پروژه ی پایانی درس پردازش زبان های طبیعی

فاز اول

کیوان بوشهری - نگین درخشان

منتور : رضا قهرمانی

بهار 1401

آدرس گیت پروژه:

https://github.com/k1booshehri/TelegramNLP

در این پروژه از دیتای موجود در تلگرام استفاده شده است. به این منظور از طریق api دیتا را جمع آوری می کنیم.
برای استفاده از api تلگرام در برنامه پایتون از کتابخانه telethon استفاده می کنیم. این کتابخانه یک ماژول telegram client است که از جدیدترین api موجود تلگرام استفاده می کند. کلاس های استفاده شده از این کتابخانه به صورت زیر است:

```
import csv
import re
import json
import asyncio
import configparser
import pandas as pd
from csv import writer
from csv import reader
from datetime import date, datetime
from telethon import TelegramClient
from telethon.errors import SessionPasswordNeededError
from telethon.tl.functions.messages import (GetHistoryRequest)
from telethon.tl.types import (PeerChannel)
```

ترتیب اینکار به صورت اتصال به یک کانال تلگرامی و دریافت پیامهای موجود در آن است. دیتای جمع آوری شده به این روش در قالب JSON است که برای استفاده ساده به csv تبدیل می کنیم.

ساختار فایلها به صورت فولدارهای data و source است. در فولدر source فایل ها به صورت زیر هستند:

اسکریپت ران شدن برنامه ها و نصب کتابخانه	Script.sh
دریافت داده از کلاینت تلگرام	Main.py
اطلاعات مربوط به کلاینت	Config.ini
Session کاربر	X.session
پیش پردازش داده	Preprocess.py

در فولدر data داده ها بصورت زیر مرتب شده اند:

دیتای json	sports_messages.json

sports_csvfile.csv	دیتای تبدیل شده به csv
sports_output.csv	دیتای تگ دار
sports_first_edit.csv	دیتای مرتب شده
sports_preprocessed.csv	دیتای پیش پردازش شده
technews_messages.json	دیتای json
technews_csvfile.csv	دیتای تبدیل شده به csv
technews_output.csv	دیتای تگ دار
technews_first_edit.csv	دیتای مرتب شده
technews_preprocessed.csv	دیتای پیش پردازش شده

برچسب ها از طریق کانال های مختلف تلگرامی و موضوعات آنها قرار داده شده اند.

پیش پردازشها:

تمام کاراکترهای اضافه از جمله @ و # و ایموجی ها حذف شده اند.

تمام stopword ها حذف شده اند.

روشها و ابزار های تفکیک و تمیز کردن داده:

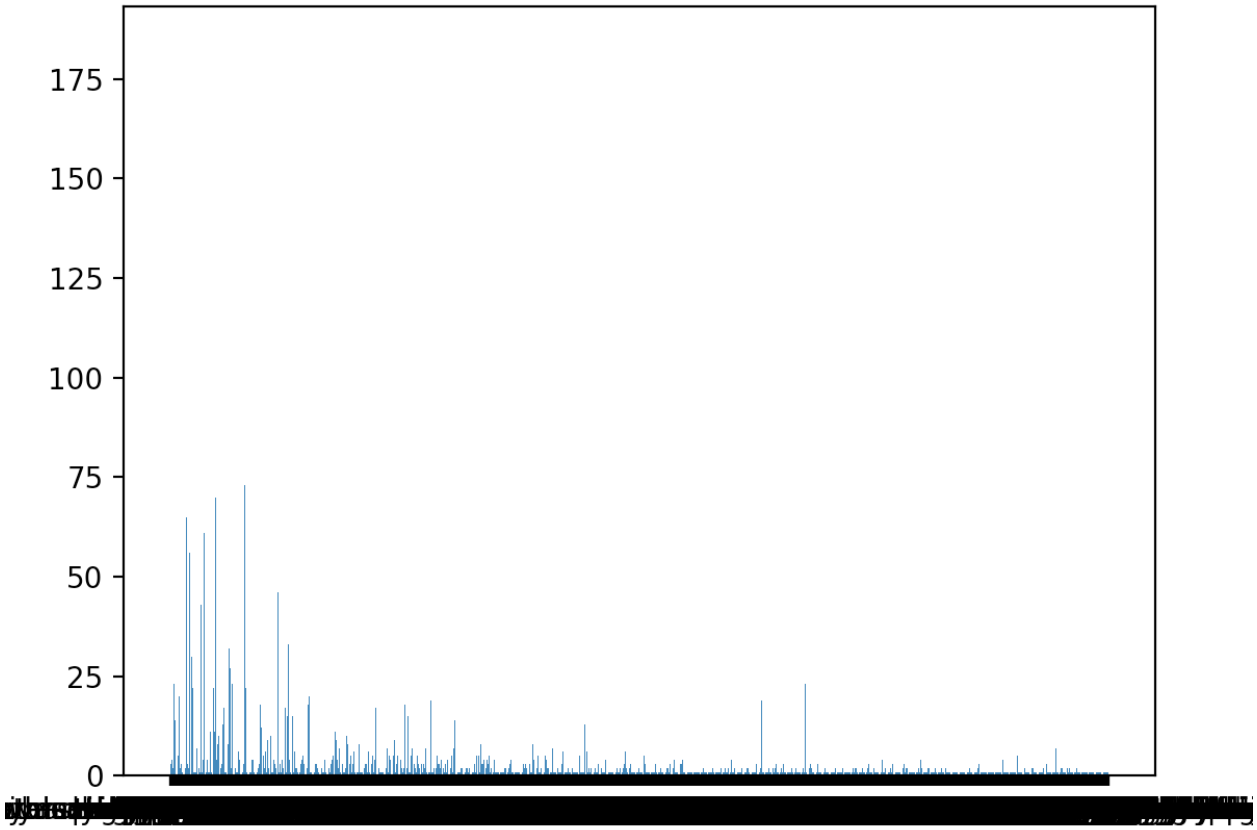
از ابزار nltk برای Tokenize و regex برای پیش پردازش داده استفاده شده است.

آمار داده ها:

```
number of words:
14682
number of types:
4636
number of sentenses:
988
number of words:
17747
number of types:
3911
number of sentenses:
1000
```

```
Client Created
Current Offset ID is: 0 ; Total Messages: 0
Current Offset ID is: 2108 ; Total Messages: 100
Current Offset ID is: 2007 ; Total Messages: 200
Current Offset ID is: 1906 ; Total Messages: 300
Current Offset ID is: 1805 ; Total Messages: 400
Current Offset ID is: 1704 ; Total Messages: 500
Current Offset ID is: 1604 ; Total Messages: 600
Current Offset ID is: 1504 ; Total Messages: 700
Current Offset ID is: 1403 ; Total Messages: 800
Current Offset ID is: 1301 ; Total Messages: 900
Client Created
Current Offset ID is: 0 ; Total Messages: 0
Current Offset ID is: 15948 ; Total Messages: 100
Current Offset ID is: 15839 ; Total Messages: 200
Current Offset ID is: 15733 ; Total Messages: 300
Current Offset ID is: 15629 ; Total Messages: 400
Current Offset ID is: 15529 ; Total Messages: 500
Current Offset ID is: 15423 ; Total Messages: 600
Current Offset ID is: 15320 ; Total Messages: 700
Current Offset ID is: 15213 ; Total Messages: 800
Current Offset ID is: 15106 ; Total Messages: 900
```

liverpool
final
india
team
first
football
beat
tennis
season
live
year
semifinal
match
title
goal
said
win
chelsea
club
score
lead
women
indian
second
champion
sunday
game
time
saturday
much dailybriefing
tournament
thomascup
thursday
kerala
rafael nadal
highlights
test
cricket
england
player
return
back
world
injury
arsenal
real madrid
open
three
villarreal
friday
victory
point
need
mens
quarterfinal
french open
mumbai indians
champions league
look
four
premier league
premierleague
realmadrid
read
playoff
take
side
mancity
defeat
details
carlos alcaraz
barcelona
semifinal
scored


175
150
125
100
75
50
25
0

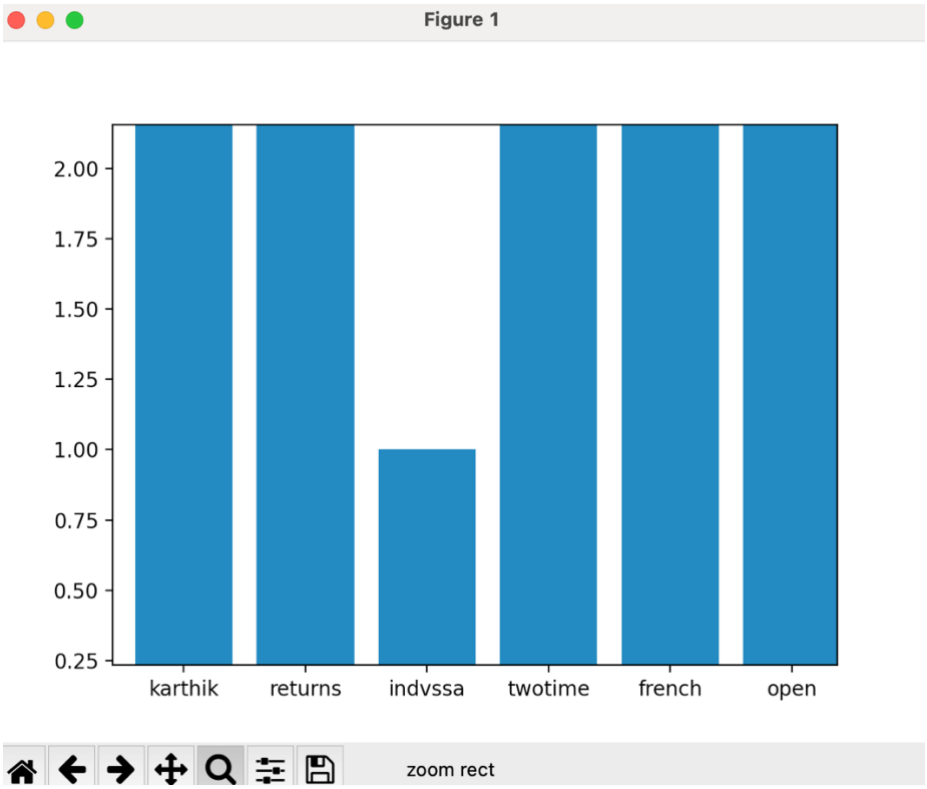
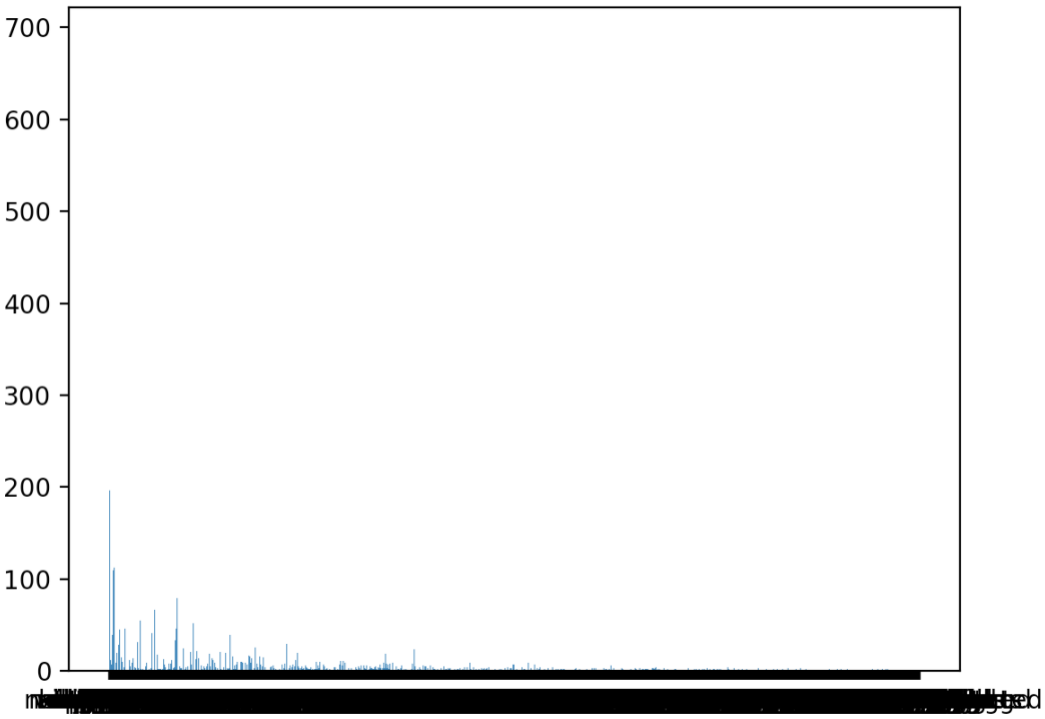
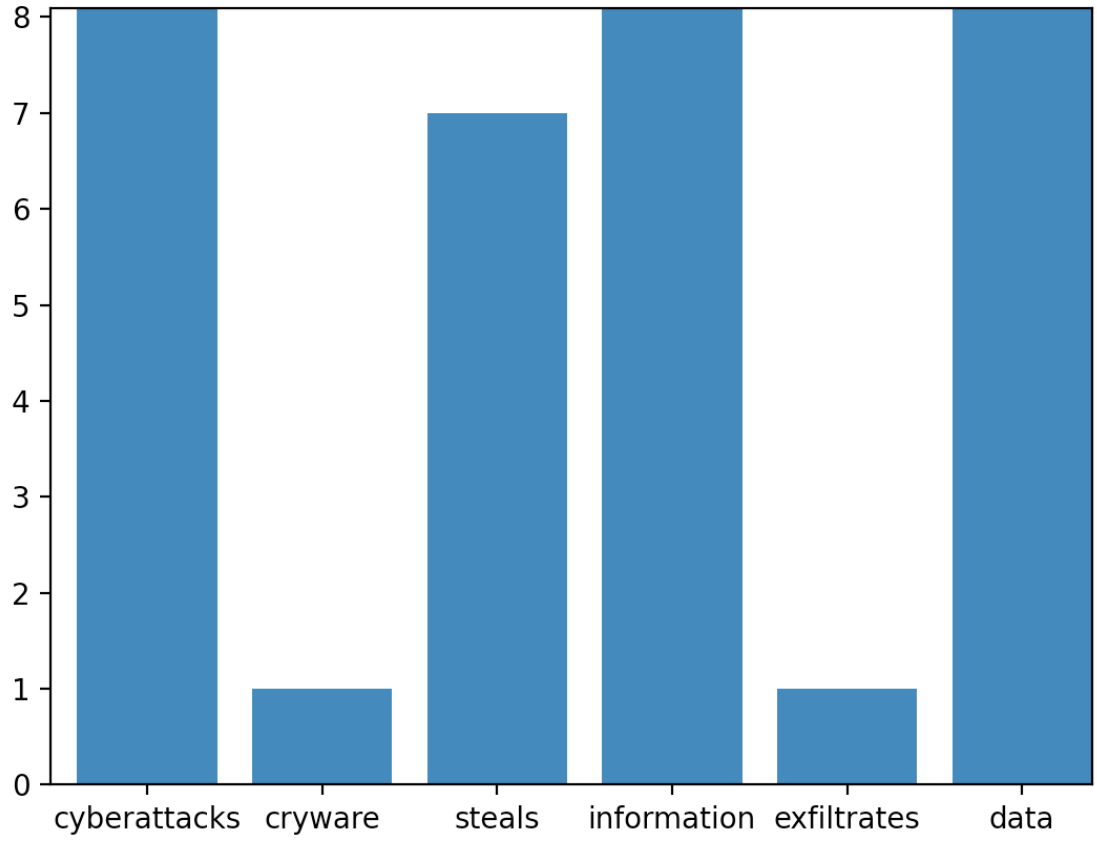


Figure 1
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
karthik
returns
indvssa
twotime
french
open
zoom rect


campaign
government
number
company
windows
platform
attacker
technique
malware
security
cyberattack
new
detail
spread
tool
apps
attacks read
hacker
actively exploited
email
russian
year
computer
researchers discovered
unpatched
read details
allow attacker
arbitrary code
gain
security vulnerabilities
android
details microsoft
targeted
data
attack
browser
patch
warning
ukraine
compromised
read researchers
targeting
victim
devices read
privacy
service
companies
infect
credential
customer
called
organization
dubbed
file
macos
activities
read
vulnerability
using
steal
researchers uncover
another
popular
apple
information
stolen
account
server
access
system
execute arbitrary
part
newly discovered
device
exploited
network
several
hacker group
systems read
feature
ransomware
critical vulnerability
exploit
released
website
issued
compromise
vulnerability affecting
user
target
microsoft
hacking
version
critical vulnerabilities
botnet
million
malicious
cryptocurrency
threat
cybercriminal
known
used
multiple
vulnerabilities affecting
details hacker
release
remotely
details researcher
update
month
warn
backdoor
google
discovered
flaw
united states
series
behind
allow
including
hacking group
software
zeroday vulnerability
vulnerabilities


700
600
500
400
300
200
100
0


8
7
6
5
4
3
2
1
0
cyberattacks
cryware
steals
information
exfiltrates
data