گزارش پروژه دوم

مقایسه تیتر اخبار صفحه اول روزنامه‌ها در دو دهه ۷۰ و ۸۰

زهرا صادقی عدل
۹۴۵۲۳۱۸۹

فاز اول

مراحل انجام پروژه:

الف)استخراج اطلاعات:

اطلاعات مربوط به سال‌های ۷۶تا ۸۰ از آرشیو روزنامه همشهری و اطلاعات مربوط به سال‌های ۸۰ تا ۸۵ از روزنامه ایران استخراج شده است.
در این فاز از داده‌های تیتر اول تمام روزهای روزنامه همشهری سال ۷۷ برای دهه ۷۰ و تمام روزهای روزنامه ایران سال ۸۵ برای دهه ۸۰ استفاده شده است.
کد مربوط به استخراج مجموعه داده در فایل data extraction به زبان پایتون موجود می باشد.این کد اطلاعات را استخراج و داده نامرتب و خام را در فایل data set ذخیره میکند.

ب)تمیزکردن دیتا:

برای تمیز کردن ابتدا تمام متن را از فونت عربی به فارسی تبدیل کردیم.سپس برای یکسان سازی نیم فاصله ها و \n را با فاصله جایگزین کردیم.در ابتدا یک لیست از لغات بدون معنی شامل افعال و حروف اضافه پرکاربرد را با تعداد محدود انتخاب و از مجموعه داده حذف نمودیم.همچینین کلمات را یکسان سازی کرده و ریشه هر کلمه را بجای ان قرار دادیم.
کد مربوط به تمیز کردن اطلاعات در فایل data cleaning موجود میباشد.این فایل اطلاعات خام را از dsts set دریافت کرده ان را تمیز و در همان فایل ذخیره میکند.

پ)شمارش و کارهای اماری:

در این مرحله کلمات به کمک هضم شناسایی و شمارش میشوند و کلمات و تعدادشان در فایلی با نام سال هر دیتاست ذخیره میگردند. سپس تعداد نسبی هر کلمه (تعداد تکرار هرکلمه تقسیم بر تعداد کل لغات) محاسبه شده و برای هر کلمه اختلاف در دو مجموعه داده محاسبه میگردد.این اختلاف در فایلهای 85-77و 77-85 ذخیره میگردند.

ت)رسم word map و مشاهده نتایج:

در این مرحله تنها کافیست فایلهای خروجی بدست امده در مراحل قبل را به تابع persianWordMap بدهیم تا با توجه به تواتر هر کلمه نتایج را ترسیم کند.
تنها نکته مورد توجه برای رسم word map ها هنگام کشیدن برای اختلاف نسبی میباشد. برای این کار برای هر کدام از فایلهای تفاضل کارهای زیر را انجام دادیم:
-انتخاب ۱۰۰ کلمه با اختلاف زیاد و رسم word map برای این ۱۰۰ کلمه
-انتخاب ۱۰۰ کلمه با اختلاف کم، تبدیل اختلاف به معکوس اختلاف(برای مشخص کردن کم بودن اختلاف) و رسم word map برای این ۱۰۰ کلمه

ج)نتایج:

سال ۱۳۸۵

سال ۱۳۷۷

-کلمات مصرف، سوخت، بنزین، طرح، مسکن، جهان، کاهش و افزایش، تصویب، هسته‌ای و ... در سال ۱۳۸۵ مهمترین اتفاقات این دهه را نمایش میدهد.

همچنین هاشمی، رفسنجانی، خاتمی، هیات، مطبوعات، جهانی، شورا، دادگاه و... بیانگر اتفاقات خاص این دهه می‌باشد.

بیشترین اختلافات:

کلماتی که در ۱۳۸۵ زیاد تکرار شده ولی در ۱۳۷۵ کم یا اصلا تکرار نشده

کلماتی که در ۱۳۷۷ زیاد تکرار شده ولی در ۱۳۸۵ کم یا اصلا تکرار نشده

-کلمات پرتکرار در سال ۱۳۸۵ و کم تکرار در سال ۱۳۷۷ بنزین، عراق، کنکور، احمدی نژاد، جنگ، اسرائیل، بهره‌برداری، لاریجانی، سوخت، مذاکران، فروشی و... میباشد که با توجه به شرایط این دهه قابل انتظار میباشد.

همچینین کلمات پرتکرار در سال ۱۳۷۷ و کم تکرار در سال ۱۳۸۵ خاتمی، هاشمی رفسنجانی، انتخابات،دیدار، خبرگان، همایش و... میباشد.

کمترین اختلافات:

کلماتی که در ۱۳۸۵ زیاد تکرار شده در ۱۳۷۵ هم بیشتر از بقیه تکرار شده

کلماتی که در ۱۳۷۷ زیاد تکرار شده ولی و ۱۳۸۵ هم بیشتر از بقیه تکرار شده

-کلمات مشترک مسکن رای، زنده باد، الودگی، جدید، قطعنامه، دلار، دیپلماتیک، فیلسوف، صهیونیستی، همکاری افغانستان، محیط و ... از کلمات مشترک دو دهه و درواقع دغدغه مشترک هر دو سال می‌باشند.