توضیحات پیش پردازش داده‌های متنی با استفاده از Regular Expression

عرفان محمّدزاده

er.mohammadzadeh@gmail.com

e-mohammadzadeh

۵ آذر ۱۴۰۴

فهرست مطالب

فهرست جداول

فهرست تصاویر

۱ مقدمه

برای اینکه بتوانیم مجموعه‌داده‌ها را با مدل تجزیه و تحلیل کنیم، قبل از آن نیاز داریم که این مجموعه‌داده‌ها را آماده‌سازی کنیم. آماده‌سازی یا پیش‌پردازش مجموعه‌داده به پاک‌سازی داده از موارد ناخواسته گفته می‌شود، مثل: پاک کردن کلمات یا علائم اضافی یا فضای خالی از مجموعه‌داده. در ابتدا نحوه خواندن مجموعه‌داده موردنظر را توضیح می‌دهیم و سپس مواردی که برای پیش‌پردازش مجموعه‌داده‌هایمان انجام شده را با جزئیات شرح می‌دهیم. قسمت پیش‌پردازش به سه بخش اصلی تقسیم می‌شود: جابجایی[۱]، حذف[۲] و بسط دادن[۳]. ترتیب مراحل پیش‌پردازش انجام شده، برابر با ترتیبی است که در جدول ۱ نوشته شده است.

#	**Preprocessing Steps**
1	Remove Null Values
2	Replace Slang words
3	Lowercase
4	Expand Contractions
5	Remove Stopwords
6	Replace URLs
7	Replace Email Addresses
8	Expand Mentions
9	Expand Hashtags

#	**Preprocessing Steps**
10	Remove Date
11	Replace Currencies
12	Replace Time
13	Remove Tags
14	Remove Cashtags
15	Remove Numbers
16	Remove Titles
17	Remove Punctuations
18	Replace Repeated Letters

جدول ۱: ترتیب مراحل پیش‌پردازش

[1] Replace
[2] Remove
[3] Expand

۲ خواندن مجموعه‌داده

در ابتدا آدرس[۱] و نام فایل‌های موردنظر به برنامه داده می‌شود، سپس فایل‌ها به ترتیب خوانده شده و در یک دیکشنری[۲] ذخیره می‌شوند. قسمت کلید[۳] در دیکشنری یکی از موارد train، test و dev می‌تواند باشد. قسمت مقدار[۴] در دیکشنری، فایل خوانده شده متناظر با کلید است. سپس مراحل پیش‌پردازش بر روی هر کدام از مقادیر این دیکشنری، اعمال می‌شود و حاصل یک داده تمیز است که در یک دیکشنری دیگر ذخیره می‌گردد. در ادامه هر کدام از مراحل پیش‌پردازش با جزئیات شرح داده شدند.

در ابتدا، ستون‌های مجموعه‌داده را با توجه به محتوای آن‌ها دو بخش می‌کنیم. یک قسمت ستون‌هایی که داده‌های عددی دارند و یک قسمت ستون‌هایی که داده‌های متنی دارند. سپس مراحل پیش‌پردازش بر روی ستون‌های حاوی داده‌های متنی اعمال می‌شوند.

۳ حذف

۱.۳ Remove Null Values

بخش اول، مربوط به حذف برخی کاراکترها، کلمات و یا علائم خاص است.

اگر مجموعه‌داده‌ای که در این بخش بررسی می‌شود، حاوی سلولی[۵] با مقدار تهی[۶] باشد، آن سطر کلاً پاک می‌شود هر چند که سایر سلول‌های آن سطر حاوی مقادیری باشند. همچنین برای اطلاع کاربر، شماره سطرهایی که سلول تهی دارند و قرار هست پاک شوند، در خروجی چاپ می‌شود.

۲.۳ Remove Stopwords

یک فایل متنی که توسط نویسنده ساخته شده، در کنار کد قرار دارد که دارای فرمت txt است. این فایل محتوی اکثر ایست‌واژه[۷]های زبان انگلیسی است با این تفاوت که کلمات منفی را به دلیل اینکه نمی‌خواهیم

[1] Path
[2] Dictionary
[3] Key
[4] Value
[5] Cell
[6] Not a Number (NaN)
[7] Stopword

English Stopwords.txt

```
1:  until, their, further, can, each, yourself, it, myself, out, were,
2:  will, but, where, ve, should've, above, your, again, up, me, those,
3:  an, very, these, needn, having, he, under, how, m, between, its,
4:  about, had, this, that'll, it's, they, hers, when, any, she, have,
5:  of, for, during, we, while, below, the, she's, through, herself,
6:  before, if, you've, other, now, that, own, off, ourselves, you're,
7:  with, whom, and, has, into, in, on, so, d, most, them, itself, same,
8:  down, you'll, is, should, because, from, yours, then, themselves,
9:  such, i, over, there, being, or, at, been, her, ours, did, here,
10: a, his, are, you'd, y, just, why, than, yourselves, our, be, which,
11: am, theirs, doing, was, s, ll, after, more, what, re, my, both, do,
12: does, all, o, to, himself, as, you, who, only, by, too, t, once,
13: against, few, ma, him, some
```

شکل ۱ : لیست ایست‌واژه های غیرمنفی در زبان انگلیسی

کلمات منفی از مجموعه‌داده حذف شوند در نظر نگرفتیم. این فایل متنی شامل ۱۴۰ ایست‌واژه انگلیسی است. شکل ۱ لیست ایست‌واژه‌های انگلیسی را نشان می‌دهد با این تفاوت که کلمات منفی از این لیست حذف شدند.

۳.۳ Remove Date

هدف از این قسمت، استخراج و حذف تاریخ است. نحوه بیان تاریخ می‌تواند متفاوت باشد، سعی شده است که اکثر فرمت‌های مهم برای بیان تاریخ را استخراج کنیم. جدول ۲ فرمت‌های کوتاه و جدول ۳ فرمت‌های کامل را که برای زمان می‌توانیم از متن استخراج بکنیم نشان می‌دهد. انواع حالت‌هایی که برای اسامی ماه‌های سال میلادی می‌توان در نظر گرفت نیز در شکل ۴ ذکر شده است. برای استخراج تاریخ از متن هم می‌توان به تنهایی از فرمت کوتاه و کامل استفاده کرد و هم می‌توان این دو فرمت را با هم ترکیب کرد. پس در نتیجه تعداد حالت‌هایی که می‌توان استخراج کرد زیاد هست و این باعث دقت بالای برنامه در هنگام پیش‌پردازش می‌شود.

جدول ۲: انواع فرمت‌های کوتاه برای زمان

#	**Short Date**		
1	01/01/2024	01-01-2024	01.01.2024
2	27/5/24	27-5-24	27.5.24
3	01/2024	01-2024	01.2024

جدول ۳: انواع فرمت‌های کامل برای زمان

Full Date			
#	First Part	Second Part	Third Part
1	Month-name*	of	Any number {4 digit}
2	Month-name	of	Any number {4 digit},
3	Any number {1-4 digit}	Month-name	
4	Any number {1-4 digit},	Month-name,	
5	Month-name	Any number {2 digit}{st, nd, rd, n-th}	Any number {1-4 digit}
6	Month-name,	Any number {2 digit}{st, nd, rd, n-th},	Any number {1-4 digit}
7	Month-name	Any number {1-4 digit}	Any number {2 digit}{st, nd, rd, n-th}
8	Month-name,	Any number {1-4 digit}	Any number {2 digit}{st, nd, rd, n-th},
9	Any number {1-4 digit}	Month-name	Any number {2 digit}{st, nd, rd, n-th}
10	Any number {1-4 digit},	Month-name,	Any number {2 digit}{st, nd, rd, n-th}

* Each month name (table 4) can be substituted

جدول ۴: انواع فرمت‌ها برای اسامی ماه‌ها

Month Names			
Jan	January	jan	january
Feb	February	feb	february
Mar	March	mar	march
Apr	April	apr	april
-	May	-	may
Jun	June	jun	june
Jul	July	jul	july
Aug	August	aug	august
Sept	September	sept	september
Oct	October	oct	october
Nov	November	nov	november
Dec	December	dec	december

Remove Tag ۴.۳

در این بخش، تگ[1] های شروع و پایان HTML و سایر تگ‌ها را شناسایی و حذف می‌کنیم. به عنوان مثال، تگ‌های روبرو در این بخش حذف می‌شوند:

```
1: <head>
2: <url>
3: </body>
4: </div>
5: <img src="image.jpg">
```

Remove Cashtag ۵.۳

طبق تعریف شرکت ایکس[2] (توییتر سابق)، «برچسب مالی یک نماد منحصربه‌فرد و مخفف برای یک رمزارز[3] یا یک شرکتی است که قبل از آن علامت دلار آمریکا قرار دارد» [۱]. از برچسب‌های مالی در برنامه ایکس استفاده می‌شود. حال در این قسمت، قصد داریم که این برچسب‌های مالی را در صورت وجود از متن استخراج و حذف کنیم. به عنوان مثال، برچسب‌های مالی روبرو و مشابه آن‌ها در این بخش حذف می‌شوند:

1 $GOOG \longmapsto$ برچسب مالی شرکت گوگل

2 $TSLA \longmapsto$ برچسب مالی شرکت تسلا

3 $PEP \longmapsto$ برچسب مالی شرکت پپسی

4 $BTC \longmapsto$ برچسب مالی رمز ارز بیت‌کوین

5 $ETH \longmapsto$ برچسب مالی رمز ارز اتریوم

[1] Tag
[2] X
[3] Cryptocurrency

Remove Number ۶.۳

پاک کردن اعداد موجود در متن، یکی دیگه از مراحل پیش‌پردازش است. اعدادی که حذف می‌شوند، چنین فرمتی می‌توانند داشته باشند:

- عدد مثبت یا منفی
- عدد کسری
- عدد اعشاری
- عددی که سه رقم سه رقم با کاراکتر ٫ یا ، جدا شده است
- عدد علمی به فرمت $\mathrm{A}E \pm \mathrm{B}$ یا $\mathrm{A}e \pm \mathrm{B}$ (A و B هر دو عدد هستند).

هنگام پاک‌سازی متن از اعداد با دو چالش مواجه شدیم که در ادامه به صورت مختصر به هر کدام اشاره‌ای شده است:

- پاک کردن همه اعداد موجود در داخل متن به طوریکه اعداد داخل ساختار

$$\triangleright\{L/M/T/E\}_{counter}\triangleleft$$

دست نخورده باقی بمانند. برای رفع این مشکل، ابتدا این ساختار را شناسایی کرده و سپس نوشته‌های داخل دو کاراکتر $\triangleleft$ و $\triangleright$ را برش داده و بعد از آن عملیات پاک کردن اعداد را انجام می‌دهیم. و در نهایت نوشته‌های برش داده شده در جای خود (بین دو کاراکتر) الصاق می‌شوند.

- دومین چالش، پاک کردن اعداد به طوریکه اعداد داخل کاراکترهای یونیکد[1] حذف نشوند. وقتی در مجموعه‌داده جمله یا کلمه‌ای غیر انگلیسی وجود داشته باشد، احتمال وجود کاراکتر یونیکد در آن کلمه یا جمله وجود دارد. کاراکترهای یونیکد معمولا دارای چنین ساختاری هستند:

\u0000

\u10FFFF

[1] Unicode Characters

بعد از حرف «u» شش کاراکتر می‌توان نوشت که بازه هر کاراکتر «اعداد بین صفر تا نه - حروف کوچک انگلیسی بین a تا f - و همچنین حروف بزرگ انگلیسی بین A تا F» است. در ادامه چندین مثال از کاراکترهای یونیکد نوشته شده است:

K\u00f6lsch ⟼ Kölsch

pur\u00e9ed ⟼ puréed

jalape\u00f1o ⟼ jalapeño

Gar\u00e7on ⟼ Garçon

با استفاده از دستوراتی ابتدا کاراکترهای یونیکد را شناسایی کرده و سپس شش کاراکتر را نادیده می‌گیریم و باقی اعداد را حذف می‌کنیم و بدین صورت دومین چالش هنگام حذف اعداد نیز رفع می‌شود.

۷.۳ Remove Title

حذف عناوین استفاده شده در متن نیز می‌تواند جزو یکی از کارهای پیش‌پردازش باشد، عناوینی مثل

Mr. Mrs. Miss. Dr. Prof. Pres. (president) ...

که به وفور در متن‌های انگلیسی تکرار می‌شوند. هفده کلمه انگلیسی برای این بخش در نظر گرفتیم که لیست کامل کلمات در عکس ۲ قابل مشاهده هستند. برای پاک کردن عناوین این نکات در نظر گرفته شدند: اولاً عدم حساسیت به حروف بزرگ و کوچک؛ ثانیاً وجود یا عدم وجود نقطه بعد از عنوان.

English Titles

```
Mr, Ms, Mrs, Miss, Dr, Prof, Sir, Ma'am, Madam, Madame,
Rev,  # Reverend, used for Christian clergy
Fr,  # Father, used for Catholic priests
Sr,  # Sister, used for Catholic nuns
Capt,  # Captain, used in the military and for pilots
Gen,  # General, used in the military
Hon,  # Honorable, used for judges and certain politicians
Pres,  # President, used for presidents of companies or organizations
```

شکل ۲: لیست عناوین زبان انگلیسی

Remove Punctuation ۸.۳

حذف علائم نگارشی و بصری اضافی از متن نیز از مراحل پیش‌پردازش به حساب می‌آید. در این قسمت، کاراکترهایی از قبیل

‘ - = [ ] \ ; ’ , . / ! @ # $ % ^& * ( ) _ + | : ” < > ?

را از مجموعه‌داده پیدا و حذف می‌کنیم.

در این قسمت نیز چالش مربوط به کاراکترهای یونیکد وجود داشت؛ کاراکتر یونیکد بدون علامت «\» معنی ندارد و نمی‌تواند به شکل موردنظر تبدیل شود. بنابراین هنگام حذف علائم نگارشی باید دقت کنیم که علامت «\» را در کاراکترهای یونیکد حذف نکنیم.

۴ جابجایی

بخش دوم، مربوط به جابجایی برخی کاراکترها، کلمات یا علائم خاص با کلمات دیگر یا کلیدهای منحصربه‌فردی است.

Slang Words ۱.۴

اولین کار در قسمت جابجایی، مربوط به جایگزینی کلمات عامیانه زبان انگلیسی با فرم کامل و رسمی است. در نوشتن متن به زبان انگلیسی، برای اینکه بتوان در سریع‌ترین زمان ممکن عملیات تایپ و ارسال پیام را انجام داد، معمولا از کلمات مخفف و عامیانه زیاد استفاده می‌شود. برای اینکه بتوان دقت مدل را افزایش داد، ما سعی کردیم تا این کلمات مخفف و عامیانه را تا حدودی شناسایی و با کلمات کامل هر کدام جایگزین کنیم.

برای این منظور یک فایل [۲] که قبلا ساخته شده بود را تغییراتی دادیم؛ در نهایت فایل حاوی یک دیکشنری است که key همان کلمه مخفف و value همان کلمه به صورت کامل است. شکل ۳ نمونه‌ای از این فایل را نشان می‌دهد. برای دسترسی به فایل کامل، به گیت‌هاب[1] نویسنده مراجعه شود.

[1] GitHub

English Slang Words.json

```
{"07734": "hello",
"2day": "today",
"2ge4": "Together",
"2morrow": "tommorrow",
"4ever": "forever",
"0noe": "Oh No",
"0vr": "over",
"10q": "thank you",
"5n": "fine",
"absnt": "absent",
"bc": "because",
"c@": "cat",
"dw": "don't worry",
...}
```

شکل ۳: بخشی از کلمات عامیانه زبان انگلیسی به همراه شکل کامل هر کدام

۲.۴ Lowercase

یکی از کارهایی که بر روی مجموعه‌داده‌ها انجام می‌گیرد، تبدیل حروف بزرگ انگلیسی به حروف کوچک انگلیسی است.

۳.۴ Contraction

یک فایل که توسط نویسنده ساخته شده (موجود در گیت‌هاب)، دارای فرمت json است. این فایل محتوی اکثر کلمات اختصار در زبان انگلیسی است. تصویر ۴ نمونه‌ای از این فایل را نشان می‌دهد. قالب کلی فایل به شکل دیکشنری هست که key همان کلمه مخفف و value شکل کامل همان کلمه مخفف است. در این قسمت از پیش‌پردازش، کلمات مختصر موجود در مجموعه‌داده را با شکل کامل جایگزین می‌کنیم.

English Contractions.json

```
{"ain't": "are not",
"aren't": "are not",
"can't've": "cannot have",
"can't": "cannot",
"'cause": "because",
"could've": "could have",
"couldn't've": "could not have",
"couldn't": "could not",
"didn't": "did not",
"doesn't": "does not",
"don't": "do not",
"dunno": "do not know",
...}
```

شکل ۴: بخشی از کلمات اختصار زبان انگلیسی به همراه شکل کامل هر کدام

URL ۴.۴

کار بعدی، استخراج مکان‌یاب منبع یکسان[1] و جابجایی آن‌ها با یک کلید منحصربه‌فردی هست که بعداً در صورت نیاز می‌توان با استفاده از کلید، نشانی وب موردنظر را بازیابی کرد. کلیدی که برای این بخش در نظر گرفته شده است دارای چنین ساختاری است:

$$\triangleright L_{counter} \triangleleft$$

به عنوان مثال، اولین نشانی وب با $\triangleright L1 \triangleleft$ جایگزین می‌شود، دومین نشانی وب با $\triangleright L2 \triangleleft$ جایگزین می‌شود و به همین ترتیب الی آخر. حرف L برگرفته از کلمه Link است. بعد از عملیات جایگزینی، نشانی وب به همراه کلید مربوطه در یک فایل ذخیره می‌شود. با این کار می‌توان نشانی وب را در صورت نیاز بازیابی کرد.

[1] Uniform Resource Locator (URL)

Email Address ۵.۴

سپس نوبت، استخراج آدرس پست الکترونیکی[1] و جابجایی آن با یک کلید منحصربه‌فردی هست که بعداً در صورت نیاز می‌توان با استفاده از کلید، آدرس ایمیل موردنظر را بازیابی کرد. برای دامنه آدرس ایمیل هیچ‌گونه محدودیتی در نظر گرفته نشده است پس هر آدرس ایمیل با نام کاربری و دامنه معتبر را شناسایی می‌کنیم. کلیدی که برای این بخش در نظر گرفته شده است دارای چنین ساختاری است:

$$\triangleright E_{counter} \triangleleft$$

به عنوان مثال، اولین آدرس ایمیل با $\triangleright E1 \triangleleft$ جایگزین می‌شود، دومین آدرس ایمیل با $\triangleright E2 \triangleleft$ جایگزین می‌شود و به همین ترتیب الی آخر. حرف E برگرفته از کلمه Email است. بعد از عملیات جایگزینی، آدرس ایمیل به همراه کلید مربوطه در یک فایل ذخیره می‌شود. با این کار می‌توان آن را در صورت نیاز بازیابی کرد.

Currency ۶.۴

یکی دیگر از مراحل پیش‌پردازش، استخراج پول و واحد پول[2] و جابجایی آن‌ها با یک کلید منحصربه‌فردی هست که بعداً در صورت نیاز می‌توان با استفاده از کلید، پول موردنظر را بازیابی کرد. کلیدی که برای این بخش در نظر گرفته شده است دارای چنین ساختاری است:

$$\triangleright M_{counter} \triangleleft$$

به عنوان مثال، اولین واحد پول با $\triangleright M1 \triangleleft$ جایگزین می‌شود، دومین واحد پول با $\triangleright M2 \triangleleft$ جایگزین می‌شود و به همین ترتیب الی آخر. حرف M برگرفته از کلمه Money است. بعد از عملیات جایگزینی، واحد پول به همراه کلید مربوطه در یک فایل ذخیره می‌شود. با این کار می‌توان آن را در صورت نیاز بازیابی کرد. نمونه‌هایی از اعداد و کلمه‌هایی که به عنوان پول شناسایی و با کلید مشخصی جایگزین می‌شوند، در زیر ذکر شده‌اند.

- علامت اختصاری‌هایی که هنگام استخراج در نظر گرفته می‌شوند، در جدول ۵ ذکر شده‌اند. درضمن چه علامت قبل از عدد ظاهر شود و چه بعد از عدد، هر دو حالت از متن استخراج می‌شود.

[1] Email
[2] Currency

#	Currency Symbol	Currency	Country
1	$	Dollar	USA
2	€	Euro	Euro Member Countries
3	£	Pound Sterling	UK
4	¥	Yen	Japan
5	₹	Rupee	India
6	¢	Cent	USA

جدول ۵: علائم اختصاری واحد پول کشورها

- معمولا هنگام نوشتن پول به صورت عددی، اعداد سه رقم سه رقم جدا می‌شوند. اینجا کاراکتر جدا کننده می‌تواند نقطه (.) یا کاما (,) یا (’) single qutation باشد.
- معمولا هنگام بیان کردن پول، عبارت‌هایی مثل «هزار»، «میلیون» و «میلیارد» به کار برده می‌شود. در جدول ۶، سیستم واحد شمارش را تعیین کردیم. به عنوان مثال، اگر در متن همراه با عدد حرف m یا کلمه Mill ظاهر شود (تفاوتی ندارد که سمت راست یا چپ عدد ظاهر شود) آنگاه آن حرف یا کلمه را میلیون در نظر می‌گیریم و عدد را به همراه آن حرف یا کلمه با کلید خاصی که در نظر گرفته شده جایگزین می‌کنیم.

با در کنار هم قرار دادن سه مورد قبلی، انواع حالت‌ها برای استخراج پول از متن بدست می‌آید. در ادامه چندین مثال را مشاهده می‌کنید:

€200 $800.00 $1,222 8$ $60k £300,75 ¢400’85 €600.90m €1000.10mill 2000.20Million€ $1800’12b

#	Unit Symbol	Unit
1	K k	Thousand
2	M, Mill, Million m, mill, million	Million
3	B, Bill, Billion b, bill, billion	Billion
4	T, Trill,Trillion t, trill, trillion	Trillion

جدول ۶: سیستم شمارش پول

Time ۷.۴

استخراج و جایگزینی زمان (ساعت) یکی دیگر از مراحل پیش‌پردازش است. کلید منحصربه‌فردی که برای این قسمت در نظر گرفته شده است داری چنین ساختاری است:

$$\triangleright T_{counter} \triangleleft$$

به عنوان مثال، اولین ساعت با $\triangleright T1 \triangleleft$ جایگزین می‌شود، دومین ساعت با $\triangleright T2 \triangleleft$ جایگزین می‌شود و به همین ترتیب الی آخر. حرف T برگرفته از کلمه Time است. بعد از عملیات جایگزینی، زمان به همراه کلید مربوطه در یک فایل ذخیره می‌شود. با این کار می‌توان آن را در صورت نیاز بازیابی کرد. نمونه‌هایی از ساعت که به عنوان زمان شناسایی و با کلید مشخصی جایگزین می‌شوند، در جدول ۷ ذکر شده‌اند.

Type	Numbers	Postfix
Short Time	Any number separated by dot(.)	AM PM am pm A.M. P.M. a.m. p.m.
Full Time	Numbers [0:23] : Numbers [0:59] : Numbers [0:59]	

جدول ۷: سیستم تشخیص ساعت

Repeated Letter ۸.۴

معمولا هنگام چت بعضی از حروف را به صورت تکراری می‌نویسند که می‌تواند دلایل متفاوتی داشته باشد، مثل:

- Michael Erard : «جبران کمبود نشانه‌های صوتی هنگام نوشتن به جای صحبت کردن»
- ابراز کنایه یا احساسات منفی
- متقاعدسازی: از نظر علمی ثابت شده است که تکرار اطلاعات احتمال تغییر ذهن افراد را افزایش می دهد.

یکی دیگر از مراحل پیش‌پردازش یافتن و حذف این حروف اضافی از متن هست. در زبان انگلیسی کلماتی داریم که دو حرف پشت سر هم تکرار شدند و آن کلمه معنادار است (مثلا Good) به همین دلیل مبنای تعداد حروف پشت سر هم، حداکثر دو است. پس اگر حرفی بیش از دو بار متوالی تکرار شده باشد، کل آن حروف را با دو حرف تعویض می‌کنیم، به عنوان مثال:

Goooooood ⟼ Good

دقت بفرمایید که در این قسمت، تنها حروف انگلیسی چک نمی‌شوند بلکه کاراکتر Space نیز چک می‌شوند و اگر در متن مجموعه‌داده، بیش از دو فاصله متوالی باشد، همگی با دو فاصله جایگزین می‌شوند. در نتیجه فضاهای خالی موجود در متن نیز در این قسمت فیلتر می‌شوند. به عنوان مثال:

`This␣␣is␣a␣␣␣␣big␣␣␣␣␣␣␣␣window.` ⟼ `This␣␣is␣a␣␣big␣␣window.`

۵ بسط دادن

بخش سوم، مربوط به گسترده کردن یا توسعه دادن برخی کلمات و یا کاراکترها است.

Mention ۱.۵

طبق تعریف شرکت ایکس: «مِنشن توییتی شامل نام کاربری حسابی دیگر است که قبل از آن، از علامت اَت ساین (@) استفاده شده است» [۱] . در این قسمت از پیش‌پردازش، هدف گسترش نام کاربری است. معمولا نام کاربری می‌تواند یک کلمه یا ترکیبی از چندین کلمه باشد که به وسیله خط فاصله[1] از

[1] Underline

همدیگر جدا شدند. هدف در این قسمت ابتدا شناسایی نام کاربری به کمک علامت @ است. سپس چک می‌کنیم که اگر از چندین کلمه و خط فاصله استفاده شده باشد، کلمات را به وسیله خط فاصله از همدیگر جدا و جایگزین نام کاربری اولیه می‌کنیم و در نهایت علامت @ را حذف می‌کنیم.

Hey, @dark_web This sentence is just for test.

Hey, dark web This sentence is just for test.

۲.۵ Hashtag

هشتگ معمولا از یک یا چندین کلمه تشکیل شده که به وسیله خط فاصله یک کلمه واحد را تشکیل می‌دهند. علامت مخصوص هشتگ «#» است. در این بخش از پیش‌پردازش، هدف پیدا کردن هشتگ‌ها و جدا کردن کلمات (در صورت امکان) و سپس جایگزین کردن با کلمه اولیه و حذف علامت هشتگ است.

Hey, #dark_web_2024 This sentence is just for test.

Hey, dark web 2024 This sentence is just for test.

۶ ذخیره مجموعه‌داده

بعد از انجام مراحل ذکر شده در قسمت‌های قبلی با ترتیبی که در جدول ۱ آمده، یک مجموعه‌داده تمیز و مرتب خواهیم داشت. فایل‌های پیش‌پردازش شده به صورت جداگانه در قالب یک فایل CSV ذخیره می‌شوند، تا در مراحل بعدی کار از آن‌ها استفاده شود.

۷ مراجع

[1] X (formerly Twitter). Glossary – X help center. https://help.x.com/en/resources/glossary, 2025. Accessed: 2025-04-05.

[2] S. J. Whitmore. tweet-collector – GitHub repository. https://github.com/sjwhitmore/tweet-collector, 2025.