بسمه تعالی

فاز اول پروژه بانک کتاب

Phoenix alliance

منتورشیپ:رضا باقریان

تهیه کنندگان:

عقیل سرمدی

فاطمه خجسته

پدرام مداح پور

امیرحسین شیخی

ارسلان جباری

هلیا اخترکاویان

کوئرا کالج تابستان ۱۴۰۲

فهرست

مقدمه و هدف:

ساخت یک بانک کتاب و استخراج و تحلیل های مهم از داده های آن و کاربردی کردن دانش کسب شده از استفاده از ابزار های دیتا آنالیز

شرح پروژه:

روزی روزگاری ارسلان که علاقه ی زیادی به برنامه نویسی داشت سر انتخاب برای شغل آینده با پدرش به مشکل خورد و پدرش بهش گفت که باید در شغل خانوادگی ما که چاپ کتاب هست فعالیت کنی. اما ارسلان مخالفت کرد و میخواست خودش رو به پدرش ثابت کنه و نشون بده که میتونه درامد چندین برابری داشته باشه با برقراری ارتباط با بزرگترین انتشارات!

و با این رزومه که تحلیل های ارزشمندی از اون انتشارات رو در کمترین زمان ممکن بهشون نشون بده که از طریق اون انتشارات بتونن درامدزایی خیلی بهتری داشته باشن!

Web Scraping:

۴ قسمت دارد گرفتن لینک هر کتاب یک هدر تعریف کرده ایم لیستی برای گرفتن لینک کتاب ها

در این بخش از پروژه داده ی حدود ۱۱۳۰۰۰ محصول از سایت ایران کتاب استخراج شده که ۱۰۰۸۰۰ تا آنها کتاب بوده است و طریقه ی استخراج این داده ها به صورت همزمان بوده است. به گونه ای که این حجم از دیتا که در حالت عادی ۱۰۰۰ لینک را در ۲۵ دقیقه استخراج می شود ،حال تنها در ۲.۵ دقیقه استخراج میشود! و در نهایت ۲۰ مگابایت داده از لینک کتاب ها استخراج شده است.

:Data Cleaning

با توجه به اینکه داده های زیادی از کتاب های مختلف ثبت نشده بود برای همسان سازی داده ها و حذف داده های اضافه جهت تحلیل درست تر داده ها موجود بخش زیادی از کار به تمیز کردن داده های موجود اختصاص داده شده است.

:DataBase

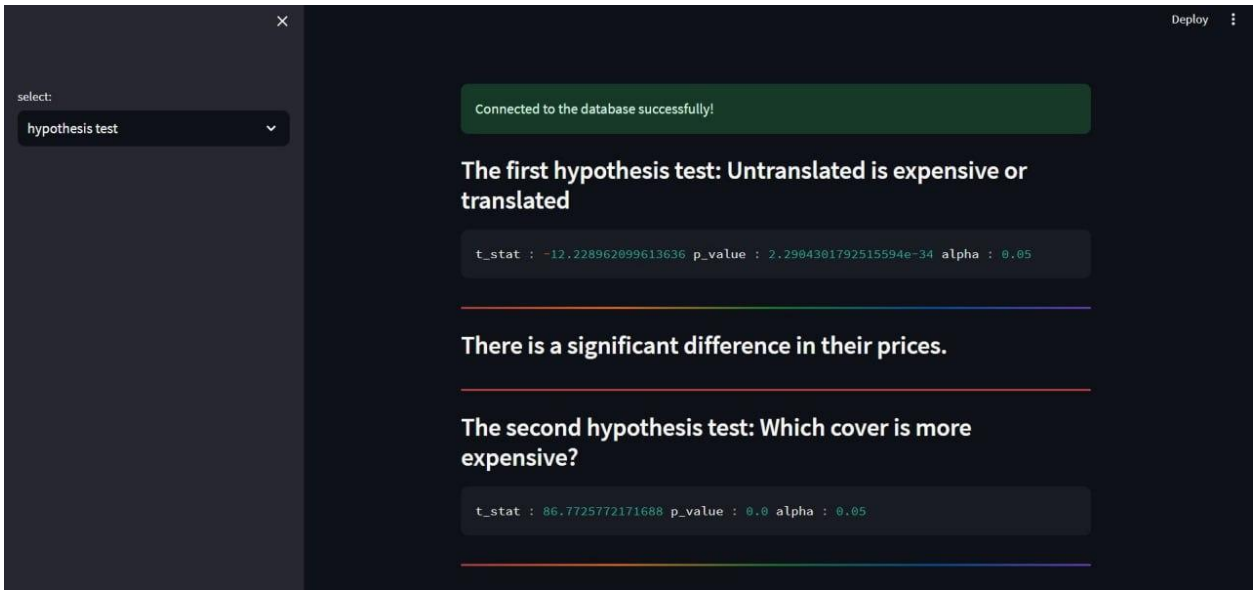
با استفاده از mysql و موجودیت به شکل کلاس پیاده سازی شده است

همچنین جدول erd تا سطح 1nf پیاده سازی شده است.جهت بهینه سازی جداول از جدول های رابطه استفاده شده است.و در روابط many to many از جدول های واسطه استفاده شده است و همچنین برای روابط one and one که اغلب میتوانند به صورت attribute باشند description ها را جدا نمودیم

قابلیت خاصی که در این بخش اضافه شده است به این صورت است که در تیبل current time بعد از هر بازه ی زمانی مشخص قیمت های کتاب ها را چک میکند و در صورت تغییر قیمت آن ها را آپدیت میکند که برای آن تریگر زده شده است و در erd نیز این قابلیت اضافه شده است. که از طرفی خود رصد کردن این تغییر قیمت ها میتواند مخزنی برای تحلیل دلایل تغییر قیمت کتاب ها از جمله شرایط چاپ و همچنین رصد علاقه مندی افراد به کتاب ها و از طرفی روانشناسی جامعه باشد.

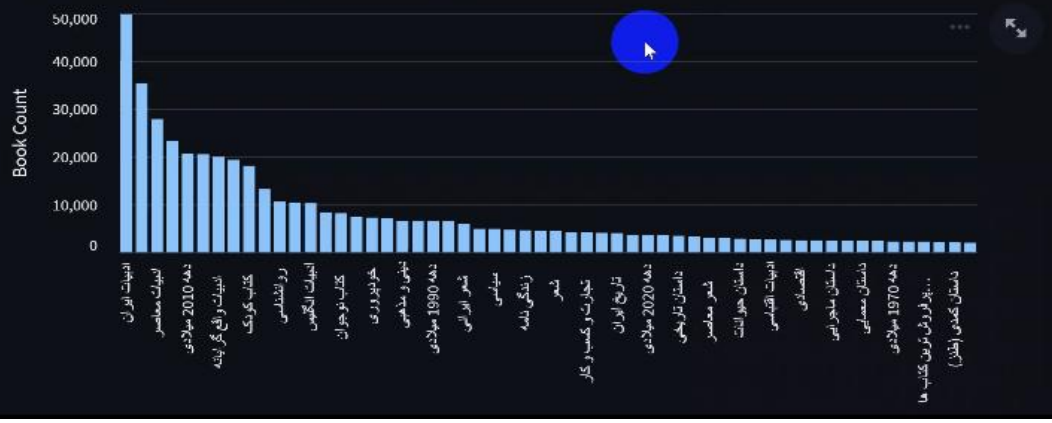
:Dashboard

تحلیل هایی نظیر ۱۰ نویسنده ی برتر یا ۱۰ مترجم برتر بر اساس تعداد کتاب،تعداد کتاب ها بر اساس هر تگ و میانگین لایک برای هر تگ را داریم. یا قطع محبوب یا تعداد کتاب های چاپ شده در هر سال و همچنین قیمت کتاب ها بر اساس سال یا صفحه ذکر شده است.



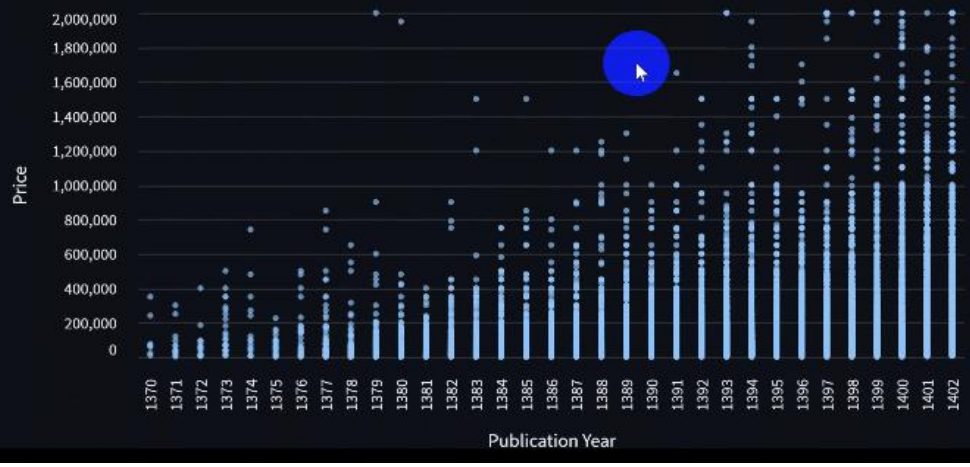
Analytical charts

1. Number of Books per Tag > 2000

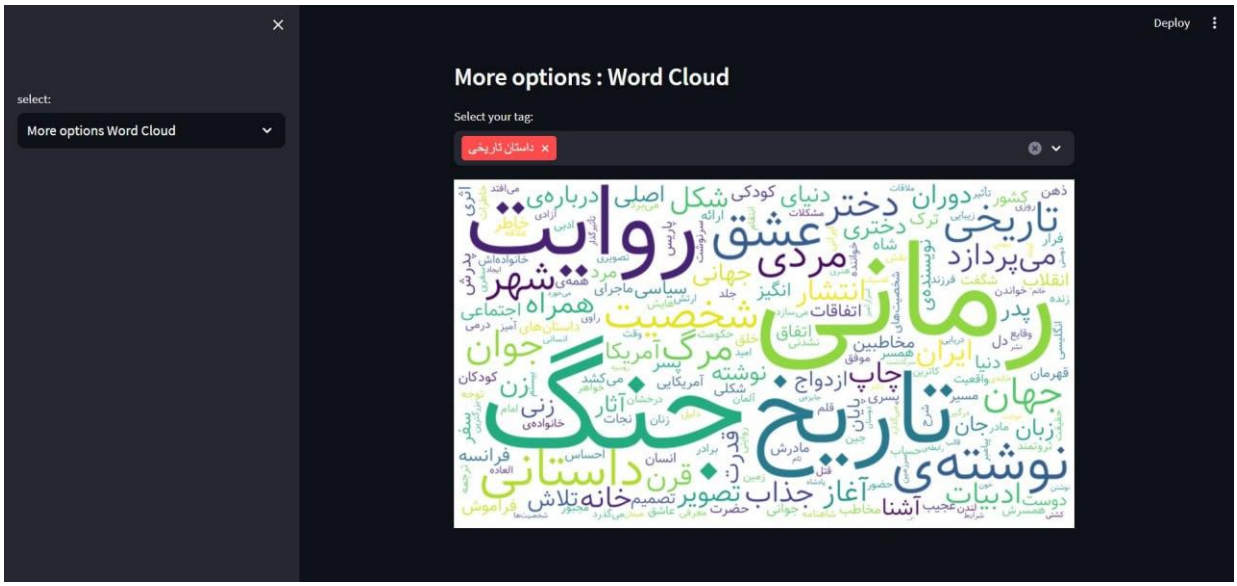


2. Top 10 Publishers by Book Count

7. Scatter Plot: Price vs. Publication Year



:More Anlaysic



با استفاده از روش های Machine Learning و الگوریتم apriori حساب میکنیم که هر تگ با چه درصدی با تگ های دیگر رابطه دارد.برای مثال اگر یک کتاب تاریخی باشد تا چه اندازه ای احتمالا تگ ادبیات معاصر را نیز خواهد داشت.

