دانشگاه علم و صنعت ایران
مبانی پردازش زبان و گفتار

فاز اول پروژه

سید حمیدرضا ثنائی کرهرودی
۹۶۵۲۱۱۳۷
دانشکده مهدسی کامپیوتر
hamid_sanaee@comp.iust.ac.ir

چکیده

موضوع پروژه مقایسه forum و discussion های بازی های معروف multiplayer به طور خاص Counter_Strike:Global_Offensive و Tom_Clancy's_Rainbow_Six_Siege است. در فاز اول پروژه قصد ما جمع آوری اطلاعات و مرتب سازی و استخراج آمار از داده های به دست آمده است.

۱ منابع

از لینک های زیر جهت استخراج داده استفاده شده است:

- Counter_Strike:Global_Offensive
- Tom_Clancy's_Rainbow_Six_Siege

۲ جمع آوری داده

ز python scrapy جهت جمع آوری داده ها استفاده شده است. روش کار به این صورت هست که در لینک که قرار گرفته شده لیستی از discussion ها درمورد بازی موردنظر با topic های مشخص قرار گرفته است که تمامی آنها به صورت paginate شده هستند و در هر صفحه میتوان لینک صفحه بعد برای ادامه کار را یافت. به ترتیب وارد هر discussion میشویم و title و comment های مورد نظر را استخراج میکنیم. این قسمت نیز paginate شده است اما امکان پیدا کردن لینک صفحه بعد وجود ندارد اما خب با کمی دقت متوجه میشویم که در صفحه بعدی تنها به مقدار ctp در url ما یک عدد اضافه میشود که با استفاده از همین می توان کار را ادامه داد.

۳ ذخیره‌سازی داده

تمامی داده های ما در پوشه data درکنار پوشه src قرار میگیرند به طور مثال داده های بازی Counter_Strike:Global_Offensive پس از استخراج با برچسب csgo داخل فایل csgo.json و داده های بازی Tom_Clancy's_Rainbow_Six_Siege با برچسب r۶ در r۶.json در پوشه data قرار میگیرند. بعد داده های تمیز شده توسط فایل clean_data در پوشه pre_processors در فایل cleane_csgo و cleaned_r۶ ریخته میشود. بعد جمله ها و کلمات توسط فایل های word_tokenizer و sent_tokenizer در پوشه pre_processors در فایل _sentences و _words ریخته میشود.

۴ پیش‌پردازش های انجام‌شده

- معیار های تمیز کردن داده: در داده های ما متن های غیر زبان انگلیسی و emoji مشکل به وجود می آورد که برای حذف متن های غیر انگلیسی از pycld۲ استفاده شده و برای حذف هاemoji با توجه به آنکه encode آنها فرق دارد مجبور شدم با regex حروف انگلیسی و اعداد را از آن جدا کنم.
- ابزار تفکیک جملات و توکن ها: برای جدا سازی جملات و کلمات از پکیج nltk استفاده شده است، از nltk.tokenize.sent_tokenize و nltk.tokenize.word_tokenize استفاده شده است.

- اندازه داده ها قبل از تمیز کردن داده ها ۳/۶ mg برای csgo و ۲/۵ mg برای r۶ هست.
  اندازه داده ها بعد از تمیز کردن داده ها ۱/۷ mg برای csgo و ۲/۳ mg برای r۶ هست.

۵ واحد برچسب‌گذاری

واحد برچسب گذاری برای داده های ما برای هر comment یا title هست. روش ما آن است که پس از دریافت هرکدام از آنها توسط scrapy برچسب مناسب (csgo،r۶) را به آن اضافه م

۶ آمار

جداول و نمودار های آماری را میتوانید در صفحه بعد مشاهده کنید. آماری که به شکل جدول در صفحه بعد میبینید را میتوانید پس از انجام مراحل اولیه با ران کردن فایل extract_stats.py یا ران کردن run.sh برای داده های خود مشاهده کنید.

	Offensive Global Strike Counter	Siege Six Rainbow Clancy’s Tom
تعداد کامنت‌های داده	۷۲۶۲	۷۲۳۸
تعداد جملات	۱۷۲۰۶	۲۰۳۰۲
تعداد کلمات	۳۲۶۶۶۳	۴۶۵۰۸۴
تعداد کلمات منحصر به فرد	۱۳۸۱۵	۱۵۸۶۸

کلمات منحصر به فرد مشترک	کلمات منحصر به فرد غیر مشترک
۵۵۷۲	۱۸۵۳۹

کلمه پر تکرار غیر مشترک برای CSGO	تعداد تکرار
valve	۲۸۵
sof	۲۲۲
CMsgSteamDatagramRouterPingReply	۲۱۳
۱۰:۲۷۰۱۹.۱۹۳.۴۵.۱۳۹	۱۲۳
Relay	۱۰۷
.Relay	۹۸
subscription	۹۷
chickens	۸۷
fang	۸۰
Bart	۷۸

کلمه پر تکرار غیر مشترک برای Rainbowً	تعداد تکرار
operators	۴۲۰
TwisterCat	۳۲۴
operator	۳۲۴
gadget	۲۵۴
siege	۲۳۳
renown	۲۲۳
Jager	۲۰۵
Ubi	۲۰۵
ubisoft	۱۸۱
Rainbow	۱۶۵

کلمه	relative_normalized_frequency_Rainbow
Ubisoft	۳۵۶/۸۰۶۰۹۰۹۴۲۷۱۱۴
rework	۱۷۲/۷۸۴۰۵۱۹۱۳۲۰۲۷۷
R۶	۱۴۸/۲۰۰۹۵۵۰۹۶۲۸۳۶۸
Siege	۱۲۵/۹۵۹۱۰۵۵۹۵۲۶۱۶۶
ongoing	۳۹۲۰۲۵۹۵۶۶۰۱۳۹.۸۶
Legendary	۸۴/۲۸۴۹۰۳۳۷۲۲۹۴۰۲
r۶	۶۳/۲۱۳۶۷۷۵۲۹۲۲۰۵۲۶
battlepass	۵۹/۷۰۱۸۰۶۵۵۵۳۷۴۹۴
armor	۵۶/۱۸۹۹۳۵۵۸۱۵۲۹۳۵۵
PvE	۵۵۴۸۷۵۶۱۳۸۶۷۶۰۲۴

relative_normalized_frequency_CSGO	کلمه
۹۵/۸۶۵۳۲۹۱۰۰۶۳۲۷۶	overwatch
۹۵/۳۹۰۷۴۸۲۶۳۵۰۰۹۲	hack
۹۳/۹۶۷۰۰۵۷۵۲۱۰۵۳۸	Broken
۸۶/۳۷۳۷۱۲۳۵۷۹۹۵۸۶	knife
۸۱/۱۵۳۳۲۳۱۴۹۵۴۵۵۶	windowed
۷۹/۲۵۴۹۹۹۸۰۱۰۱۸۱۷	prime
۶۶/۷۳۷۹۳۰۲۲۱۶۶۵۷۶	VAC
۵۸/۳۷۳۴۴۲۹۶۷۲۱۶۹۸	websites
۵۵/۵۲۵۹۵۷۹۴۴۴۲۵۹۱	}
۵۴/۱۰۲۲۱۵۴۳۳۰۳۰۳۷	agent