هفته هشتم:

- پروژه بارگذاری داده های بزرگ با حجم ۴ گیگ با کمک لاگستش در الستیک سرچ
- برای باگذاری داده های بزرگ بر روی الستیک سرچ مانند لاگهای سیتم ها باید از لاگستش استفاده کرد برای همین چون دادگان خبری بزرگ بود از لاگستش برای بارگذاری ان در الستیک سرچ استفاده شد.

- یادگیری کار با فیلتر های لاگستش
- چون دادگان خبری با فرمت جیسون تودرتو بود باید با استفاده از فیلتر های لاگستش ان را به یک جیسون یک سطحی تبدیل کرد تا بشود بر روی ان کویری زد

- تهیه گزارش کاراموزی

- پروژه نهایی:
- دادگان خبری را با استفاده از شبکه عصبی پردازش کنید و embed کنید و سپس در elasticsearch ایندکس کنید و یک موتور جستجو بسازید

- پیش پردازش ها و پردازش ها بر روی فیلد content که حاوی خود متن خبر است انجام شده است و برا اساس شباهت ان بازیابی میشود

- ابتدا برای پیش پردازش داده ها از کتابخانه hazm برای نرمال کردن داده های خبری استفاده شد و از یک لیست کلمات ایست برای حذف کلمات بی معنی با کمک کتابخانهnltk استفاده شد و برای تبدیل داده های جیسون چند خطی به یک خطی از regex استفاده شده است.

- سپس بر روی داده های پیش پردازش شده از fasttext استفاده شد تا بصورت یادگیری بدون ناظر مدل اموزش داده شود.

- پس از وزنهای مدل شبکه عصبی در یک فایل برای استفاده های بعدی مانند embedding ذخیره شد

- در نهایت از فایل اصلی داده و فایل داده های پیش پردازش شده و وزن های شبکه عصبی استفاده شد و داده های پیش پردازش شده به صورت بردارهایی با ابعاد ۳۰۰ همراه با متن و عنوان و سایر فیلدهای خبر در فایل دیگری به اسم embedded data ذخیره شد.

- سپس داده های embed شده با کمک لاگستش و فیلتر های مناسب در دیتابیس elasticsearch ایندکس شدند

- برای جستجو از موتور جستجو کویری کاربر ابتدا مانند قبل پیش پردازش میشود و بردار ان نیز با همان مدل و وزنها بدست میاید و سپس برای elasticsearch فرستاد میشود تا با استفاده از cosine similarity نردیک ترین داده ها برگردانده شود.

- نمونه ای از پرسجو و جواب:

- query = ریس جمهور ایران

1. به روحانی تبریک گفت

2. پیام تبریک رئیس جمهور ونزوئلا به روحانی
3. رهبر ارامنه جهان به روحانی شادباش گفت

- query = ریس جمهور امریکا

1. ترامپ: بزودی رئیس FBI را معرفی می کنم
2. هشدار جدی اروپا به ترامپ

- نمای پروژه نهایی

ایندکس کردن