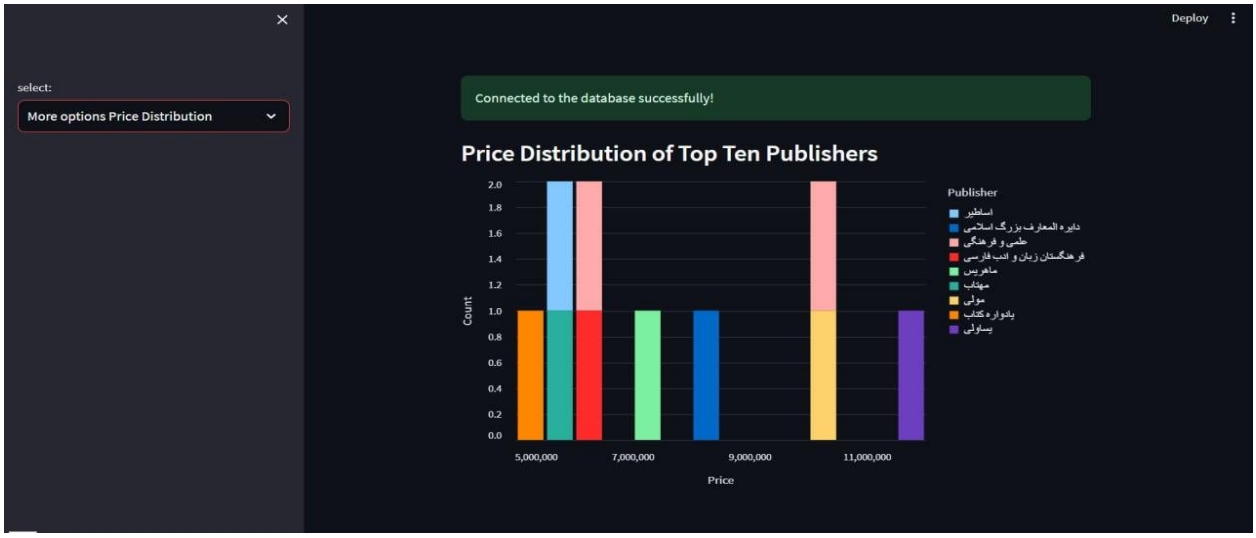
همچنین یک word cloud ای طراحی شده است که توضیحات مربوط به کتاب های مربوط به یک تگ خاص را دریافت کرده و بر اساس تعداد کلمات تکرار شونده عکس یک کلود را درست میکند

همچنین برای تحلیل بیشتر با توجه به درخواست نویسندگان برای پیدا کردن انتشارات مناسب،او این قابلیت را دارد که معیار مد نظر خود را پیدا کند مثل نشر محبوب یا چاپ بیشتر آن ناشر یا قیمت کتاب های مدنظر.

در راستای درخواست مشتریان فرض کنید مشتری ای داریم که میخواهد کتاب عاشقانه خریداری کند او با توجه به نظر کاربران(لایک آنها) یا بر اساس نویسندگانی که کتاب هایشان بیشتر از همه تجدید چاپ شده اند یا کتاب هایی که قیمت بیشتری دارند و او میتواند هر یک از این معیار ها را انتخاب کرده و کتاب مد نظر خود را پیدا کند.

همچنین ما با تست فرضیه این تحلیل را داشتیم که آیا ترجمه شدن کتاب تاثیر در قیمت آن دارد که با توجه به اینکه مقدار pvalue از آلفا کوچکتر است تفاوت معناداری بین قیمت کتاب ترجمه شده و کتاب عادی وجود داشت.

همچنین در تست فرضیه بعدی متوجه شدیم جلد سخت قیمت بسیار بیشتری از جلد نرم دارد و افراد حاضرند برای آن قیمت بیشتری پرداخت کنند.



User Interface:

قابلیتی که این پروژه را متمایز می سازد این است که جهت فهم بیشتر تحلیل ها گرافی از تگ ها طراحی شده است که در آن هر نود یک تگ بوده و هر یال ارتباط آن تگ ها در کتاب هاست و وزن یال هم تعداد تکرار آن ها است

منابع و مباحث:

https://github.com/realBagher/iran_ketab_analysis_pa/tree/main#project-overview