به نام خدا

TOPIC MODELING

چکیده:

در این وضیفه توییت های تاریخ ۱۳۹۸/۱۱/۳ تا تاریخ ۱۳۹۸/۱۱/۱۳ که در ارتباط با موضوع کرونا بودند را پس از پیش‌پردازش های لازم بر روی متن توییت ها، با استفاده از مدل Latent Dirichlet Allocation در ۱۰ موضوع دسته بندی کردیم.

مجموع دادگان:

برای ورودی مدل(ویژگی های ورودی) از متن توییت‌های پردازش شده استفاده شد. در پیش پردازش داده ها ابتدا علائم نگارشی فارسی و انگلیسی، اعداد، لینک‌ها، تگ‌ها و کلمات انگلیسی از توییت‌ها حذف شدند همچنین حروف عربی با فارسی و کلمات جمع مکسر با مفرد متناظرشان جایگزین شدند؛ در مرحله بعد با استفاده از داده های یک فایل[1] کلمات اضافه (stop word) و کلماتی که از نظر معنایی در topic moedling تاثیر چندانی ندارند از متن پیام حذف شدند.

از این داده ها با استفاده از کتابخانه countVectorizer به بردار هایی تبدیل می‌کنیم همچنین از کلماتی که در کمتر از ۵ سند و یا بیشتر از ۹۰٪ استاد آمده باشند صرف نظر می‌کنیم. حد بالای تعداد کلمات را ۵۰۰۰ درنظر می‌گیریم.

در نهایت ماتریسی با ابعاد ۶۷۰۲۲ سطر و ۵۰۰۰ ستون حاصل می‌شود که ورودی مدل یادگیری ما است

[1] کلمات به کوشش آقای وحید خرازی تهیه شدند

مدل

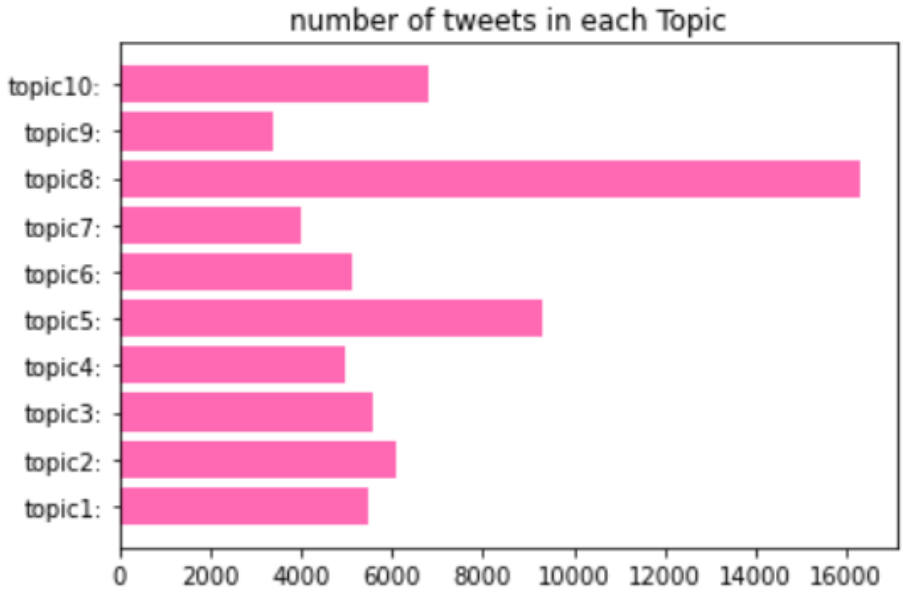
همانطور که گفته شد از مدل LDA استفاده شده است.به طور کلی از مدل LDA به عنوان روش استخراج موضوعات پنهان، کاهش اثر پراکندگی داده، و ساخت ویژگی برای آموزش بیشتر مدل های طبقه بندی داده‌های متنی استفاده می‌کنیم. در این مسئله نیز از این مدل استفاده شده و تعداد موضوعات را ۱۰ درنظر گرفتیم.

بعد از دادن داده های ورودی و آموزش مدل کلمات پرتکرار در هر موضوع به شرح زیر بدست آمد.

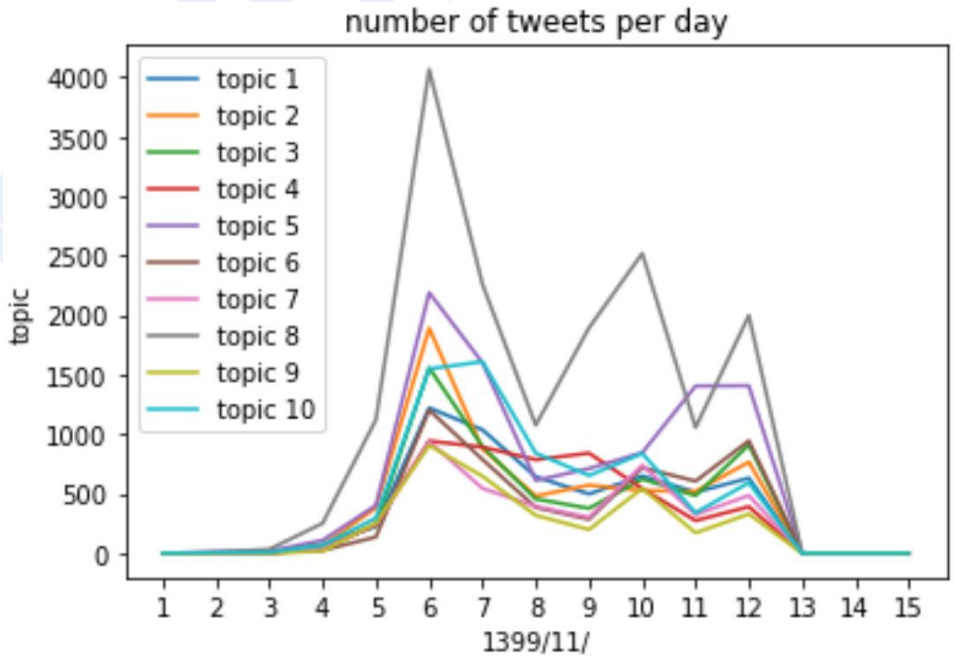
موضوع	لغات پرتکرار
کرونا در جهان	منتظر', 'پیشنهاد', 'اینه', 'تایلند', 'سقوط', 'گلوله', 'نگران', 'الله', 'ماه', 'تب', 'جهان', 'امریکا', 'ایران'
نگرانی	خدا', 'کاری', 'کتاب', 'انتقال', 'بخاطر', 'قرار', 'گرفت', 'ویروسی', 'نگرانی', 'مسیولین', 'میگن', 'ویروس'
حکومت	خواب', 'مملکت', 'اثر', 'ایرانی', 'ملت', 'زنده', 'اینا', 'جنگ', 'ترسید', 'طب', 'خطرناک', 'ترس', 'مرد', 'ویروس', 'جمهوری'
آموزش و پرورش	' ,استان', 'هاوشینگ', 'مقاطع', 'تحصیلی', 'استاندار', 'سفر', 'کار', 'کشتن', 'یاد', 'الوده', 'گسترش', 'مغزی', 'ضربه', 'کشته', 'ویروس'
تازه های علم	زمین', 'کنترل', 'تعداد', 'فیلم', 'گزارش', 'رژیم', 'جان', 'مسری', 'دانشمند', 'علم', 'کشور', 'درمان', 'سال', 'چین', 'شیوع', 'ویروس'
وضعیت اضطراری	رییس', 'عمومی', 'مرکز', 'انتشار', 'وزارت', 'سازمان', 'هشدار', 'وضعیت', 'سرباز', 'جهانی', 'هواپیما', 'دلیل', 'بهداشت', 'امریکایی', 'فوری', 'بیماری, ' 'ویروس'
شایعات	اربیل', 'راه', 'عسل', 'اشعه', 'حوزه', 'حیوان', 'بدن', 'موشک', 'اسم', 'منتشر', 'پاکستان', 'ابتلا', 'پیشگیری', 'تبریزی', 'دوست', 'انسان', 'تایید', 'ویروس'
پیدایش	خورد', 'وزیر', 'بهداشت', 'پزشکی', 'مرگ', 'ورود', 'خفاش', 'کشور', 'ووهان', 'بیمارستان', 'شهر', 'قرنطینه', 'ایران', 'نفر', 'چین', 'ویروس'
دین و کرونا	اسلام', 'خبری', 'شرکت', 'کشید', 'نامحرم', 'استارت', 'افراد', 'اتش', 'کره', 'میترسونین', 'فرد', 'کشنده', 'اخیرا', 'زد', 'ویروس', 'مبتلا'
تدبیرها	کشور', 'مرز', 'ویروس', 'داروی', 'بیمار', 'خبر', 'خودمون', 'بود', 'چینی', 'امام', 'کشف', 'دست', 'دنیا', 'ایران', 'وارد'

تحلیل نتایج حاصل

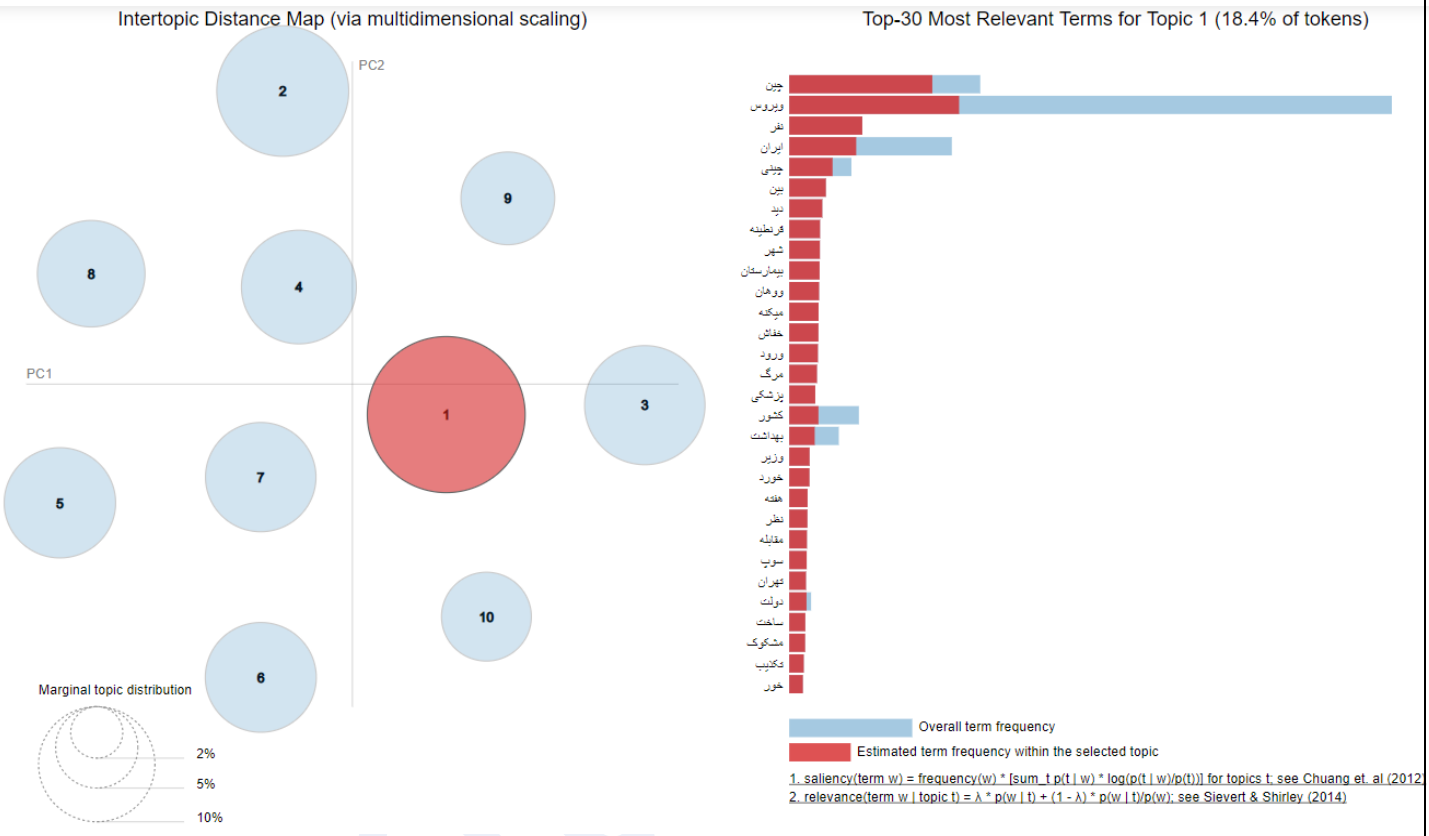
تعداد توییت‌های هر موضوع در نمودار زیر نمایش داده شده.



فراوانی تعداد توییت‌ها در هر موضوع در بازه زمانی ۱ بهمن تا ۱۵ بهمن هم در نمودار زیر نشان داده شده :



توضیع دو بعدی موضوعات همچنین توضیع توکن‌ها در هر موضوع در نتایج قابلم شاهده است[۲]. برای نمونه در موضوع ۶:



با بررسی نتایج مشخص می‌شود داده ها به خوبی تفکیک و دسته‌بندی شده‌اند و در فضا از یکدیگر فاصله گرفته‌اند. اما تعداد موضوعات قابل افزایش بود.

کد و مجموعه‌دادگان در آدرس زیر موجود می‌باشد.

Topic_Modeling

[2] مشاهده جزئیات و نتایج سایر موضوعات در لینک

METODATA