مسئله اول: دسته‌بندی متون

در پوشه *Classification* یک مجموعه داده مربوط به توییتر قرار دارد. در این مجموعه داده ویژگی‌های زیر موجودند:

- متن توییت‌ها
- ویژگی‌های تکمیلی مربوط به توییت مانند *is_retweet, is_quote* و غیره
- متغیر هدف با نام *relevant* که دو مقدار مرتبط(*Relevant*) و غیرمرتبط (*Irrelevant*) دارد

هدف از این پروژه ارائه و پیاده‌سازی مدلی برای دسته‌بندی توییت‌ها به دو دسته مرتبط و غیرمرتبط است. انتظار می‌رود که مراحل زیر برای انجام این کار مدنظر قرار گیرد:

- پیش‌پردازش متون
- تعریف ویژگی‌ها (مهندسی خصوصیات)
- پیاده‌سازی مدل دسته‌بندی با روش‌های شناخته شده مانند *RandomForest* و *XGBoost* و غیره
- با توجه به اینکه این مجموعه داده متوازن نیست تکنیک‌هایی مانند *Undersampling*، *Oversampling* و *SMOTE* روی مجموعه دادگان اعمال شده و نتایج مقایسه شوند.

خروجی مورد نظر گزارشی با رعایت اصول و استانداردهای نگارشی برای تدوین متون تخصصی است که شامل توضیح کامل مسئله، مجموعه داده، روش‌ها و تکنیک ها، نتایج و تحلیل نتایج می‌باشد.

مسئله دوم: شبکه ریتوییت

در پوشه *Retweet Network* تعدادی فایل با پسوند *json* قرار دارد که کلیه اطلاعات مربوط به توییت‌های یک واقعه خاص را جمع‌آوری کرده است. در این مجموعه داده به ازای هر توییت ویژگی‌های مورد نظر عبارتند از:

- شناسه منحصر به فرد کاربری که آن توییت را نوشته (*user_id*)
- اگر توییت مورد نظر ریتوییت باشد(*is_retweet: true*)، شناسه توییت‌کننده اصلی که توییتش ریتوییت شده (*retweeted_from_id*)
- نام کاربری که آن توییت را نوشته (*screen_name*)
- اگر توییت مورد نظر ریتوییت باشد(*is_retweet: true*)، نام توییت‌کننده اصلی که توییتش ریتوییت شده (*retweeted_from_sn*)

هدف از این پروژه ساخت شبکه ریتوییت برای این مجموعه دادگان و ارزیابی این شبکه است. برای این منظور ابتدا فایل‌های *json* را خوانده و تجمیع کرده و تبدیل به یک دیتافریم می‌کنیم. سپس با شناسایی ریتوییت‌ها و ویژگی‌های فوق، لیست مجاورت افراد را می‌سازیم. به عبارتی اگر یک کاربر توییت کاربر دیگر را ریتوییت کرده باشد نام این دو کاربر (*screen_name* ، *retweeted_from_sn*) در یک لیست مجاورت قرار می‌گیرد که در آن ستون اول نام کاربر اصلی و ستون دوم نام کاربری که مطلبش ریتوییت شده و ستون سوم وزن یا تعداد دفعات تکرار این اتفاق است.

این ماتریس مجاورت ورودی نرم‌افزار *Gephi* برای مصورسازی و تشخیص انجمن‌های شبکه است. با این لیست ابتدا شبکه را مصور کرده و سپس با ابزار *community detection* انجمن‌ها را شناسایی کنید. همچنین اطلاعات شبکه مانند تعداد رئوس و یال‌ها و معیارهای مرکزیت متداول مانند درجه رئوس، *betweenness*، *PageRank* و غیره را برای رئوس این شبکه محاسبه کنید و در قالب نمودار نمایش دهید.

خروجی مورد نظر گزارشی با رعایت اصول و استانداردهای نگارشی برای تدوین متون تخصصی است که شامل توضیح کامل مسئله، مجموعه داده، روش‌ها و تکنیک ها، نتایج و تحلیل نتایج می‌باشد.

مسئله سوم: مدلسازی موضوعی

در پوشه *Topic Modeling* فایلی مربوط به بخشی از توییت‌های فارسی منتشرشده مرتبط با موضوع کرونا قرار دارد. هدف از این پروژه این است که با پردازش این توییت‌ها و با استفاده از روش *Topic Modeling* این توییت‌هارا از نظر موضوعی دسته‌بندی کنید. برای این منظور ابتدا توییت‌ها را پیش‌پردازش کرده و بعد از توکن کردن، آن‌ها را به عنوان ورودی به مدل *LDA* که در پایتون برای دسته‌بندی موضوعی مورد استفاده قرار می‌گیرد بدهید. این کار را برای تعداد ۱۰ تاپیک انجام دهید. خروجی این مدل عبارتست از:

- موضوعی که به هر توییت نسبت داده شده
- لیست ۲۰ لغت پرتکرار هر موضوع

حال با بررسی لغات پرتکرار هر موضوع باید بتوانید نامی به آن تاپیک (که در حال حاضر اعداد ۰ تا ۹ هستند) نسبت دهید. سپس نمودار فراوانی هر تاپیک را بر اساس توییت‌ها رسم کنید. برای آشنایی بیشتر با روش کار و خروجی‌های مورد نظر به گزارش *http://metodata.ai/files/9910-Metodata-Report-Et-Al-AgendaSetting-E2.pdf* مراجعه کنید و بخش‌های مقدماتی آن را مطالعه کنید. همچنین از این گزارش به عنوان الگویی برای ساختار گزارش نویسی نیز استفاده کنید.

خروجی مورد نظر گزارشی با رعایت اصول و استانداردهای نگارشی برای تدوین متون تخصصی است که شامل توضیح کامل مسئله، مجموعه داده، روش‌ها و تکنیک ها، نتایج و تحلیل نتایج می‌باشد.

انتظار می‌رود شما بتوانید در طی انجام مراحل این سه پروژه موضوعات ناآشنا و ناشناخته را یاد بگیرید و دانش خود را افزایش دهید. مهلت انجام این پروژه‌ها تا تاریخ ۱ شهریور ۱۴۰۰ است و لازم به ذکر است که هیچ یک از این مسائل جزو پروژه‌های فعلی شرکت نبوده و بخش کوچکی از پروژه‌های تحقیقاتی و کاری و آموزشی است که در گذشته انجام شده و به پایان رسیده است.