محمدمهدی نظری – ۹۹۳۱۰۶۱

برای انجام بخش اول این فاز ۴ تابع تعریف کردم تا متن ورودی را به ترتیب نرمال کرده ( مواردی مانند درست کردن تاریخ ، درست کردن فاصله و نیم فاصله ، تبدیل اعداد فارسی به انگلیسی ،تبدیل نسخه فینگیلیش به فارسی و ... ) در مرحله بعدی توکن ها را در یک لیست جمع آوری میکند . در مرحله بعدی حروف یا کلمات پرتکرار مثل با – به - از و ... و علایِم نگارشی مثل نقطه ویرگول و... را از لیست توکن ها حذف میکند . مرحله آخر کلمات با ریشه یکسان را به یک کلمه تبدیل میکند مثلا صورت مختلف صرف افعال .

برای دو بخش اول از parsivar و بخش آخر از تابع های hazm استفاده شده است. برای حذف stopword ها از یک فایل تکست استفاده شده که عبارات موجود درآن را با توکن ها مقایسه میکنیم و اگر در لیستمان موجود بود حذف میکنیم.

نتیجه برای متن ورودی "من به تاریخ ۱۱ شهریور به دانشگاه می روم. man be tarikhe 11 shahrivar be danshgah miravam" به شکل زیر است :

```
Normal shode :  من به تاریخ 11 شهریور به دانشگاه می‌روم . منع به tarikhe 11 شهریور به دانشگاه miravam
Token haye be dast amade :  ['می', 'دانشگاه', 'به', 'شهریور', '11', 'تاریخ', 'به', 'من', 'منع\u200cبه', '.', 'روم', 'tarikhe', '11', 'دانشگاه', 'به', 'شهریور',
Hazf kalamate por tekrar :  ['می', 'دانشگاه', 'شهریور', '11', 'تاریخ', 'من', 'منع\u200cبه', 'روم', 'tarikhe', '11', 'دانشگاه', 'شهریور', 'miravam']
Hazf kalamate ba rishe yeksan :  ['می', 'دانشگاه', 'شهریور', '11', 'تاریخ', 'من', 'منع\u200cبه', 'رو', 'tarikhe', '11', 'دانشگاه', 'شهریور', 'miravam']
```

میبینیم که بصورت کلی به جواب خوبی میرسیم مثلا نیم فاصله در می روم را درست کرده - عدد فارسی را انگلیسی کرده – فینگلیش را تا حد خوبی درست برگردانده اما جاهایی که اعراب دارد مثل اعراب کسره در تاریخ را تشخیص نداده یا man را به جای من ، منع ترجمه کرده .

در بخش تهیه توکن ها هم مشکلاتی مثل مورد تکراری یا اشتباهی درنظر گرفتن کلمه مرکب مثلا دراینجا به من را یک کلمه حساب کرده (\u200c برای فاصله بین کلمات مرکب است که در اصل همان نیم فاصله است)

در بخش حذف stopword درست عمل کرده و "به" را حذف کرده در بخش ریشه یابی هم رفتن را به رو تبدیل کرده .

برای مشکلات ذکر شده دربخش های بعدی بصورت دستی حل میشود.