گزارش فاز اول

بنفشه کریمیان

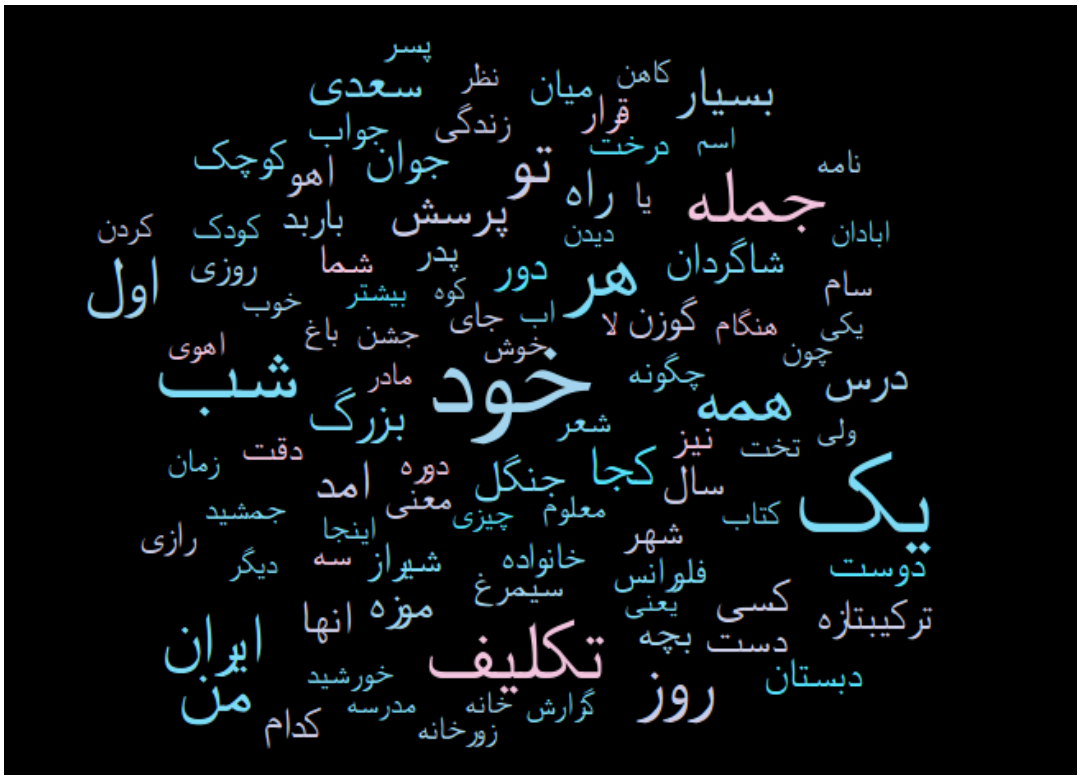
94521189

1. **دیتاست**

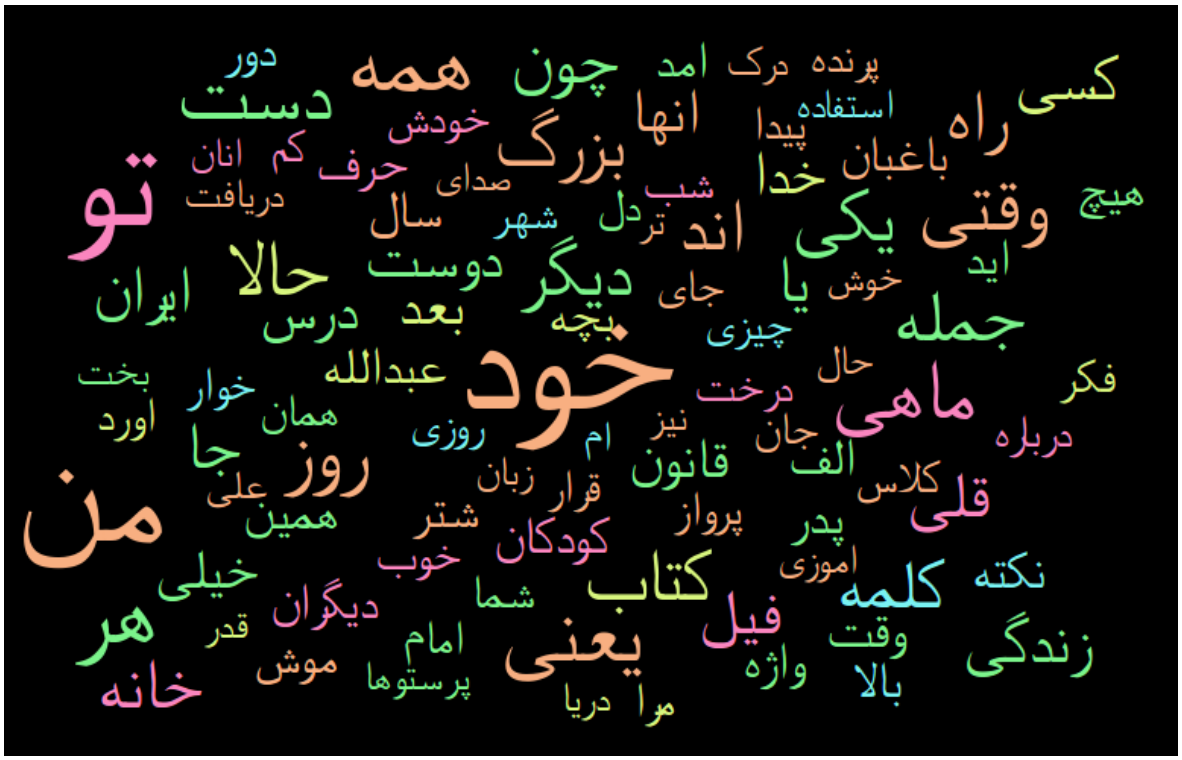
دیتاست انتخابی برای این پروژه کتاب ادبیات چهارم دبستان برای سال 1343 و 1391 است .
برای گرفتن متن این دو کتاب به دلیل عکس بودن از ابزار googleDrive استفاده‌شده‌است.

2. **فاز اول**

برای فاز اول ابتدا نیم فاصله‌ها را با فاصله جایگزین کردیم. سپس یک لیست از نشانه‌ها مثل: .,
: و ... و لیست دیگری از حروف اضافه‌ی فارسی مثل: با, به, در و اعداد تشکیل‌داده و این لیست‌ها را با فاصله جایگزین کردیم. سپس کارکترهای عربی مثل ۀ, ي و اعراب‌ها و نشانه‌ی 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 مشاهده‌شد که فعل‌هایی مثل است, بود و ... تواتر زیادی داشتند که برای جلوگیری از این مشکل فعل‌ها را نیز حذف کردیم.
ابتدا دو واژه‌نمای زیر برای کلمات متواتر هر دیتاست به‌دست‌آمدند.

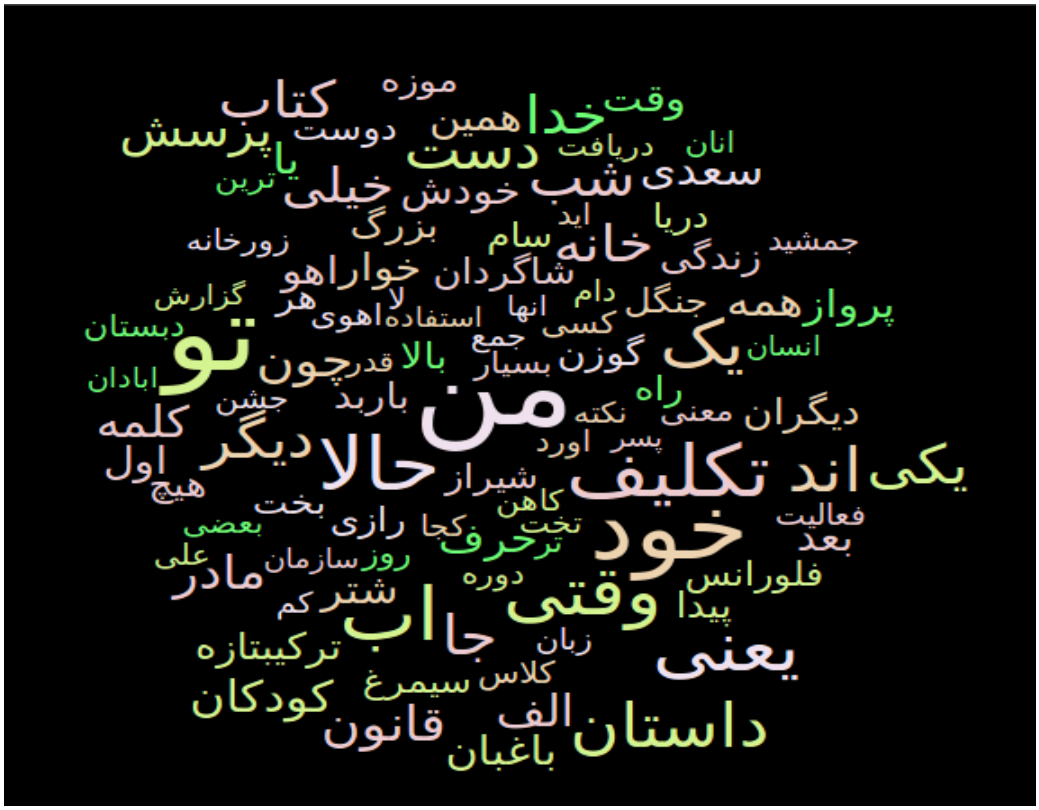


واژه‌نما1 کلمات متواتر پس از تمیز کردن دیتا در کتاب چهارم دبستان سال 1343

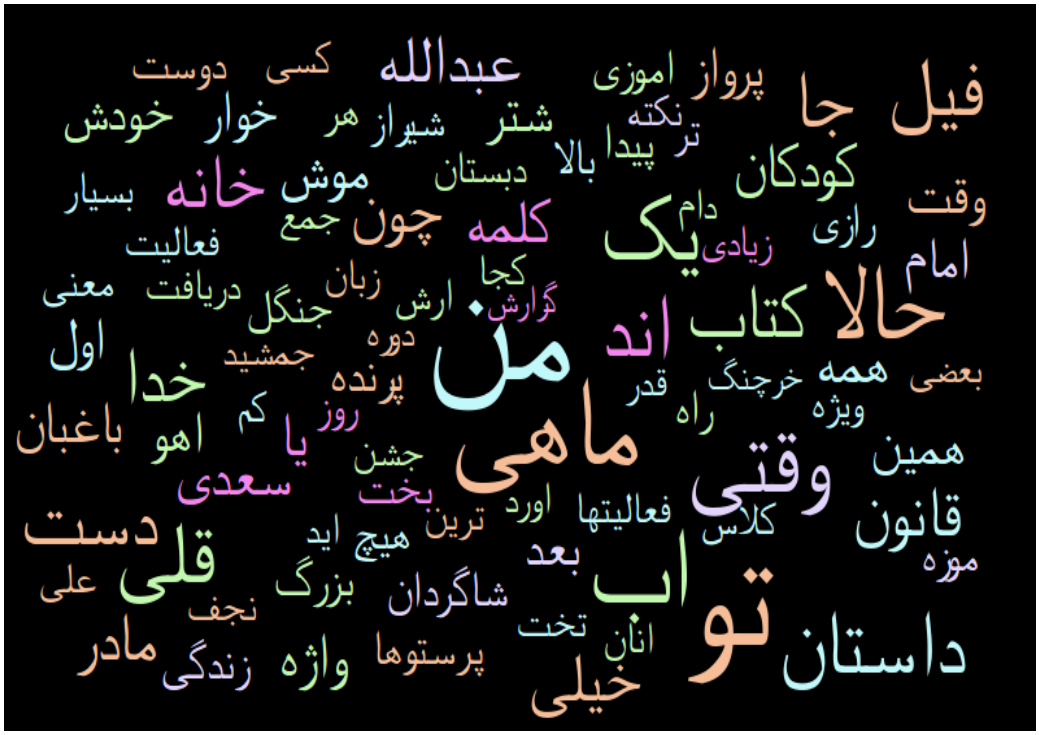


واژه‌نما 2 کلمات متواتر پس از تمیز کردن دیتا در کتاب چهارم دبستان سال 1391

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

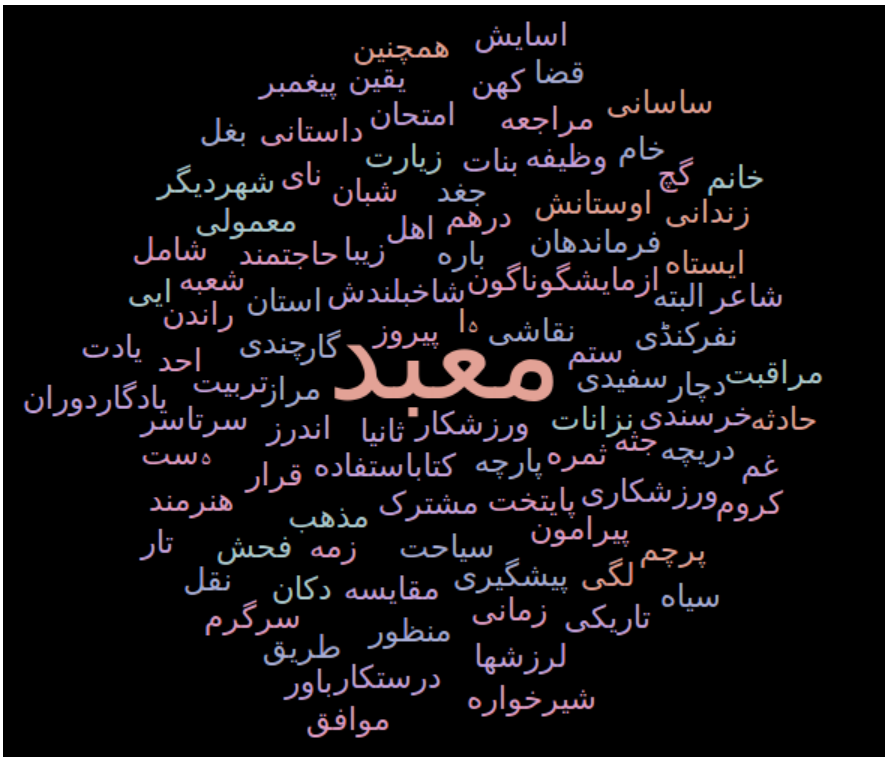


واژه‌نما 3 کلمات کتاب سال 1343 با اختلاف تواتر آن‌ها در کتاب سال 1391

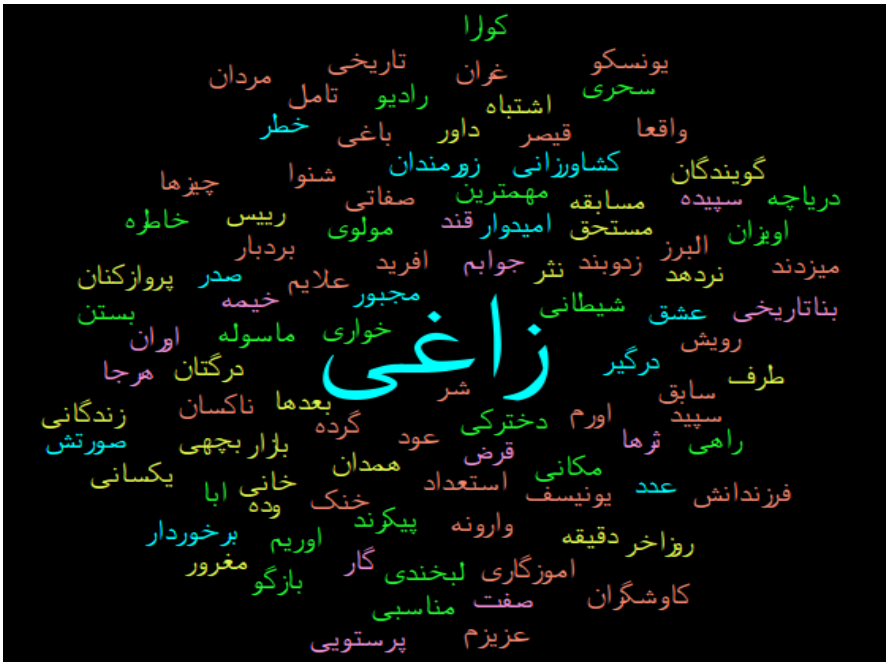


واژه‌نما 4کلمات کتاب سال 1391 با اختلاف تواتر آن‌ها در کتاب سال 1343

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.



واژه‌نما 5 کلمات بااختلاف تواتر کم در کتاب سال 1343



واژه‌نما 6 کلمات بااختلاف تواتر کم در کتاب سال 1391