تمرین دوم:

مقایسه‌ی پست‌های فالوئرهای دو شخصیت مطرح در توئیتر

- درس: پردازش زبان‌های طبیعی
- استاد: آقای دکتر صالح اعتمادی
- دانشجو: امیرحسین کاظم‌نژاد - ۹۴۵۲۳۲۱۶

دیتاست

- جمع‌آوری

این دیتاست از فالوئرهای دو فرد به نام‌های

«حمید رسایی - نماینده‌ی اصولگرای مجلس سابق» @hamidrasaee

«دکتر محمدجواد اکبرین - از فعالین اصلاح‌طلب و اپوزسیون» @MJ_Akbarin

جمع‌آوری شده است.

نحوه‌ی جمع‌آوری به این صورت است که برای هر کدام از این دو نفر پست‌هایشان را جمع‌آوری می‌کنیم و برای هر پست، کسانی که آن‌را ریتوئیت کردند پیدا می‌کنیم. سپس از بین این افراد آن‌هایی که فالوئر هستند انتخاب می‌شوند و توئیت‌های آنان جمع‌آوری می‌شود.

برای انتخاب افراد هم این نکته رعایت شد که اکانت آن‌ها متعلق به خبرگزاری‌ها و خبرنگارها نباشد تا از بایاس شدن دیتا جلوگیری شود.

برای جمع‌آوری توئیت‌های هر فرد هم ابتدا بررسی می‌کنیم که چند درصد آن‌ها ریتوئیت است، اگر بیشتر ۲۰ درصد باشد فقط توئیت‌های تولید شده توسط آن فرد را بر می‌داریم.

- تمیز کردن

دیتاست موجود برای پردازش به نسبت دیتاست کثیفی است. بنابراین نیاز به کار زیادی در این قسمت داریم.

برای تمیز کردن ابتدا تمام لینک‌های توئیت را حذف می‌کنیم سپس تمامی کارکترهای انگلیسی را نیز پاک می‌کنیم. بعد از آن منشن‌ها را نیز از بین می‌بریم و سپس تمام کارکترهای عربی را جایگزین می‌کنیم. برای اعداد هم تمامی ارقام (از جمله فارسی و عربی) را با معادل انگلیسی آن جایگزین می‌کنیم.

در مرحله‌ی بعد تمامی علامت‌های جمع «ها» را با کلمه‌ی‌قبل + نیم‌فاصله + ها جایگرین می‌کنیم. این کار را برای تمامی «می» استمراری هم انجام می‌دهیم.
برای سایر نیم‌فاصله‌ها هم آن‌ها را حذف می‌کنیم تا تمام کلمات ما یک‌دست شود.
سپس برای هر کدام لیستی از کلمات اضافه و فعل‌های بی‌معنی تشکیل داده و این کلمات را از دیتاست حذف می‌کنیم.
برای هر توئیت هم دقت می‌کنیم که حتما زبان آن فارسی باشد. (این فیلد توسط خود توئیتر فراهم شده‌است.)

دیتاست از لحاظ آماری

دیتاست مربوط به حمید رسایی شامل ۴۹۵۱۱۰۴ عبارت و ۲۸۳۰۵۱ عبارت یکتا می‌باشد.
دیتاست مربوط به محمدجواد اکبرین شامل ۱۸۱۹۴۹۶۹ عبارت و ۶۴۴۳۳۹ عبارت یکتا می‌باشد.

ابزارهای مورد استفاده

- کتاب‌خانه‌ی python-twitter و tweepy برای برقراری ارتباط با توئیتر
- کتاب‌خانه‌ی persian پایتون برای یکسان‌سازی حروف عربی
- کتاب‌خانه‌ی hazm پایتون، برای توکنایز و جدا کردن کلمات
- ابزار Kumo برای تولید wordmap

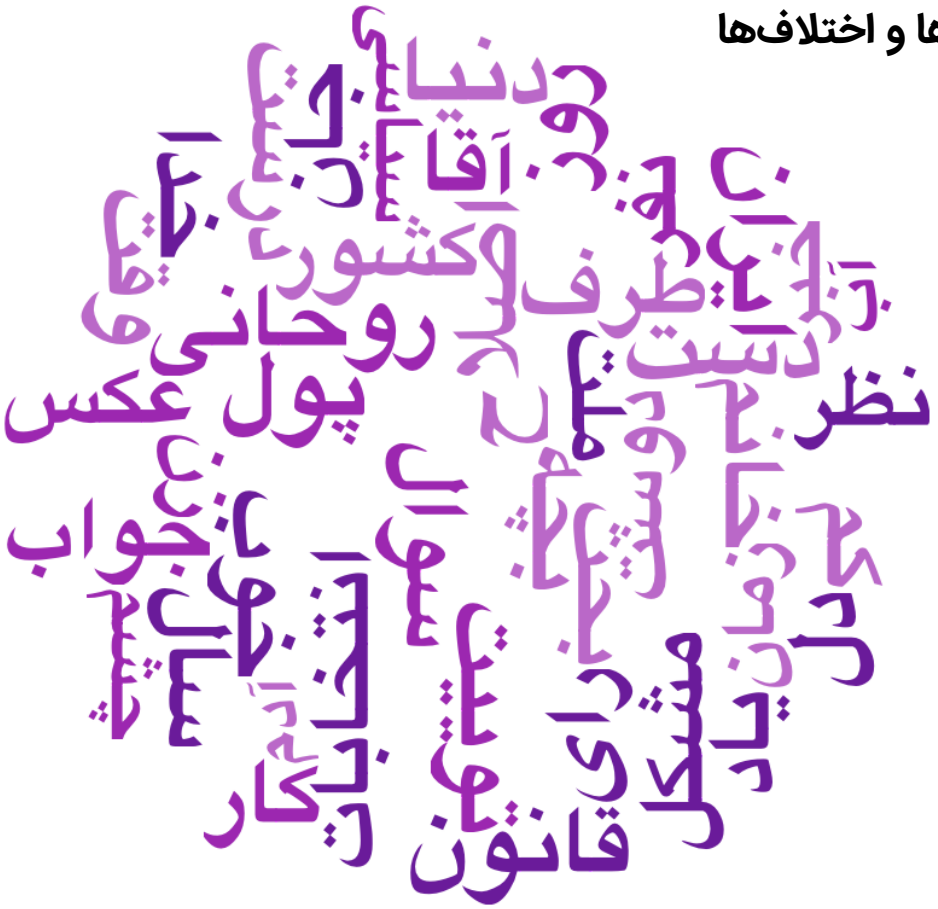
نتایج

گروه رسایی

گروه اکبرین

همانطور که دیده می‌شود، گروه رسایی محتوایی بیشتر سیاسی دارند و شاید استفاده‌ی آن‌ها از توئیتر مقداری هدفمند و باياس‌شده باشد. اما در گروه اکبرین شاید مسائل کمی اجتماعی‌تر و غیر سیاسی‌تر باشد.

اشتراک‌ها و اختلاف‌ها



کلمات مشترک و پر استفاده بین هر دو گروه

کلماتی که در گروه رسایی بیشتر دیده شده‌اند و در گروه اکبرین کم

کلماتی که در گروه اکبرین بیشتر دیده‌شده‌اند و در گروه رسایی کم

کلمه‌های مشترک:

ایران - روحانی - مجلس - آمریکا - ملت - آب - انقلاب - کار - پول - قانون، که خب به نسبت جالب است و نشان‌دهنده‌ی دغدغه‌ی مشترک است.

خدا، که می‌توان از شیوه‌ی و طرز بیان ناشی باشد.

کلمه‌های مشترک اما با سایز متفاوت:

روحانی در گروه رسایی بیشتر از هر چیزی مورد توجه است و در گروه مقابل خیر.

آمریکا در گروه رسایی بسیار بزرگ‌تر از گروه مقابل است.

دولت در گروه رسایی > گروه اکبرین

امام در گروه رسایی >> گروه اکبرین

انقلاب در گروه رسایی > گروه اکبرین

ایران و کار در گروه اکبرین > گروه رسایی

نظام در گروه اکبرین < گروه رسایی

انتخابات در گروه اکبرین > گروه رسایی

زندگی در گروه اکبرین > گروه رسایی

زن در گروه اکبرین > گروه رسایی

کلمه‌های متفاوت:

گروه رسایی:

- شهید
- رهبری
- الله
- برجام
- اسرائیل
- انقلابی
- ترامپ

گروه اکبرین:

- حکومت
- حال
- دانشگاه
- دختر
- جامعه
- کتاب