

**پروژه ی پایانی درس پردازش زبان های طبیعی**

**فاز اول**

کیوان بوشهری - نگین درخشان

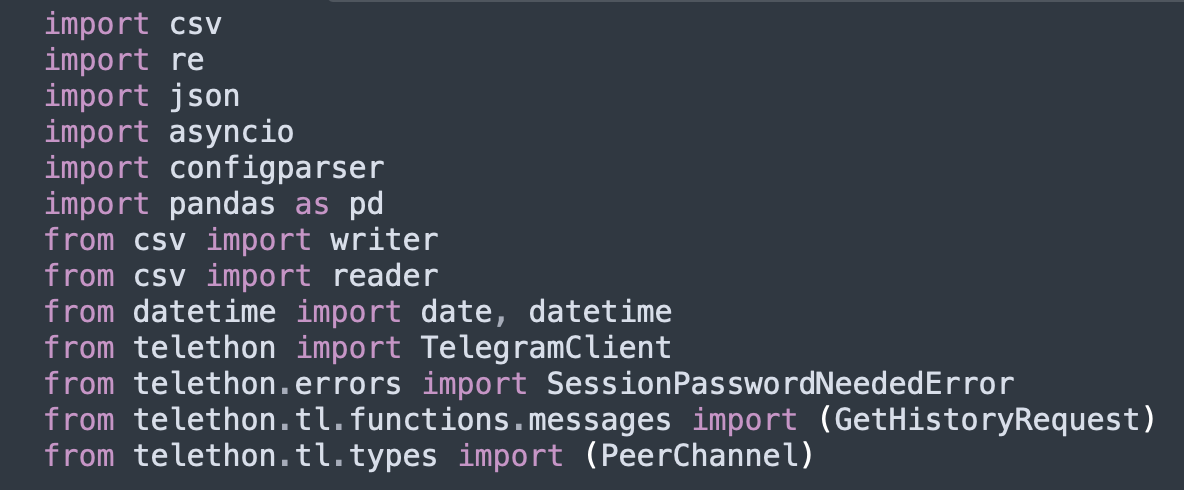
منتور : رضا قهرمانی

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:



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 |
| دیتای تبدیل شده به csv | sports\_csvfile.csv |
| دیتای تگ دار | sports\_output.csv |
| دیتای مرتب شده | sports\_first\_edit.csv |
| دیتای پیش پردازش شده | sports\_preprocessed.csv |
| دیتای json | technews\_messages.json |
| دیتای تبدیل شده به csv | technews\_csvfile.csv |
| دیتای تگ دار | technews\_output.csv |
| دیتای مرتب شده | technews\_first\_edit.csv |
| دیتای پیش پردازش شده | technews\_preprocessed.csv |

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:**

