info            1938
http://abcnews.go.com             1771
http://www.nigeriatoday.ng        1290
https://www.yahoo.com              946
https://www.nigeriatoday.ng        872
http://www.israelnationalnews.com  746
https://www.nytimes.com            735
http://www.wafb.com                650
http://uproxx.com                  624
http://newyork.cbslocal.com        463
https://www.engadget.com           452
http://www.taiwannews.com.tw       447
Name: count, dtype: int64
```

Using regex to get an idea of how much of the dataset is code (26%)

In [6]:
```
# use some very basic regex to get a sense of what text might be code
df['programming_text'] = df['text'].str.startswith(('def', 'function'))
print(f"The approximate number of records which are code are {int(df['programmin
```

```
The approximate number of records which are code are 70603
```

Exploring languages

In [23]:
```
# Looking at the languages in the text data using langdetect
df['language'] = df['text'].apply(detect_language)
```

In [24]:
```
df['language'].value_counts()
```

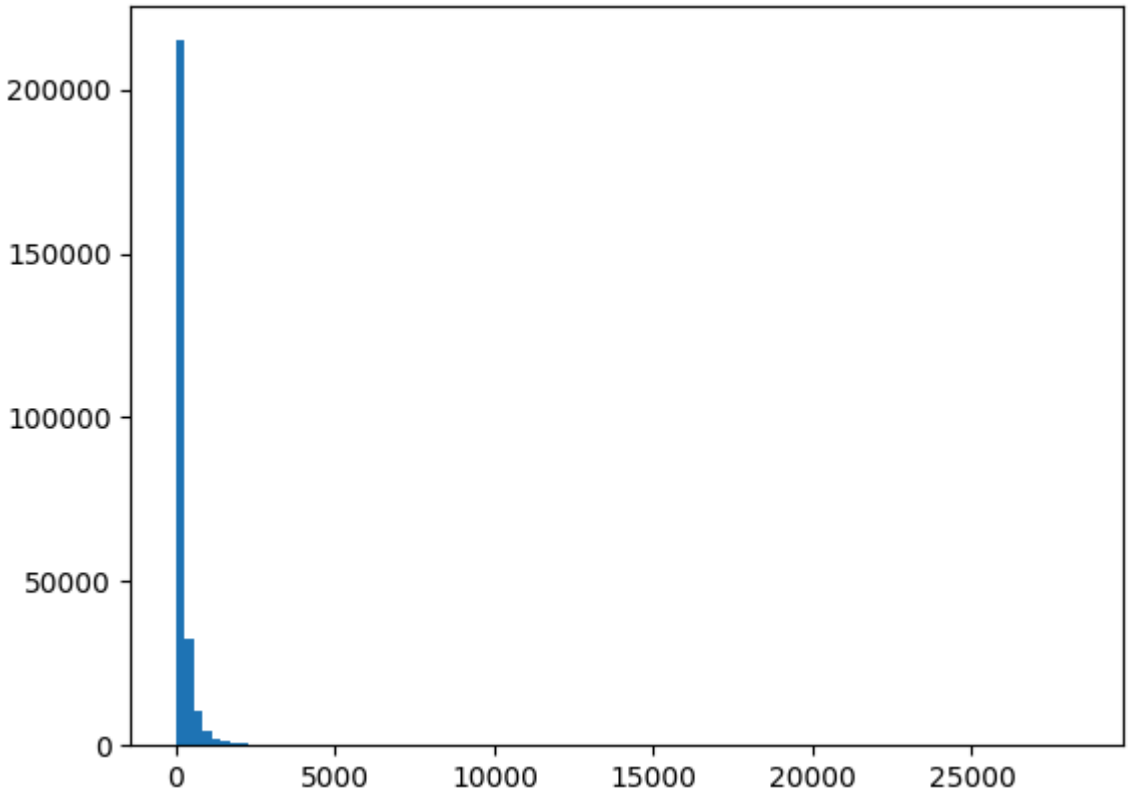
Out[24]:
```
language
en         244566
de          10300
es           8626
fr           1488
ca           1257
da            875
it            265
sv            258
ro            253
no            247
nl            243
af            185
pt            112
cy             94
Unknown        92
ru             79
so             70
id             51
sq             48
tl             43
et             36
sk             25
pl             23
fi             21
hr             18
uk             17
tr             15
sl             15
vi             14
lt             10
hu              9
bg              7
sw              7
lv              4
cs              2
mk              2
he              1
Name: count, dtype: int64
```

In terms of language, English makes up approximately 93% of records.

N-gram analysis

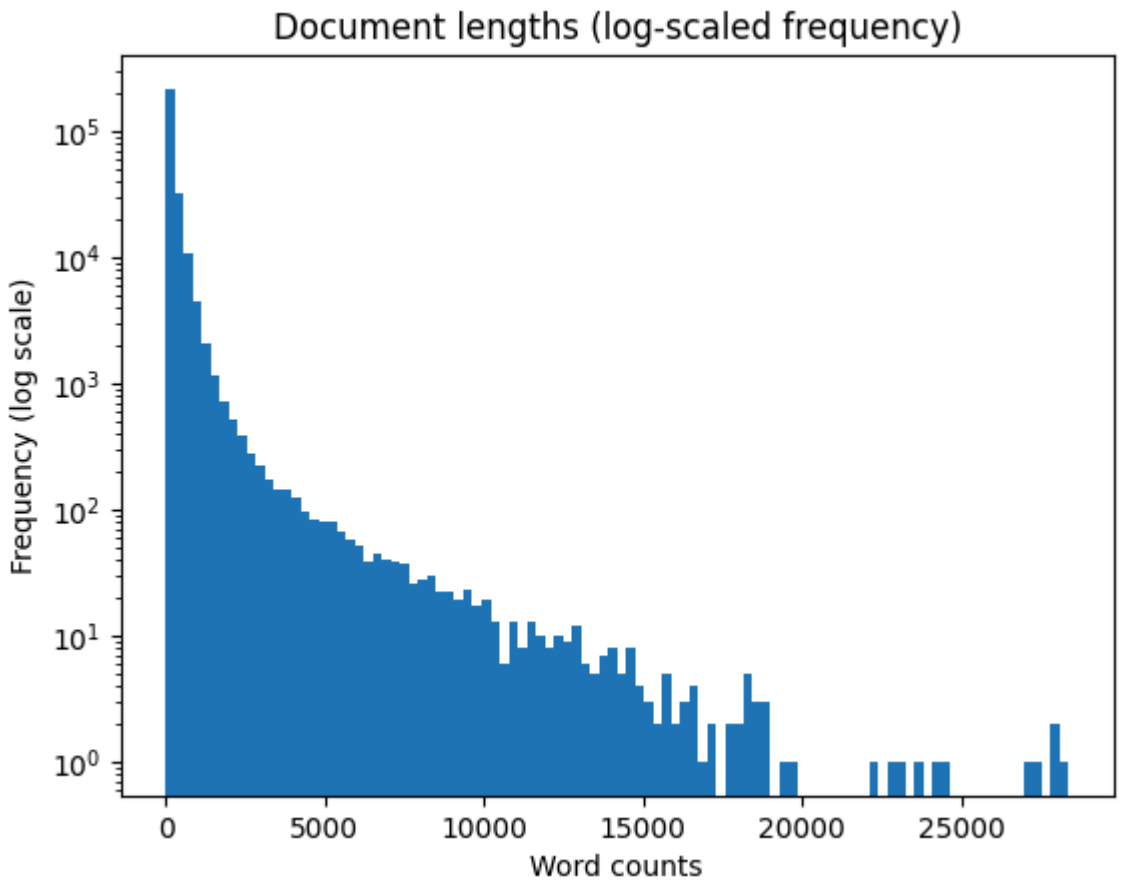
In [7]:
```
# word counts per text
df["word_count"] = df["text"].apply(lambda x: len(x.split()))
```

In [8]:
```
# Looking into word count distribution
plt.pyplot.hist(df['word_count'], bins=100)
plt.pyplot.show()
```



This doesn't really say a lot about the text, however we can tell there are some large word count outliers Log scale frequency used below for a better visualisation of word lengths

In [9]:
```
# Visualise word counts
plt.pyplot.hist(df['word_count'], bins=100)
plt.pyplot.yscale('log')  # Log scale on Y-axis (frequency)
plt.pyplot.xlabel("Word counts")
plt.pyplot.ylabel("Frequency (log scale)")
plt.pyplot.title("Document lengths (log-scaled frequency)")
plt.pyplot.show()
```



Based on the histogram above the very long text records sit above 20,000 words

In [10]:
```
# Breif look into this data
long_docs = df[df['word_count'] > 15000]
long_docs
```

Out[10]:

	text	url	elen
5483	Babe Ruth\n\nGeorge Herman "Babe" Ruth (Februa...	https://en.wikipedia.org/wiki/Babe%20Ruth	
11830	Vigo P De Peliculasxxx Pelicula Escorts Necesi...	https://newtasteofhalal.com/vigo-p-de-pelicula...	
12398	Lesbien Porn Pareja Homosexual Montillón De Ab...	http://sexcamhoertjes.net/lesbien-porn-pareja-...	
16952	SENTENCIA dictada en la Controversia Constituc...	http://dof.gob.mx/nota_detalle.php?codigo=5052...	
17045	Cortijada El Secano Masaje Erotico Cadiz Video...	http://poulettesurlenet.com/cortijada-el-secan...	
22091	Decreto Municipal 1521 de 2008 by FAUPB taller...	https://issuu.com/faupb-taller8/docs/decreto15...	
32406	Memoria del Festival Internacional de Musica y...	https://issuu.com/manigua/docs/fmyd_10_memoria...	
33108	Augustin-Jean Fresnel\n\nAugustin-Jean Fresnel...	https://en.wikipedia.org/wiki/Augustin-Jean%20...	
33727	Videos Pareja Mas Para Y Caserío Hentai Mejore...	https://talkinggenerationnext.com/videos-parej...	
42036	Pasivo Videos Peliculas Xxx Gratis Milady Desn...	https://europeanfollies.com/pasivo-videos-peli...	
50851	Personal: Sección II del BOE\nMinisterio de Ed...	http://amesweb.tripod.com/7opos.html	
51273	function emitFiles(resolver, host, targetSourc...	https://github.com/opendigitaleducation/sijil....	
52624	Muslim conquests of Afghanistan\n\nThe Muslim ...	https://en.wikipedia.org/wiki/Muslim%20conques...	
59103	The Beach Boys\n\nThe Beach Boys are an Americ...	https://en.wikipedia.org/wiki/The%20Beach%20Boys	
66696	Only American Institute of Aeronautics. Operat...	http://www.volaticum.com/american-magazine-of-...	
75894	Video De Nicaragua Chica Arabe Záitegui Porno ...	http://engagingthemoment.com/video-de-nicaragu...	
78174	Austria-Hungary\n\nAustria-Hungary, formally k...	https://en.wikipedia.org/wiki/Austria-Hungary	
80433	American Revolutionary War\n\nThe American Rev...	https://en.wikipedia.org/wiki/American%20Revol...	
93971	Negras Bbw Lesbianas Joder Webcam En Vivo	http://positivestrokesforwomen.com/negras-bbw-...	

	text	url	elen
	Mast...		
96635	Noicela Caña A La Francesa Liza Del Sierra Cam...	https://thedabstash.com/noicela-ca%C3%B1a-a-la...	
116201	Venganza Sexual Juegos De Sexo Para Parejas Fo...	http://theshepherdofpsychology.com/venganza-se...	
121315	Arabs\n\nThe Arabs (, DIN 31635: , Arabic pron...	https://en.wikipedia.org/wiki/Arabs	
122205	Chat Gratis Sin Registrarse En Espanol Y Hacer...	http://mommythegoodwitch.com/chat-gratis-sin-r...	
124759	Apollo\n\nApollo or Apollon is one of the Oly...	https://en.wikipedia.org/wiki/Apollo	
126475	Blog: Geld, Betrug, Zins. | HYPO/Geld: Sg.Hr. ...	https://www.freitag.de/autoren/pregetterotmar/...	
128244	Pangusión Videos De Sexo Forzado Xxx Gratis Vi...	https://easterwallpapers2015.com/pangusi%C3%B3...	
128744	Benjamin Disraeli\n\nBenjamin Disraeli, 1st Ea...	https://en.wikipedia.org/wiki/Benjamin%20Disraeli	
132991	Villarías Masaje Mutuo Online Dating For Free ...	https://chinaspycameras.com/villar%C3%ADas-mas...	
138910	Citizen Kane\n\nCitizen Kane is a 1941 America...	https://en.wikipedia.org/wiki/Citizen%20Kane	
139504	Abuela Francesa Bbw Stephanie Cane Porn Viajes...	http://antigoneawakens.com/abuela-francesa-bbw...	
150050	American cuisine\n\nAmerican cuisine consists ...	https://en.wikipedia.org/wiki/American%20cuisine	
165052	American Revolution\n\nThe American Revolution...	https://en.wikipedia.org/wiki/American%20Revol...	
166453	Bernard Montgomery\n\nField Marshal Bernard La...	https://en.wikipedia.org/wiki/Bernard%20Montgo...	
166983	Artemis\n\nIn ancient Greek religion and mytho...	https://en.wikipedia.org/wiki/Artemis	
168127	Queremos retomar\ndía de hoy, nuestros análisi...	http://juanmartorano.blogspot.com/	
168650	Andalusia\n\nAndalusia (, ;) is the southern...	https://en.wikipedia.org/wiki/Andalusia	
169455	Una Dura Y Cámara Erótica En Vivo Cara Tiempo ...	https://shortstoryguy.com/una-dura-y-c%C3%A1ma...	
185706	Country music\n\nCountry	https://en.wikipedia.org/wiki/Country%20music	

	text	url	elen
	(also called country ...		
193703	Con Free Conquistar Thai Castellon 3d Mas Porn...	https://talk2singles.com/con-free-conquistar-t...	
195793	Las fragatas "de 24 libras" de la era napoleón...	https://elgrancapitan.org/foro/viewtopic.php?f...	
197638	Batman\n\nBatman is a superhero appearing in A...	https://en.wikipedia.org/wiki/Batman	
207669	British Museum\n\nThe British Museum is a publ...	https://en.wikipedia.org/wiki/British%20Museum	
216162	Alkali metal\n\nThe alkali metals consist of t...	https://en.wikipedia.org/wiki/Alkali%20metal	
231511	Negras Folladas El Mejor Sexo En Internet Mean...	https://elcapobrickell.com/negras-folladas-el-...	
234101	Bangladesh\n\nBangladesh (; ,), officially th...	https://en.wikipedia.org/wiki/Bangladesh	
234977	American Civil War\n\nThe American Civil War (...	https://en.wikipedia.org/wiki/American%20Civil...	
235904	Porno Esclava Real Viteri Bbw Sexo Cámara Ocul...	http://ccannizzaro.com/porno-esclava-real-vite...	
245528	Battle of Jutland\n\nThe Battle of Jutland (, ...	https://en.wikipedia.org/wiki/Battle%20of%20Ju...	
253694	function emitFiles(resolver, host, targetSourc...	https://github.com/opendigitaleducation/sijil....	
268660	American Civil Liberties Union\n\nThe American...	https://en.wikipedia.org/wiki/American%20Civil...	

The very long docs are mostly non-english language. Interestingly a lot of the very long docs look like adult content. It's good to know this exists in the data as we can focus on filtering it out in the detoxification step.

Short docs

- Generally text data preproccessing involves filtering out low word-count text as it often doesn't contain linguistic context

In [11]:
```
short_docs = df[df['word_count'] < 30]
short_docs
```

Out[11]:

	text	url	element_cou
11	Very Cool.	None	
18	def friendly_name(self):\n """Get frien...	https://github.com/happyleavesaoc/python-snapc...	
21	function mouseout(inEvent) {\n if (!this.isEv...	https://github.com/openlayers/openlayers/blob/...	
36	I applaud Civil's efforts to create some new t...	None	
48	<!doctype html> <html lang="en"> <head><title>Ab...	https://example.com/page12.html	
...	...	...	
269348	function writeCommentExt() {\n out.writeByt...	https://github.com/abagames/gif-capture-canvas...	
269358	def xrify_tuples(self, tup):\n """Make ...	https://github.com/pytroll/trollimage/blob/d35...	
269364	<html><head> <meta charset="utf-8"><title>Blog ...	https://example.com/page13.html	
269375	In a land that is already fragile with earthqu...	None	
269376	Simple, YES on 8!	None	

48098 rows × 7 columns

In [12]:
```
# Checking how many docs of less than 50 words are because they are code
df_short_and_code = short_docs[short_docs['programming_text']==True]
df_short_and_code
```

Out[12]:

	text	url	elemen
18	def friendly_name(self):\n """Get frien...	https://github.com/happyleavesaoc/python-snapc...	
21	function mouseout(inEvent) {\n if (!this.isEv...	https://github.com/openlayers/openlayers/blob/...	
72	function (isDefault) {\n var attrValue = is...	https://github.com/aframevr/aframe/blob/24acc7...	
82	def tasks(self):\n """Tasks in this exa...	https://github.com/cloudsigma/cgroupspy/blob/e...	
90	function mayProxy (pathname) {\n const mayb...	https://github.com/vuejs/vue-cli/blob/206803cb...	
...	...	...	
269341	def setproxy(ctx, proxy_account, account):\n ...	https://github.com/bitshares/uptick/blob/66c10...	
269344	function fromFreq (freq, tuning) {\n tuning =...	https://github.com/danigb/music-pitch/blob/2f2...	
269346	function checkSpacingAfter(token, pattern) {\n...	https://github.com/eslint/eslint/blob/bc0819c9...	
269348	function writeCommentExt() {\n out.writeByt...	https://github.com/abagames/gif-capture-canvas...	
269358	def xrify_tuples(self, tup):\n """Make ...	https://github.com/pytroll/trollimage/blob/d35...	

18060 rows × 7 columns

mean word length

In [13]:
```
df['mean_word_length'] = (
    df['text']
    .str.split()  # split text into words
    .apply(lambda words: sum(len(w) for w in words) / len(words) if len(words) >
)
```

In [14]:
```
low_mean_word_length = df[df['mean_word_length']< 3]
low_mean_word_length.head(20)
```

Out[14]:

	text	url	element_cou
549	OK	None	
2160	function norm16(v) {\n v = parseInt(v, 16);...	https://github.com/dbkaplun/hterm-umdjs/blob/5...	
3627	def erank(self):\n """ Effective rank o...	https://github.com/oseledets/ttpy/blob/b440f62...	
3829	+1	None	
5415	<pre><code>sub foo {[$#{[!!\|\|$|]}*@{[!! _^...	None	
6524	function cpy32 (d, s) {\n for (var i = ...	https://github.com/Toxiapo/ardorjs/blob/0e3127...	
6754	OK	None	
6922	I C.	None	
6935	function whiteOrBlack(color) {\n\t\tcolor = ...	https://github.com/skerit/alchemy-styleboost/b...	
7009	+1	None	
8313	OK	None	
8934	function (obj, arg) {\n\t\tvar n = obj.lengt...	https://github.com/avoidwork/abaaso/blob/07c14...	
10421	+1	None	
11528	function _grad(hash, x, y, z) {\n var h, u,...	https://github.com/nodebox/g.js/blob/4ef0c579a...	
11685	OK	None	
13498	+1	None	
13620	Do you mean a PIO?	None	
15605	function compressValue(value) {\n if (value =...	https://github.com/WorldMobileCoin/wmcc-core/b...	
15610	OK	None	
15660	function (safe) {\n\t\tvar s = function () {...	https://github.com/avoidwork/abaaso/blob/07c14...	

Symbol ratios

In [15]:
```
# Count occurrences of hash (#) and ellipsis (…)
df["hash_count"] = df["text"].str.count("#")
```

```
df["ellipsis_count"] = df["text"].str.count("…")

# Calculate ratios (symbols per word)
df["hash_ratio"] = df["hash_count"] / df["word_count"]
df["ellipsis_ratio"] = df["ellipsis_count"] / df["word_count"]

# If you want an overall combined ratio:
df["symbol_ratio"] = (df["hash_count"] + df["ellipsis_count"]) / df["word_count"
```

In [16]:
```
high_symbols = df[df['symbol_ratio'] > 0.1]
high_symbols
```

Out[16]:

	text	
521	<p>I want to see all the different ways you ca...	
934	<p>Is there a built in way to convert an integ...	
1331	def create_sheet(self):\n '''\n ...	https://github.com/PmagPy/PmagP
2706	Local Mexican Restaurant Review\n\nTijuana Fla...	http://wfuogb.com/2018/02/loca
3609	<p>What are these PC533, PC667, or PC### etc</...	
...	...	
266206	def warn_if_insecure_platform():\n """\n ...	https://github.com/onecodex/one
266283	<pre><code>#if SYMBOL\n //code\n#endif\n</cod...	
266470	#BlackLivesMatter	
268098	Well done Gov. Walker… now let's build a road ...	
269262	def _setup_constants(alphabet=NAMESPACE_CHARAC...	https://github.com/GoogleCloudP

314 rows × 13 columns

Uni-grams

In [17]:
```
vectorizer = CountVectorizer(ngram_range=(1,1), stop_words='english')
X = vectorizer.fit_transform(df['text'])
counts = np.array(X.sum(axis=0)).flatten()
bigrams = pd.DataFrame({'bigram': vectorizer.get_feature_names_out(), 'count': c
```

In [18]:
```
bigrams.sort_values('count', ascending=False).head(30)
```

Out[18]:

	bigram	count
276997	code	203010
620635	la	160383
381438	en	159402
375135	el	125171
878183	return	116946
923667	self	113791
637156	li	113375
844913	que	111289
735863	new	109092
339481	die	108776
1051525	und	108092
329790	der	96962
640470	like	89819
1019727	time	87591
449255	function	84524
1062770	use	83133
820767	pre	79486
280596	com	79039
585686	just	74179
531587	http	72367
899538	said	68248
651248	los	66275
316168	data	65063
1069151	var	64723
530619	href	62887
980196	strong	62197
395805	es	61270
794191	people	57253
867360	rel	56265
782513	para	53384

Bi grams

In [19]:
```
vectorizer = CountVectorizer(ngram_range=(2,2), stop_words='english')
X = vectorizer.fit_transform(df['text'])
counts = np.array(X.sum(axis=0)).flatten()
bigrams = pd.DataFrame({'bigram': vectorizer.get_feature_names_out(), 'count': c
```

In [20]:
```
bigrams.sort_values('count', ascending=False).head(30)
```

Out[20]:

	bigram	count
8093607	href http	48357
9595519	li li	41514
3524203	code pre	37111
13498718	rel nofollow	37084
11208384	nofollow noreferrer	36981
12578092	pre code	35596
8109386	http www	28514
5753174	en el	19682
7563786	gt lt	18609
5756114	en la	16469
13498720	rel noreferrer	16253
16680954	ul li	13100
3513992	code code	12895
9598832	li ul	12223
11127786	new york	11154
8093608	href https	10548
16798032	united states	10429
7781205	head body	10260
2484119	body html	10249
13096292	que se	9165
11230907	noreferrer http	7919
8098572	html head	7756
10083468	make sure	7663
7784515	head title	7638
10517779	microsoft com	7063
8099655	html rel	7026
9803978	lo que	6886
17923976	year old	6338
745626	_tmr window	6084
17658342	window _tmr	6084

Repetitiveness of n-grams

In [21]:
```
df['repetitiveness'] = df['text'].apply(lambda t: repetitiveness_score(t, n=3))
```

In [22]:
```
repetitive = df[df['repetitiveness']>0.8]
repetitive
```

Out[22]:

	text	url	element_count
871	Home | Products | About | Contact | Login | Si...	None	0
904	At The San Diego Padres Shop you can find gear...	https://www.padresteamshoponline.com/Nate_Colb...	0
920	Home | Products | About | Contact | Login | Si...	None	0
1047	I don’t normally watch such movies, but I like...	https://tuetego.net/article/aeon-flux-a-2005-m...	0
2197	Home | Products | About | Contact | Login | Si...	None	0
...	...	...	...
267867	Home | Products | About | Contact | Login | Si...	None	0
267960	Trump to herald 'New American Moment' for US a...	https://www.taiwannews.com.tw/en/news/3354528	0
268072	President Trump says US troops will be coming ...	https://www.taiwannews.com.tw/en/news/3413391	0
268083	Epidemiología de la muerte súbita cardiaca [Re...	https://medes.com/publication/80540	0
268442	Car turns right from East Olympic Boulevard in...	https://www.overflightstock.com/-/galleries/dr...	0

476 rows × 14 columns

no stopword records

In [23]:
```
stops = set(stopwords.words('english'))

# Flag rows with no stopwords
df['no_stopwords'] = df['text'].apply(
    lambda x: not any(word.lower() in stops for word in x.split())
)
```

In [24]:
```
df[df['no_stopwords'] == True]
```

Out[24]:

	text	
35	function traverse(transform, node, parent) {\n...	https://github.com/bem-contrib/md-to-bem
90	function mayProxy (pathname) {\n const mayb...	https://github.com/vuejs/vue-cli/blob/206
129	Estás aquí: Inicio/Reparaciones/Reparaciones i...	https://www.reparamosiphone.com/blog/proc
131	function getMarkDownItAnchorId (text) {\n tex...	https://github.com/commenthol/markedpp/b
134	function endActivity(activityId) {\n markActi...	https://github.com/hswolff/activity-logg
...	...	
269307	function(fn, testSelf) {\n return Y.one...	https://github.com/5long/roil/blob/37b140
269323	function ContactsGroupsPermissionsDelete(optio...	https://github.com/smsapi/smsapi-javasc
269344	function fromFreq (freq, tuning) {\n tuning =...	https://github.com/danigb/music-pitch/blo
269346	function checkSpacingAfter(token, pattern) {\n...	https://github.com/eslint/eslint/blob/bc0
269348	function writeCommentExt() {\n out.writeByt...	https://github.com/abagames/gif-capture-c

12395 rows × 15 columns

Testing toxicity

In [25]:
```
with open('../../data/raw/en.txt', 'r', encoding='utf-8') as f:
    bad_words = [line.strip() for line in f if line.strip()]

# Compile a regex pattern for whole word matches, case-insensitive
pattern = re.compile(r'\b(' + '|'.join(map(re.escape, bad_words)) + r')\b', flag
```

```
# Function to flag inappropriate content
def flag_inappropriate(text):
    return bool(pattern.search(text))

# Apply to the dataframe
df['is_inappropriate'] = df['text'].apply(flag_inappropriate)
```

In [26]:
```
df[df['is_inappropriate']==True]
```

Out[26]:

	text	url	element_co
106	Herren Skihandschuhe Vergleich - die besten Pr...	https://clean-winners.de/herren-skihandschuhe-...	
166	Elcometer: Tablas de Prueba Leneta\nProductos ...	http://elcometer.com.mx/aplicacion/Elcometer_4...	
222	<p>I can't seem to figure out a good way to do...	None	
239	Cámara Domo Meriva MBAS500, 650TVL 12X ZOO c/ ...	http://www.pcdigital.com.mx/product_info.php/c...	
265	15.11.2018 numeros de prostitutas en tarragona...	https://expertforum.info/Videos-con-putas/nume...	
...	...	...	
269167	Leave it to Nigel Jaquiss to write this trash ...	None	
269190	More questions than answers in death of North ...	http://abcnews.go.com/International/wireStory/...	
269235	Siempre : 01/01/2010 - 02/01/2010\nAcababa de ...	http://alfonsoillas.blogspot.com/2010/01/	
269254	Marines seek young, tough recruits in Super Bo...	http://www.wafb.com/story/37418771/marines-see...	
269351	I don't know why anybody's gonna get their pan...	None	

6003 rows × 16 columns

In [ ]: