به نام خدا

دانشگاه اصفهان

دانشکده مهندسی کامپیوتر

پروژه پایانی درس شبکه‌های اجتماعی

نام و نام خانوادگی اعضای تیم:

کسرا صمدی دارستانی – ۹۹۳۶۲۳۰۳۰

مهسا آقایی دهنوی – ۹۹۳۶۲۳۰۰۳

نام استاد:

جناب آقای امیرحسین خانیکی

لینک گیت‌هاب پروژه:

https://github.com/kasraSMD/SN_Project

دلیل استفاده از دیتاست[1]:

در ابتدا، تلاش‌هایی برای استفاده از API توییتر[2] و اینستاگرام[3] را آغاز کردیم، اما متأسفانه موفقیت‌آمیز نبودند. علت این امر به دو عامل برمی‌گردد. عامل اول، دسترسی به API توییتر با هزینه‌ای بالا همراه بود. همچنین، اینستاگرام نیز API خود را محدود کرده بود، بنابراین امکان استفاده از آن نیز وجود نداشت. با توجه به محدودیت‌های زمانی و نیز توضیحات استاد درباره این مسئله و موافقت ایشان برای استفاده از دیتاست، تصمیم گرفتیم از یک دیتاست آماده استفاده کنیم.

توضیحات پروژه و راه‌حل‌های پیاده‌سازی:

- برای انجام این پروژه، از کتابخانه‌های pandas، numpy و matplotlib.pyplot استفاده کردیم. این کتابخانه‌ها ابزارهای قدرتمندی برای تحلیل داده‌ها، محاسبات عددی و تجسم‌سازی ارائه می‌دهند.
- در ادامه، داده‌ها را از فایل "tweets_dataFrame.csv" خواندیم. یک نکته قابل توجه در این مرحله این است که دیتاست اصلی حاوی حدود ۶۰۰ هزار رکورد[4] بود، اما ما آن را به تعداد حدود ۱۴۷ هزار رکورد کاهش دادیم. این کاهش حجم دیتاست به منظور سهولت در پردازش و کارایی بیشتر در طول تحلیل‌های بعدی انجام شده است.
- همچنین، رکوردهایی که داده‌هایی با مقدار null داشتند را نیز حذف کردیم. این اقدام انجام شد تا دیتاست ما فقط شامل داده‌های کامل و قابل اطمینان باشد.

[1] dataset
[2] twitter
[3] Instagram
[4] Record

- اطلاعات مربوط به ستون‌های هر رکورد و تعداد null های مربوط به آن‌ها:

```
ID                          0
user_name                   0
user_location           85969
user_description        27934
user_created                0
user_friends                0
user_friends_username       0
user_favourites             0
user_verified               0
date                        0
text                        0
hashtags                16213
source                      0
dtype: int64
```

با توجه به توضیحات و خواسته پروژه، ۵۰۰ نود را از یوزرنیم[5]های یونیک[6] و غیر تکراری انتخاب کرده (resample کرده‌ایم) و آن‌ها را به یک گراف تبدیل کردیم (ایجاد گراف بوسیله‌ی کتابخانه network انجام شده است). در این گراف نودها همان آیدی[7] ها و یال ها، ارتباط بین آیدی‌ها هستند.

- نودها را بر اساس معیار های اهمیت و مرکزیت آنها مرتب کردیم و ۵ نود با اهمیت و مرکزیت بالاتر را معرفی کردیم.

```
Top 5 nodes based on Degree Centrality:
Node: 35821, Degree Centrality: 0.012261463796286108
Node: 59543, Degree Centrality: 0.012249239006858106
Node: 77074, Degree Centrality: 0.012249239006858106
Node: 39908, Degree Centrality: 0.0122125646385741
Node: 82851, Degree Centrality: 0.0122125646385741
```

[5] Username
[6] Unique
[7] ID

- برای هر توییت[8]، تمامی حروف متن را به حالت Lowercase تبدیل کردیم و علائم نگارشی آن را حذف کردیم. این اقدام انجام شد تا کلمات در توییت‌ها به شکل استاندارد و یکنواختی باشند. سپس متن توییت را به واحدهای کوچکتر تقسیم کرده و آن‌ها را به صورت توکن‌ها شناسایی کردیم (Tokenize) و سپس نتیجه را در ستون جدیدی به نام text_tokens ذخیره کردیم.
- سپس، Stopwords فارسی و انگلیسی را از متن حذف کردیم. Stopwords کلماتی هستند که در تحلیل متنی معمولاً ارزش اطلاعاتی کمتری دارند و می‌توانند اثر منفی بر تحلیل‌ها داشته باشند. با حذف این کلمات، متمرکز شدن بر کلمات با اهمیت بیشتر و کاربردی‌تر انجام می‌شود.
- سپس، ۱۰۰ تا از متداول‌ترین و پرکاربردترین کلمات و همچنین ۱۰۰ مورد از کم کاربردترین کلمات در متن توییت‌ها را شناسایی کردیم و آن‌ها را نمایش دادیم.

نمونه ای از کلمات پرکاربرد و تعداد تکرار آنها:

```
iran: 53870
iranprotests: 37241
people: 24437
voice: 23332
regime: 22917
support: 22015
```

نمونه ای از کلمات کم کاربرد و تعداد تکرار آنها:

```
httpstcos1hh3zh83q: 1
httpstco4llxnmomrz: 1
httpstcobajn7e8byw: 1
httpstcoruhu1sfdma: 1
httpstco81x2vm9sjp: 1
httpstco3v6cdlvdbb: 1
httpstcow74reoxzrp: 1
```

[8] tweet

- برای پیداکردن هشتگ[9]های استفاده شده، آمار تعداد تکرار هر هشتگ را محاسبه کردیم. هشتگ‌ها کلماتی هستند که معمولاً در توییت‌ها برای بیان موضوعات و دسته‌بندی‌های مختلف استفاده می‌شوند. با پیداکردن هشتگ‌های پرتکرار، می‌توانیم موضوعات مهم و پربحث را شناسایی کنیم.
- سپس، آیدی کاربران فعال را بر اساس تعداد توییت‌های آنها پیدا کردیم. در واقع، با فرض اینکه کاربران با تعداد توییت بیشتر فعالیت بیشتری دارند، کاربرانی را که تعداد توییت‌های بالاتری داشته‌اند، به عنوان کاربران فعال شناسایی کردیم. این اقدام می‌تواند به ما در درک الگوها و رفتارهای کاربران در شبکه اجتماعی کمک کند.

آی‌دی برخی از کاربران فعال و تعداد توییت‌های آنها:

```
Active Users Based on The Number Of Tweets:
ID=326: 14837
ID=290: 4420
ID=4: 2249
ID=30: 1725
ID=186: 1656
ID=15: 1606
ID=97: 1346
ID=3: 1149
ID=20: 924
ID=468: 850
ID=27: 742
ID=84: 737
ID=286: 699
ID=141: 696
ID=8: 679
ID=210: 670
ID=31: 631
ID=7: 630
ID=90: 561
ID=79: 530
ID=133: 483
```

- ستون‌های 'Positive words' , 'sentiment_label' , 'compound_score' , 'Negative words'را برای پردازش متن به عنوان ۴ ستون جدید به دیتاست اضافه کردیم که درادامه توضیحات مربوط به هر کدام از آنها را بیان میکنیم:
  - ستون 'compound_score': مقدار اولیه این ستون ۰ است. در ادامه برای متن هر توییت امتیازدهی صورت میگیرد و مقدار این امتیاز در این ستون قرار می‌گیرد.

[9] hashtag

- ستون 'Positive words': لیست کلمات مثبت موجود در هر توییت را استخراج می‌کنیم. این کلمات معمولاً واژگانی هستند که به احساسات و نظرات مثبت اشاره می‌کنند.
- ستون 'Negative words': لیست کلمات منفی موجود در هر توییت را استخراج می‌کنیم. این کلمات معمولاً واژگانی هستند که به احساسات و نظرات منفی اشاره می‌کنند.
- ستون 'sentiment_label': با توجه به امتیاز محاسبه شده در ستون 'compound_score'، یک لیبل برای آن در نظر میگیریم به صورت زیر:

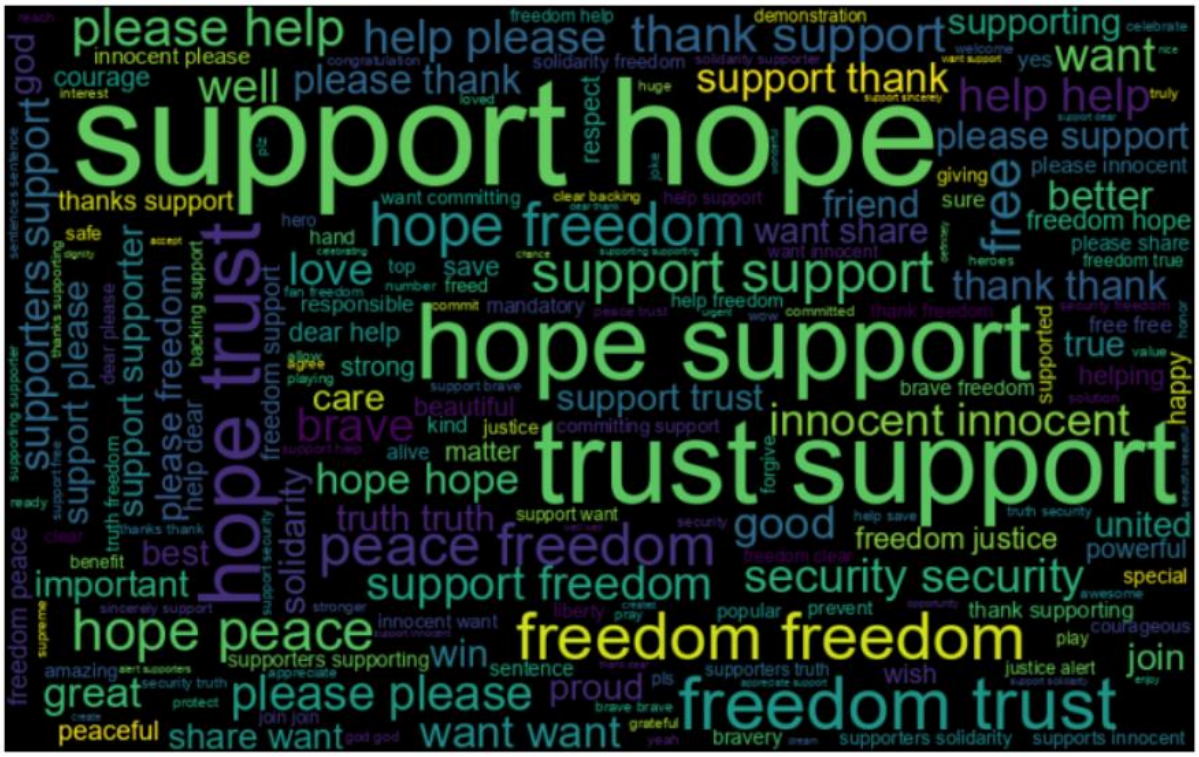
```
# if compound_score > 0:
#         main_df.at[index, 'sentiment_label'] =
"Positive"
#     elif compound_score < 0:
#         main_df.at[index, 'sentiment_label'] =
"Negative"
#     else:
#         main_df.at[index, 'sentiment_label'] =
"Neutral"
```

در واقع در صورتی که امتیاز محاسبه مثبت باشد لیبل Positive، در صورتی که امتیاز محاسبه شده منفی باشد، لیبل Negative و در صورتی که این امتیاز برابر با صفر باشد لیبل Neutral به منظور بار معنایی خنثی در نظر گرفته می‌شود.

نکته مهم و قابل توجه در این بخش: به دلیل زمان‌بر بودن اجرای این بخش (حدود ۱۰ ساعت)، نتیجه (DataFrame جدید) را در فایل " Positive_Negative_Words.csv" ذخیره کردیم و برای استفاده بعدی از این فایل استفاده می‌کنیم. این کار اجرای مجدد این بخش را غیر ضروری می‌کند.

- در ادامه تمامی کلمات مثبت و منفی مربوط به تمامی رکوردها را به کمک ستون‌های مربوط به کلمات مثبت و منفی (که در قسمت‌های قبل ایجاد کردیم و توضیحات مربوط به آنها ارائه شد) پیدا کردیم و به کمک آن، ابر کلمات مثبت و منفی را ساختیم. که در ادامه تصویر آن‌ها قابل مشاهده است.

ابر کلمات مثبت:



ابر کلمات منفی: