تمرین 1

7 آبان

مهدیه ساجدی‌پور

1. مقدمه

هدف این تمرین بررسی مراحل مختلف در پیش پردازش متن فارسی و همچنین آشنایی با ابزارهای مختلف پردازش زبان فارسی هست. سه ابزار مختلف بر روی متن داده شده استفاده می‌شوند تا تفاوت‌های آنها در نحوه ی پیش پردازش مقایسه شود.

2. توضیحات

پیش پردازش متون به سه بخش نرمال سازی، ریشه‌یابی و بن‌یابی، و توکن‌بندی تقسیم می‌شود.
ما در این تمرین، این مراحل را با ابزارهای هضم، دادما، و پارسیوار انجام می‌دهیم.
متن آزمایشی به این صورت است:
"

عنوان مقاله: صفحهٔ اصلی

متن ویکی‌پدیا فارسی

من john هستم و در ب.م.م گیری تخصص دارم!

كلمات عربي مانند اصلاح كاف و يا ي برای توکنایزر ما اَهمیت دارند.

ما می دانیم که در تاریخ ۲۰ سپتامبر ۲۰۰۴ (۲۹ شهریور،۱۳۸۳) مقاله های "ویکی پدیا" در ۱۰۵ زبان به یک میلیون رسید.
که این مقالات شامل زمـــان های پیشین نیستند.

در ویکی‌پدیا فارسی ممکن است (گاهی) فاصله پرانتز ها رعایت نشده باشد، یا حتی ممکن است درباره محبوب ترین های فارسی صحبت شده باشد.
در اینجا یک ایمیل آزمایشی از من sh@sbu.ac.ir قرار دارد.

برای اطلاعات بیشتر می‌توانید به وبسایت ویکی‌پیدا فارسی به آدرس http://wikipedia.com سر بزنید.
(داخل پرانتز بگویم، این یک متن تستی است. حداقل به من اینطور گفته شده است.)

مجلهٔ تایم در گزارش سال ۲۰۰۶ خود، جیمی ویلز را در گروه ۱۰۰ فرد تأثیرگذار سال اعلام کرد.

همچنین در همین سال ویکی پدیای روسی برندهٔ جایزهٔ رانِت (روسی: Премия Рунета) در بخش «دانش و آموزش» شد. این جایزه از طرف دولت اعطا می شود.

همچنین ویکی پدیا جایزهٔ یک میلیون دلاریِ مدیریت پروژه را از همایش صفاجو دریافت کرد.

پلتفورم اهداف توسعه پایدار United Nations:

چندین پروژهٔ متن-آزاد دارد که وظایف غیردانشنامه ای را انجام می دهند

"

a. نرمال سازی

نرمال سازی از مراحل ابتدایی در پردازش هر زبانی است. زبان فارسی دارای متونی سخت برای نرمال سازی است که دلایل مختلفی دارد. از جمله، فاصله‌ها ممکن است به طرق مختلفی نمایان شوند؛ به طور مثال کلمه درخت‌ها می‌تواند به چند صورت درختها، درخت‌ها و درخت ها نوشته شود. قانون ه کسره در بسیاری از متون رعایت نمیشود و ... . در نرمال‌سازی فاصله‌های بین کلمات، علائم نگارشی، تکرار در حروف، اختصارها و ... اصلاح می‌شوند تا به یک استاندارد مشترک بین پیش‌پردازش و پردازش که همان یونیکد است برسیم.

i. متن نرمال شده توسط هضم:

'عنوان مقاله: صفحه اصلی\n\n<p>متن ویکی\u200cپدیا فارسی< / p>\n\nمن john هستم و در ب. م. م گیری تخصص دارم! \n\nکلمات عربی مانند اصلاح کاف و یا\u200cی برای توکنایزر ما اهمیت دارند. \n\nما می\u200cدانیم که در تاریخ ۲۰ سپتامبر ۲۰۰۴ (۲۹ شهریور، ۱۳۸۳) مقاله\u200cهای «ویکی پدیا» در ۱۰۵ زبان به یک\u200cمیلیون رسید. \nکه این مقالات شامل زمان\u200cهای پیشین نیستند. \n\nدر ویکی\u200cپدیا فارسی ممکن است (گاهی) فاصله پرانتز\u200cها رعایت نشده باشد، یا حتی ممکن است درباره محبوب\u200cترین\u200cهای فارسی صحبت شده باشد. \nدر اینجا یک ایمیل آزمایشی از من sh@sbu. ac. ir قرار دارد. \n\nبرای اطلاعات بیشتر می\u200cتوانید به وبسایت ویکی\u200cپیدا فارسی به آدرس http: / / wikipedia. com سر بزنید. \n (داخل پرانتز بگویم، این یک متن تستی است. حداقل به من اینطور گفته\u200cشده است.) \n\nمجله تایم در گزارش سال ۲۰۰۶ خود، جیمی ویلز را در گروه ۱۰۰ فرد تأثیرگذار سال اعلام کرد. \n\nهمچنین در همین سال ویکی پدیای روسی برنده جایزه رانت (روسی: Премия Рунета) در بخش «دانش و آموزش» شد. این جایزه از طرف دولت اعطا می\u200cشود. \n\nهمچنین ویکی پدیا جایزه یک\u200cمیلیون دلاری مدیریت پروژه را از همایش صفاجو دریافت کرد. \n\nپلتفورم اهداف توسعه پایدار United Nations: \n\nچندین پروژه متن-آزاد دارد که وظایف غیردانشنامه\u200cای را انجام می\u200cدهند'

ii. متن نرمال شده توسط پارسیوار

'عنوان مقاله : صفحهٔ اصلی < p > متن ویکی\u200cپدیا فارسی < /p > من john هستم و در ب . م . م\u200cگیری تخصص دارم ! کلمات عربی مانند اصلاح کاف و یا ی برای توکنایزر ما اهمیت دارند . ما می\u200c\u200cدانیم که در تاریخ 20 سپتامبر 2004 (1383 ، y0m6d29) مقاله\u200cهای " ویکی پدیا " در 105 زبان به 1000000 رسید . که این مقالات شامل زمان\u200cهای پیشین نیستند . در ویکی\u200cپدیا فارسی ممکن است (گاهی) فاصله پرانتز\u200cها رعایت\u200cنشده\u200cباشد ، یا حتی ممکن است درباره محبوب ترین\u200cهای فارسی صحبت\u200cشده\u200cباشد . در اینجا 1 ایمیل آزمایشی از من . sh@sbu . ir . ac قرار دارد . برای اطلاعات بیشتر می\u200cتوانید به وبسایت ویکی\u200cپیدا فارسی به آدرس http :// wikipedia . com سر بزنید . (داخل پرانتز بگویم ، این 1 متن تستی است . حداقل به من اینطور گفته\u200cشده\u200cاست .) مجلهٔ تایم در گزارش سال y0m0d2006 خود ، جیمی ویلز را در گروه 100 فرد تاثیرگذار سال اعلام کرد . همچنین در همین سال ویکی پدیای روسی برندهٔ جایزهٔ رانت (روسی : Премия Рунета) در بخش « دانش و آموزش » شد . این جایزه از طرف دولت اعطا می\u200cشود . همچنین ویکی پدیا جایزهٔ 1000000 دلاری مدیریت پروژه را از همایش صفاجو دریافت کرد . پلتفورم اهداف توسعه پایدار United Nations : چندین پروژهٔ متن - آزاد دارد که وظایف غیردانشنامه\u200cای را انجام می\u200c\u200cدهند'

iii. متن نرمال شده توسط دادما

'عنوان مقاله: صفحهٔ اصلی متن ویکی\u200cپدیا فارسی من john هستم و در ب. م. م گیری تخصص دارم! کلمات عربی مانند اصلاح کاف و یا ی برای توکنایزر ما آهمیت دارند. ما می \u200cدانیم که در تاریخ <NUM> سپتامبر <NUM>) <NUM> شهریور، <NUM>) مقاله های "ویکی پدیا" در <NUM> زبان به یک میلیون رسید. که این مقالات شامل زمـــان های پیشین نیستند. در ویکی\u200cپدیا فارسی ممکن است (گاهی) فاصله پرانتز ها رعایت نشده باشد، یا حتی ممکن است درباره محبوب ترین های فارسی صحبت شده باشد. در اینجا یک ایمیل آزمایشی از من <EMAIL> قرار دارد. برای اطلاعات بیشتر می\u200cتوانید به وبسایت ویکی\u200cپیدا فارسی به آدرس <URL> سر بزنید. (داخل پرانتز بگویم، این یک متن تستی است. حداقل به من اینطور گفته شده است.) مجلهٔ تایم در گزارش سال <NUM> خود، جیمی ویلز را در گروه <NUM> فرد تاثیرگذار سال اعلام کرد. همچنین در همین سال ویکی پدیای روسی برندهٔ جایزهٔ رانِت (روسی: Премия Рунета) در بخش «دانش و آموزش» شد. این جایزه از طرف دولت اعطا می شود. همچنین ویکی پدیا جایزهٔ یک میلیون دلاری مدیریت پروژه را از همایش صفاجو دریافت کرد. پلتفورم اهداف توسعه پایدار United Nations: چندین پروژهٔ متن-آزاد دارد که وظایف غیردانشنامه ای را انجام می دهند'

iv. مقایسه:

نرمالایزر هضم بر خلاف دو ابزار دیگر، \n را با next line جابه‌جا می‌کند، در صورتی که آن دو next line را با white space جایگزین می‌کنند.

نرمالایزر پارسیوار <p> را به سه کلمه تبدیل میکند اما دیگر ابزار ها آن را یک کلمه درنظر میگیرند.
هیچ کدام کلمات مختصر شده را تشخیص نمی‌دهند.
تنها پارسیوار به قبل علائم نگارشی فاصله می‌افزاید.
دادما از دیگر ابزارها در تشخیص نیم‌فاصله‌ها و تصحیح‌شان ضعیف‌تر عمل می‌کند.
هر سه در تبدیل ي به ی درست عمل می‌کنند اما دادما اعراب را حذف نمی‌کند.
دادما در مرحله نرمالایز کردن قابلیت تبدیل اعداد و url و ادرس ایمیل را به token های مشخص شده دارد.
پارسیوار میتواند تاریخ شمسی را نرمالایز کند.
دادما بر خلاف دو ابزار دیگر قابلیت حذف کاراکتر ـــ در زمـــان را ندارد.
در پارسیوار فاصله ی پرانتز ها تصحیح نمی‌شود(استاندارد این کتابخانه با بقیه متفاوت است).
تنها دادما میتواند ایمیل و url را تشخیص دهد، بقیه قبل و بعد از . فاصله می‌گذارند.
دادما افعال چندبخشی را تشخیص نمی‌دهد، هضم تنها افعال را به صورت دو بخشی تشخیص می‌دهد، اما پارسیوار میتواند افعال سه بخشی را نیز تشخیص دهد.
هضم برخلاف دو ابزار دیگر، أ را به ا تبدیل نمی‌کند.
تنها هضم ء را حذف می‌کند.
پارسیوار ترکیب متن-آزاد را به سه توکن نبدیل می‌کند.
به نظر می‌رسد پارسیوار متن نرمال شده‌ی بهتری را تولید می‌کند.

b. توکن‌بندی

ساده‌ترین راه برای توکن‌بندی استفاده از فاصله به عنوان جدا کننده است. توکن‌بندی به ما کمک می‌کند بتوانیم از ابزارهای ریشه‌یابی و بنیابی استفاده کنیم و نیز روابط بین کلمات در هر جمله را به دست بیاوریم. همچنین همچنین هر توکن به عنوان یک لغت در لغت‌نامه مدل ذخیره می‌شود. بهتر در توکن‌بندی از ابزاری استفاده شود که کلمات چندبخشی را تشخیص داده و به عنوان یک توکن در نظر بگیرد.

i. توکن‌بندی با هضم

```
['عنوان', 'مقاله', ':', 'صفحه', 'اصلی', '<p>متن',
'ویکی\u200cپدیا', 'فارسی<', '/', 'p>', 'من', 'john', 'هستم',
'و', 'در', 'ب', '.', '4', 'NUM', 'م', 'گیری', 'تخصص',
'دارم', '!', 'کلمات', 'عربی', 'مانند', 'اصلاح', 'کاف', 'و',
'یا\u200cی', 'برای', 'توکنایزر', 'ما', 'اهمیت', 'دارند', '.',
'ما', 'می\u200cدانیم', 'که', 'در', 'تاریخ', '2', 'NUM',
'سپتامبر', '2', 'NUM', '(', '4', 'NUM', 'شهریور', '،', 'NUM',
```

```
4'', ')', 'مقاله\u200cهای', '«', 'ویکی', 'پدیا', '»', 'در',
'3', 'NUM', 'زبان', 'به', 'یک\u200cمیلیون', 'رسید', '.', 'که',
'این', 'مقالات', 'شامل', 'زمان\u200cهای', 'پیشین', 'نیستند',
'.', 'در', 'ویکی\u200cپدیا', 'فارسی', 'ممکن', 'است', '(',
'گاهی', ')', 'فاصله', 'پرانتز\u200cها', 'رعایت', 'نشده_باشد',
'،', 'یا', 'حتی', 'ممکن', 'است', 'درباره',
'محبوب\u200cترین\u200cهای', 'فارسی', 'صحبت', 'شده_باشد',
'.', 'در', 'اینجا', 'یک', 'ایمیل', 'آزمایشی', 'از', 'من',
'ir', 'ac', '.', 'sh@sbu', '.', 'قرار', 'دارد', '.', 'برای',
'اطلاعات', 'بیشتر', 'می\u200cتوانید', 'به', 'وبسایت',
'ویکی\u200cپیدا', 'فارسی', 'به', 'آدرس', '/', '/', ':', 'http',
'com', '.', 'wikipedia', '/', '', 'سر', 'بزنید', '.', '(',
'داخل', 'پرانتز', 'بگویم', '،', 'این', 'یک', 'متن', 'تستی',
'است', '.', 'حداقل', 'به', 'من', 'اینطور', 'گفته\u200cشده',
'است', '.', ')', 'مجله', 'تایم', 'در', 'گزارش', 'سال',
'4', 'NUM', 'خود', '،', 'جیمی', 'ویلز', 'را', 'در', 'گروه',
'3', 'NUM', 'فرد', 'تأثیرگذار', 'سال', 'اعلام', 'کرد', '.',
'همچنین', 'در', 'همین', 'سال', 'ویکی', 'پدیای', 'روسی',
'برنده', 'جایزه', 'رانت', '(', 'روسی', ':', 'Премия',
'Рунета'', ')', 'در', 'بخش', '«', 'دانش', 'و', 'آموزش',
'»', 'شد', '.', 'این', 'جایزه', 'از', 'طرف', 'دولت', 'اعطا',
'می\u200cشود', '.', 'همچنین', 'ویکی', 'پدیا', 'جایزه',
'یک\u200cمیلیون', 'دلاری', 'مدیریت', 'پروژه', 'را', 'از',
'همایش', 'صفاجو', 'دریافت', 'کرد', '.', 'پلتفورم', 'اهداف',
'توسعه', 'پایدار', 'United', 'Nations', ':', 'چندین', 'پروژه',
'متن-آزاد', 'دارد', 'که', 'وظایف', 'غیردانشنامه\u200cای', 'را',
'انجام', 'می\u200cدهند']
```

ii. توکن‌بندی با پارسیوار

```
['عنوان', 'مقاله', ':', 'صفحهٔ', 'اصلی', '>', 'p', '<', 'متن',
'ویکی\u200cپدیا', 'فارسی', '/>', 'p', '<', 'من', 'john',
'هستم', 'و', 'در', 'ب', '.', 'م', '.', 'م\u200cگیری',
'تخصص', 'دارم', '!', 'کلمات', 'عربی', 'مانند', 'اصلاح', 'کاف',
'و', 'یا', 'ی', 'برای', 'توکنایزر', 'ما', 'اهمیت', 'دارند', '.',
'ما', 'می\u200c\u200cدانیم', 'که', 'در', 'تاریخ', '20',
'سپتامبر', '2004', '(', '1383', '،', 'y0m6d29', ')',
'مقاله\u200cهای', '"', 'ویکی', 'پدیا', '"', 'در', '105',
'زبان', 'به', '1000000', 'رسید', '.', 'که', 'این', 'مقالات',
'شامل', 'زمان\u200cهای', 'پیشین', 'نیستند', '.', 'در',
'ویکی\u200cپدیا', 'فارسی', 'ممکن', 'است', '(', 'گاهی', ')',
'فاصله', 'پرانتز\u200cها', 'رعایت\u200cنشده\u200cباشد', '،',
'یا', 'حتی', 'ممکن', 'است', 'درباره', 'محبوب',
'ترین\u200cهای', 'فارسی', 'صحبت\u200cشده\u200cباشد', '.',
'در', 'اینجا', '1', 'ایمیل', 'آزمایشی', 'از', 'من', 'sh@sbu',
'ir', 'ac', '.', '', 'قرار', 'دارد', '.', 'برای', 'اطلاعات',
'بیشتر', 'می\u200cتوانید', 'به', 'وبسایت', 'ویکی\u200cپیدا',
'فارسی', 'به', 'آدرس', '.', 'wikipedia', '://', 'http',
'com'', 'سر', 'بزنید', '.', '(', 'داخل', 'پرانتز', 'بگویم', '،',
```

```
'این', '1', 'متن', 'تستی', 'است', '.', 'حداقل', 'به', 'من',
'اینطور', 'گفته\u200cشده\u200cاست', '.)', 'مجلهٔ', 'تایم', 'در',
'گزارش', 'سال', 'y0m0d2006', 'خود', '،', 'جیمی', 'ویلز',
'را', 'در', 'گروه', '100', 'فرد', 'تاثیرگذار', 'سال', 'اعلام',
'کرد', '.', 'همچنین', 'در', 'همین', 'سال', 'ویکی', 'پدیای',
'روسی', 'برندهٔ', 'جایزهٔ', 'رانت', '(', 'روسی', ':', 'Премия
Рунета'', ')', 'در', 'بخش', '«', 'دانش', 'و', 'آموزش',
'»', 'شد', '.', 'این', 'جایزه', 'از', 'طرف', 'دولت', 'اعطا',
'می\u200cشود', '.', 'همچنین', 'ویکی', 'پدیا', 'جایزهٔ',
'1000000', 'دلاری', 'مدیریت', 'پروژه', 'را', 'از', 'همایش',
'صفاجو', 'دریافت', 'کرد', '.', 'پلتفورم', 'اهداف', 'توسعه',
'پایدار', 'United', 'Nations', ':', 'چندین', 'پروژهٔ', 'متن',
'-', 'آزاد', 'دارد', 'که', 'وظایف', 'غیردانشنامه\u200cای', 'را',
'انجام', 'می\u200cدهند']
```

iii. مقایسه

* امکان استفاده از توکنایزر دادما وجود نداشت.

همانطور که در مرحله نرمال‌سازی پارسیوار از هضم بهتر عمل کرده‌است، در این مرحله نیز بهتر عمل می‌کند.

در هضم به دلیل فاصله نگذاشتن قبل از علائم نگارشی، برخی کلمه‌ها همراه با علائم به عنوان یک توکن در نظر گرفته‌شده‌اند.

در هضم، در مرحله ی توکن‌بندی تبدیل ایمیل و اعداد به توکن‌های مخصوص را داریم، که با توجه به فاصله‌گذاری‌های نامناسب در مرحله‌ی قبل، قادر به تشخیص ایمیل و url نمی‌باشد.

هر دو توکنایزر کلماتی را که با نیم‌فاصله به هم متصل شده‌اند را یک توکن در نظر می‌گیرند، که باتوجه به قوی‌تر بودن پارسیوار در این موضوع در مرحله قبل، در این مرحله این پارسیوار متن را به صورت درست‌تری توکنایز کرده‌است.

c. ریشه‌یابی و بن‌یابی

ریشه‌یابی به معنای حذف بخش‌های افزوده شده به کلمه و تبدیل آن به حالت ساده می‌باشد دایره لغات مدل حالت منسجم تری داشته باشد. این عمل برای زبان انگلیسی دارای الگوریتمی به نام porter می‌باشد، اما در کتابخانه‌های فارسی به صورت کاملا ابتدایی تعریف شده است. خروجی ریشه‌یابی گاها کلمات بامعنی نیستند؛ به این دلیل که حذف پسوند تنها با یک شرط انجام می‌شود و در این عملیات کلمه‌ای مثل عالی نیز تحت تاثیر ریشه‌یاب به عال تبدیل می‌شود.(مثال: کتاب‌هایشان -> کتاب)

در فرآیند بن‌یابی بن‌یاب به دنبال یک کلمه با معنی می‌باشد. تفاوت ریشه‌یاب و بن‌یاب در همین مسئله است که کلمات تولید شده بعد از stem کردن ممکن است بامعنی نباشند اما بعد از lemmatize کردن همیشه کلمات با معنی داریم. Lemmatizer های پیشرفته بر اساس morphological parsing پیاده‌سازی شده‌اند، اما کتابخانه های فارسی براساس لیستی از کلمات مرجع به همراه ریشه آن این کار را انجام می‌دهد.

i. ریشه‌یابی با هضم

['عنو', 'مقاله', ':', 'صفحه', 'اصل', '<p>متن',
'ویکی\u200cپدیا', 'فارسی<', '/', 'p>', '<p', 'من', 'john', 'هس',
'و', 'در', 'ب', '.', '4', 'NUM', '', 'گیر', 'تخصص', 'دار',
'!', 'کل', 'عرب', 'مانند', 'اصلاح', 'کاف', 'و', 'یا', 'برا',
'توکنایزر', 'ما', 'اهم', 'دارند', '.', 'ما', 'می\u200cدان', 'که',
'در', 'تاریخ', '2', 'NUM', 'سپتامبر', 'NUM', '(', 'NUM', '4', 'NUM',
2'', 'شهریور', '،', '4', 'NUM', ')', 'مقاله', '«', 'ویک',
'پدیا', '»', 'در', '3', 'NUM', 'زب', 'به', 'یک\u200cمیلیون',
'رسید', '.', 'که', 'این', 'مقال', 'شامل', 'زمان', 'پیشین',
'نیستند', '.', 'در', 'ویکی\u200cپدیا', 'فارس', 'ممکن', 'اس',
')', 'گاه', '(', 'فاصله', 'پرانتز', 'رعا', 'نشده_باشد', '،', 'یا',
'حت', 'ممکن', 'اس', 'درباره', 'محبوب\u200cترین', 'فارس',
'صحب', 'شده_باشد', '.', 'در', 'اینجا', 'یک', 'ایمیل', 'آزمایش',
'از', 'من', 'ir', '.', 'ac', '.', 'sh@sbu', 'قرار', 'دارد',
'.', 'برا', 'اطلاع', 'ب', 'می\u200cتوانید', 'به', 'وبسا',
'ویکی\u200cپیدا', 'فارس', 'به', 'آدرس', '/', '/', 'http', ':',
com', '.', 'wikipedia', '/', '', 'سر', 'بزنید', '.', '(',
'داخل', 'پرانتز', 'بگو', '،', 'این', 'یک', 'متن', 'تست', 'اس',
'.', 'حداقل', 'به', 'من', 'اینطور', 'گفته\u200cشده', 'اس',
'.', ')', 'مجله', 'تا', 'در', 'گزار', 'سال', '4', 'NUM',
'خود', '،', 'جیم', 'ویلز', 'را', 'در', 'گروه', '3', 'NUM',
'فرد', 'تأثیرگذار', 'سال', 'اعلا', 'کرد', '.', 'همچنین', 'در',
'همین', 'سال', 'ویک', 'پدیا', 'روس', 'برنده', 'جایزه', 'ران',
'(', 'روس', ':', 'Рунета', 'Премия', ')', 'در', 'بخ',
'«', 'دان', 'و', 'آموز', '»', 'شد', '.', 'این', 'جایزه', 'از',
'طرف', 'دول', 'اعطا', 'می\u200cشود', '.', 'همچنین', 'ویک',
'پدیا', 'جایزه', 'یک\u200cمیلیون', 'دلار', 'مدیر', 'پروژه', 'را',
'از', 'هما', 'صفاجو', 'دریاف', 'کرد', '.', 'پلتفور', 'اهداف',
'توسعه', 'پایدار', 'Nations', 'United', ':', 'چندین', 'پروژه',
'متن-آزاد', 'دارد', 'که', 'وظایف', 'غیردانشنامه', 'را', 'انجا',
'می\u200cدهند']

ii. بن‌یابی با هضم

['عنو', 'مقاله', ':', 'صفحه', 'اصل', '<p>متن',
'ویکی\u200cپدیا', 'فارسی<', '/', 'p>', '<p', 'من', 'john', 'هس',
'و', 'در', 'ب', '.', '4', 'NUM', '#هست', 'گیر', 'تخصص',
'دار', '!', 'کل', 'عرب', 'مانند', 'اصلاح', 'کاف', 'و', 'یا',
'برا', 'توکنایزر', 'ما', 'اهم', 'داشت#دار', '.', 'ما',
'می\u200cدان', 'که', 'در', 'تاریخ', '2', 'NUM', 'سپتامبر',
'2', 'NUM', '(', 'NUM', '4', 'شهریور', '،', '4', 'NUM',
')', 'مقاله', '«', 'ویک', 'پدیا', '»', 'در', '3', 'NUM',
'زب', 'به', 'یک\u200cمیلیون', 'رسید', '.', 'که', 'این', 'مقال',
'شامل', 'زمان', 'پیشین', 'نیستند', '.', 'در', 'ویکی\u200cپدیا',
'فارس', 'ممکن', 'اس', '(', 'گاه', ')', 'فاصله', 'پرانتز',
'رعا', 'شد#شو', '،', 'یا', 'حت', 'ممکن', 'اس', 'درباره',

```
'محبوب', 'فارس', 'صحب', 'شد#شو', '.', 'در', 'اینجا', 'یک',
'ایمیل', 'آزمایش', 'از', 'من', '.', 'ac', '.', 'sbu', '.', 'sh@
ir'', 'قرار', 'داشت#دار', '.', 'برا', 'اطلاع', 'ب',
'توانست#توان', 'به', 'وبسا', 'ویکی\u200cپیدا', 'فارسی', 'به',
'آدرس', 'com', '.', 'wikipedia', '.', '/', '/', ':', 'http',
'سر', 'زد#زن', '.', '(', 'داخل', 'پرانتز', 'بگو', '،', 'این',
'یک', 'متن', 'تست', 'اس', '.', 'حداقل', 'به', 'من',
'اینطور', 'گفته\u200cشده', 'اس', '.', ')', 'مجله', 'تا', 'در',
'گزار', 'سال', '4', 'NUM', 'خود', '،', 'جیم', 'ویلز', 'را',
'در', 'گروه', '3', 'NUM', 'فرد', 'تأثیرگذار', 'سال', 'اعلا',
'کرد#کن', '.', 'همچنین', 'در', 'همین', 'سال', 'ویک', 'پدیا',
'روس', 'برنده', 'جایزه', 'ران', '(', 'روس', ':', 'Премия
Рунета'', ')', 'در', 'بخ', '«', 'دان', 'و', 'آموز', '»',
'شد#شو', '.', 'این', 'جایزه', 'از', 'طرف', 'دول', 'اعطا',
'شد#شو', '.', 'همچنین', 'ویک', 'پدیا', 'جایزه',
'یک\u200cمیلیون', 'دلار', 'مدیر', 'پروژه', 'را', 'از', 'هما',
'صفاجو', 'دریاف', 'کرد#کن', '.', 'پلتفور', 'اهداف', 'توسعه',
'پایدار', 'United', 'Nations', ':', 'چندین', 'پروژه',
'متن-آزاد', 'داشت#دار', 'که', 'وظایف', 'غیردانشنامه', 'را', 'انجا',
'داد#ده']
```

iii. ریشه و بن‌یابی با پارسیوار

```
['عنوان', 'مقاله', ':', 'صفحهٔ', 'اصلی', '>', 'p', '<', 'متن',
'ویکی\u200cپدیا', 'فارسی', '/>', 'p', '<', 'من', 'john',
'هست', 'و', 'در', 'ب', '.', 'م', '.', 'م\u200cگیری',
'تخصص', 'داشت&دار', '!', 'کلمات', 'عربی', 'مانند', 'اصلاح',
'کاف', 'و', 'یا', 'ی', 'برای', 'توکنایزر', 'ما', 'اهمیت',
'داشت&دار', '.', 'ما', 'دانست&دان', 'که', 'در', 'تاریخ', '20',
'سپتامبر', '2004', '(', '1383', '،', 'y0m6d29', ')', 'مقاله',
'"', 'ویکی', 'پدیا', '"', 'در', '105', 'زبان', 'به',
'1000000', 'رسید', '.', 'که', 'این', 'مقالات', 'شامل', 'زمان',
'پیشین', 'نیستند', '.', 'در', 'ویکی\u200cپدیا', 'فارسی', 'ممکن',
'اس', '(', 'گاهی', ')', 'فاصله', 'پرانتز',
'رعایت\u200cنشده\u200cباشد', '،', 'یا', 'حتی', 'ممکن', 'اس',
'درباره', 'محبوب', 'ترین\u200cهای', 'فارسی',
'صحبت\u200cشده\u200cباشد', '.', 'در', 'اینجا', '1', 'ایمیل',
'آزمایشی', 'از', 'من', 'ir', 'ac', '.', 'sbu', '.', 'sh@
',
'قرار', 'داشت&دارد', '.', 'برای', 'اطلاعات', 'بیشتر',
'توانست&توان', 'به', 'وبسایت', 'ویکی\u200cپیدا', 'فارسی', 'به',
'آدرس', 'com', '.', 'wikipedia', '.', '://', 'http', 'سر',
'زد&زن', '.', '(', 'داخل', 'پرانتز', 'گفت&گو', '،', 'این', '1',
'متن', 'تست', 'اس', '.', 'حداقل', 'به', 'من', 'اینطور',
'گفته\u200cشده\u200cاست', '.', ')', 'مجلهٔ', 'تایم', 'در', 'گزارش',
'سال', 'y0m0d2006', 'خود', '،', 'جیمی', 'ویلز', 'را', 'در',
'گروه', '100', 'فرد', 'تاثیرگذار', 'سال', 'اعلام', 'کرد', '.',
'همچنین', 'در', 'همین', 'سال', 'ویکی', 'پدیای', 'روسی',
'برندهٔ', 'جایزهٔ', 'ران', '(', 'روسی', ':', 'Премия
Рунета'', ')', 'در', 'بخش', '«', 'دانش', 'و', 'آموزش',
```

```
'»', 'شد', '.', 'این', 'جایزه', 'از', 'طرف', 'دولت', 'اعطا',
'شد&شو', '.', 'همچنین', 'ویکی', 'پدیا', 'جایزهٔ', '1000000',
'دلاری', 'مدیریت', 'پروژه', 'را', 'از', 'همایش', 'صفاجو',
'دریافت', 'کرد', '.', 'پلتفورم', 'اهداف', 'توسعه', 'پایدار',
'United', 'Nations', ':', 'چندین', 'پروژهٔ', 'متن', '-',
'آزاد', 'داشت&دارد', 'که', 'وظایف', 'غیردانشنامه\u200cای', 'را',
'انجام', 'داد&ده']
```

iv. مقایسه

* پارسیوار دارای بن‌یاب و ریشه‌یاب جداگانه نیست و این کار را به صورت ترکیبی در `FindStems` انجام می‌دهد.

در هضم stemmer اگر کلمات با ["ات", "ان", "ترین", "تر", "م", "ت", "ش", "یی", "ی", "ها", "ٔ", "ا", ""] تمام شده‌باشند، آن بخش هارا حذف می‌کند و ۀ را به ه تبدیل می‌کند. به همین دلیل دارای خطای بسیار است. به طور مثال کلمه ی اعلام را به اعلا تبدیل می‌کند.

Lemmatizer هضم ابتدا کلمه را با لیستی ازکلمات ذخیره شده مقایسه می‌کند، در صورت مچ شدن همین کلمه را برمی‌گرداند. در غیر این صورت، در فایلی از افعال به دنبال کلمه می‌گردد تا ریشه را بازگرداند. سپس یک بار دیگر کلمه را stem می‌کند و با لغات ذخیره شده چک می‌کند. در نهایت یا stem شده را برمی‌گرداند یا خود کلمه را.

Stemmizer پارسیوار خیلی ضعیف عمل می‌کند. به خصوص برای کلماتی که دارای نیم‌فاصله می‌باشند.

در مراحل مختلف پسوند و پیشوندهای افعال و دیگر کلمات را حذف می‌کند.

تمرکز آن بیشتر روی افعال می‌باشد و توانایی ریشه یابی کلمات عادی آن بسیار پایین است؛ به طور مثال نمی‌تواند غیردانشنامه‌ای را ساده سازی کند.

همچنین در بن‌یابی فعل ساده‌ای مثل هستند شکست می‌خورد.

3. نتیجه‌گیری

پکیج دادما دارای باگ‌های بسیاری بود و کارکردن با آن بسیار سخت!

پارسیوار و هضم دارای تفاوت‌هایی در مراحل مختلف بودند که باید در هنگام انتخاب به آنها توجه کرد و با توجه به نوع متن و همچنین تسک مورد نظر یک کدام را برگزید. به طور کلی پارسیوار در مرحله توکنایز کردن بهتر عمل کرد اما هضم برای ریشه‌یابی و بن‌یابی بهتر عمل می‌کند.

4. منابع

https://github.com/roshan-research/hazm/tree/master
https://github.com/Dadmatech/DadmaTools/tree/main

https://github.com/ICTRC/Parsivar/tree/master
https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf