گزارش فاز اول

بنفشه كريميان

94521189

1. دیتاست

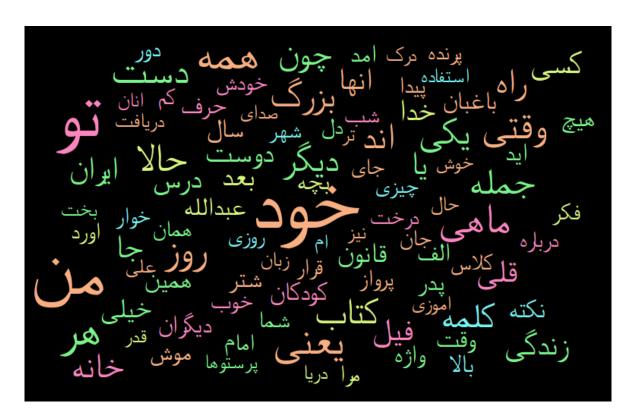
دیتاست انتخابی برای این پروژه کتاب ادبیات چهارم دبستان برای سال 1343 و 1391 است. برای گرفتن متن این دو کتاب به دلیل عکس بودن از ابزار googleDrive استفاده شده است.

2. فاز اول

برای فاز اول ابتدا نیم فاصله ها را با فاصله جایگزین کردیم. سپس یک لیست از نشانه ها مثل: ., و ... و لیست دیگری از حروف اضافه ی فارسی مثل: با, به, در و اعداد تشکیل داده و این لیست ها را با فاصله جایگزین کردیم. سپس کارکتر های عربی مثل هٔ, ی و اعراب ها و نشانه ی جمع را حذف کردیم. (در این بین تمام حروف انگلیسی نیز به دلیل احتمال وجود خطا در OCR حذف شدند) پس از این مراحل wordmap را رسم کردیم اما مشاهده شد که فعل هایی مثل است, بود و ... تواتر زیادی داشتند که برای جلوگیری از این مشکل فعل ها را نیز حذف کردیم. ابتدا دو واژهنمای زیر برای کلمات متواتر هر دیتاست به دست آمدند.



واژهنما کلمات متواتر پس از تمیز کردن دیتا در کتاب چهارم دبستان سال 1343

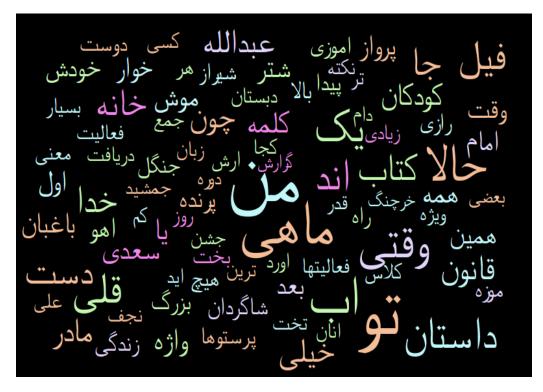


واژهنما 2 كلمات متواتر پس از تميز كردن ديتا در كتاب چهارم دبستان سال 1391

سپس تواتر هر کلمه را در آوردیم و دو واژهنمای زیر را رسم کردیم که اولی اختلاف تواتر کلمات کتاب سال 1391 با کلمات کتاب سال 1391 با کلمات کتاب سال 1391 با 1343 است. (برای کلماتی که در دیتاست دیگری وجود نداشتند تواتر خود کلمه منهای عدد صفر شدهاست)



واژهنما 3 كلمات كتاب سال 1343 با اختلاف تواتر آنها در كتاب سال 1391



و در آخر دو واژهنما با همان روش بالا ولی با این تفاوت که عدد بدست آمده را از بزرگترین عدد کم میکنیم تا کلماتی که اختلاف کمی دارند قابل مشاهدهباشند.

```
اسایش همچنین
                                           قضا کھن
                      يقين پيغمبر
                                             ساسانی مراجعه
                 امتح<sup>ان</sup> <sub>دا</sub>ستانی بغل
                                 خام وظيفه بنات زيارت
                              شىان
                                   اوستانش <sub>در</sub>هَم <sub>اه</sub>ل
             ایستاه ازمایشگوناگون شاخبلندش استان
شاعر البته
                                 نقاشی <sup>ها</sup> پیروز
                                                    مراقبت<sub>دچار</sub> سفیدی
یادگاردوران
           حادثهخرسندي نِزانات ورزشكار ۖ ثانيا اندرز ً سرتاسر
           ته رست ورو
دریچه <sup>جثه</sup> ثمره پارچه کتاباستفاده قرار
غم دریچه کتاباستفاده قرار
                         كرومورزشكاري پايتخت مشترك مذهب
                رزشدرت پید
پیرامون سیاحت زمه فحس
پرچم لگی پیشگیری مقایسه دکان نقل
ان سرگرم
                      سیاه <sup>تحی پی</sup>
سیاه تاریکی زمانی
<sub>ا</sub> منظور طریق
                                           لرزشها
                        درستكارباور
                                          شيرخواره
                          موافق
```

واژهنما 5 كلمات بااختلاف تواتر كم در كتاب سال 1343



واژهنما 6 كلمات بااختلاف تواتر كم در كتاب سال 1391