تمرین اول:

1- نرمال‌سازی:

الف) نرمال‌سازی به گونه های مختلفی انجام می‌شود. یکی از ساده ترین انواع آن تبدیل همه حروف به بزرگ یا کوچک که در انگلیسی رایج است. بعد از آن میتوان به پاک کردن علائم نگارشی اشاره کرد(.؟!:"....). کار دیگری که عموما برای نرمال‌سازی متن انجام می‌شود پاکسازی حروف توقف یا ایست واژه است و این کلمات هیچ ارتباطی با مفهوم کلی محتوا ندارند و حذف آنها معنای متن را تغییر نمی‌دهد (در اکثر اوقات). کار دیگری که می‌توان انجام داد پاک کردن عدد ها و علامت های خاص مثل ($#^*@) است.

ب) از سه ابزار هضم و پارسیوار و دادما استفاده کردیم

متن اولیه:

عنوان مقاله: صفحهٔ اصلی

متن ویکی‌پدیا فارسی

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

کلمات عربي مانند اصلاح کاف و يا ي برای توکنایزر ما أهمیت 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.

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

همچنین ویکی پدیا جایزهٔ یک میلیون دلاري 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

متن نرمال شده توسط هضم:

عنوان مقاله: صفحه اصلی

<p>متن ویکی‌پدیا فارسی< / p>

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

کلمات عربی مانند اصلاح کاف و یای برای توکنایزر ما اهمیت دارند.

ما می‌دانیم که در تاریخ ۲۰ سپتامبر ۲۰۰۴ (۲۹ شهریور، ۱۳۸۳) مقاله‌های «ویکی پدیا» در ۱۰۵ زبان به یک‌میلیون رسید.
که این مقالات شامل زمان‌های پیشین نیستند.

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

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

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

در بخش «دانش و آموزش» شد. این جایزه از طرف دولت اعطا می‌شود (روسی: Премия Рунета) همچنین در همین سال ویکی پدیای روسی برنده جایزه رانت.

همچنین ویکی پدیا جایزه یک‌میلیون دلاری مدیریت پروژه را از همایش صفاجو دریافت کرد.

United Nations: پلتفورم اهداف توسعه پایدار

چندین پروژه متن-آزاد دارد که وظایف غیردانشنامه‌ای را انجام می‌دهند

همانطور که دیده میشود تغییر خاصی ایجاد نشده و ضعیف عمل کرده.

## متن نرمال شده توسط پارسیوار:

عنوان مقاله : صفحهٔ اصلی
< p > متن ویکی‌پدیا فارسی </ p >
من john هستم و در ب . م . مگیری تخصص دارم !
کلمات عربی مانند اصلاح کاف و یا ی برای توکنایزر ما اهمیت دارند .
ما می‌دانیم که در تاریخ 20 سپتامبر 2004 ( 29 شهریور ، 1383 ) مقاله‌های " ویکی پدیا " در 105 زبان به یک میلیون رسید .
که این مقالات شامل زمان‌های پیشین نیستند .
در ویکی‌پدیا فارسی ممکن است ( گاهی ) فاصله پرانتزها رعایت‌نشده‌باشد ، یا حتی ممکن است درباره محبوب ترین‌های فارسی صحبت‌شده‌باشد .
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 گفته‌شده‌است .)
مجلهٔ تایم در گزارش سال 2006 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 می‌شود .
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 غیردانشنامه‌ای را انجام می‌دهند

در اینجا تغییرات قابل مشاهده تر است و خط های خالی حذف شده اند.

## حال به دادما نگاهی می‌اندازیم:

عنوان مقاله صفحهٔ اصلی متن ویکی‌پدیا فارسی من john هستم و در ب م م گیری تخصص دارم کلمات عربی مانند اصلاح کاف و یا ی برای توکنایزر ما اَهمیت دارند ما می دانیم که در تاریخ 20 سپتامبر 2004 29 شهریور 1383 مقاله های ویکی پدیا در 105 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 <EMAIL> قرار دارد برای اطلاعات بیشتر می‌توانید به وبسایت ویکی‌پیدا فارسی به آدرس سر بزنید داخل پرانتز بگویم این یک متن تستی است حداقل به من اینطور گفته شده است مجلهٔ تایم در گزارش سال 2006 خود جیمی ویلز را در گروه 100 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 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

همانطور که مشاهده میشود همه خط ها به هم چسبیده و موارد اضافی تا حد خوبی پاک شده و نرمالسازی بسیار بهتر از دو ابزار دیگر انجام شده.

2- ریشه‌یابی و بن‌یابی:

الف) در ریشه یاب به تک واژه ها نگاه میشود وبه اطراف آن کاری نداریم و وند های تصریفی و اشتقاقی از کلمه حذف میشوند. ولی در بن یابی بر اساس نوع کلمه و بستری که در آن به کار رفته است عمل را انجام میدهیم و فقط وند های تصریفی ممکن است حذف شوند. پس بن یاب کلمه را به شکل پایه معنی دار آن تبدیل میکند ولی در ریشه یابی عموما صرفا چند حرف آخر یک کلمه حذف میشود.

ب) ابتدا در هضم بررسی میکنیم:

['عنو', 'مقاله:', 'صفحه', 'اصل', '<p>متن', 'ویکی\u200cپدیا', 'فارسی<', '/', 'p>', 'من', 'john', 'هس', 'و', 'در', 'ب.', 'م.', '"', 'گیر', 'تخصص', 'دارم!', 'کل', 'عرب', 'مانند', 'اصلاح', 'کاف', 'و', 'یا', 'برا', 'توکنایزر', 'ما', 'اهم', 'دارند.', 'ما', 'می\u200cدان', 'که', 'در', 'تاریخ', '۲۰', 'سپتامبر', '۲۰۰۴', '(۲۹', 'شهریور،', '۱۳۸۳)', 'مقاله', '«ویک', 'پدیا»', 'در', '۱۰۵', 'زب', 'به', 'یک\u200cمیلیون', 'رسید.', 'که', 'این', 'مقال', 'شامل', 'زمان', 'پیشین', 'نیستند.', 'در', 'ویکی\u200cپدیا', 'فارس', 'ممکن', 'اس', '(گاهی)', 'فاصله', 'پرانتز', 'رعا', 'نشده', 'باشد،', 'یا', 'حت', 'ممکن', 'اس', 'درباره', 'محبوب\u200cترین', 'فارس', 'صحب', 'شده', 'باشد.', 'در', 'اینجا', 'یک', 'ایمیل', 'آزمایش', 'از', 'من', 'sh@sbu.', 'ac.', 'ir', 'قرار', 'دارد.', 'برا', 'اطلاع', 'ب', 'می\u200cتوانید', 'به', 'وبسا', 'ویکی\u200cپیدا', 'فارس', 'به', 'آدرس', 'com', 'wikipedia.', '/', '/', 'http:', 'سر', 'بزنید.', '(داخل', 'پرانتز', 'بگویم،', 'این', 'یک', 'متن', 'تست', 'است.', 'حداقل', 'به', 'من', 'اینطور', 'گفته\u200cشده', 'است.)', 'مجله', 'تا', 'در', 'گزار', 'سال', '۲۰۰۶', 'خود،', 'جیم', 'ویلز', 'را', 'در', 'گروه', '۱۰۰', 'فرد', 'تأثیرگذار', 'سال', 'اعلا', 'کرد.', 'همچنین', 'در', 'همین', 'سال', 'ویک', 'پدیا', 'روس', 'برنده', 'جایزه', 'ران', '(روسی:', 'Премия', 'Рунета)', 'در', 'بخ', '«دان', 'و', 'آموزش»', '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', 'United', 'Nations', 'چندین', 'پروژه', 'متن-آزاد', 'دارد', 'که', 'وظایف', 'غیردانشنامه', 'را', 'انجا', 'می\u200cدهند']

متن نرمال شده به این گونه ریشه یابس میشود و همانگونه که گفته شد بعضی کلمه ها معانی خود را از دست داده اند زیرا بخشی از آخر آنها حذف شده.

['عنو', 'مقاله:', 'صفحه', 'اصل', '<p>متن', 'ویکی\u200cپدیا', 'فارسی<', '/', 'p>', 'من', 'john', 'هس', 'و', 'در', 'ب.', 'م.', '#هست', 'گیر', 'تخصص', 'دارم!', 'کل', 'عرب', 'مانند', 'اصلاح', 'کاف', 'و', 'یا', 'برا', 'توکنایزر', 'ما', 'اهم', 'دارند.', 'ما', 'می\u200cدان', 'که', 'در', 'تاریخ', '۲۰', 'سپتامبر', '۲۰۰۴', '(۲۹', 'شهریور،', '۱۳۸۳)', 'مقاله', '«ویک', 'پدیا»', 'در', '۱۰۵', 'زب', 'به', 'یک\u200cمیلیون', 'رسید.', 'که', 'این', 'مقال', 'شامل', 'زمان', 'پیشین', 'نیستند.', 'در', 'ویکی\u200cپدیا', 'فارس', 'ممکن', 'اس', '(گاهی)', 'فاصله', 'پرانتز', 'رعا', 'شد#شو', 'باشد،', 'یا', 'حت', 'ممکن', 'اس', 'درباره', 'محبوب', 'فارس', 'صحب', 'شده', 'باشد.', 'در', 'اینجا', 'یک', 'ایمیل', 'آزمایش', 'از', 'من', 'sh@sbu.', 'ac.', 'ir', 'قرار', 'دارد.', 'برا', 'اطلاع', 'ب', 'توانست#توان', 'به', 'وبسا', 'ویکی\u200cپیدا', 'فارس', 'به', 'آدرس', 'http:', 'com', 'wikipedia.', '/', '/', 'سر', 'بزنید.', '(داخل', 'پرانتز', 'بگویم،', 'این', 'یک', 'متن', 'تست', 'است.', 'حداقل', 'به', 'من', 'اینطور', 'گفته\u200cشده', 'است.)', 'مجله', 'تا', 'در', 'گزار', 'سال', '۲۰۰۶', 'خود،', 'جیم', 'ویلز', 'را', 'در', 'گروه', '۱۰۰', 'فرد', 'تأثیرگذار', 'سال', 'اعلا', 'کرد.', 'همچنین', 'در', 'همین', 'سال', 'ویک', 'پدیا', 'روس', 'برنده', 'جایزه', 'ران', '(روسی:', 'Премия', 'Рунета)', 'در', 'بخ', '«دان', 'و', 'آموزش»', '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', 'United','Nations':', 'چندین', 'پروژه', 'متن-آزاد', 'داشت#دار', 'که', 'وظایف', 'غیردانشنامه', 'را', 'انجا', 'داد#ده']

همانطور که میبینیم به دلیل خوب نرمال نشدن متن بن واژه ها به خوبی یافت نشده و فقط تعداد محدودی از کلمات بن یابی شده اند.

حال پارسیوار را برسی میکنیم:

['عنوان', 'مقاله', ':', 'صفحهٔ', 'اصلی', '<', 'p', '>', 'متن', 'ویکی\u200cپدیا', 'فارسی', '</', 'p', '>', 'من', 'john', 'هست', 'و', 'در', 'ب', '.', 'م\u200cگیری', 'تخصص', 'داشت&دار', '!', 'کلمات', 'عربی', 'مانند', 'اصلاح', 'کاف', 'و', 'یا', 'ی', 'برای', 'توکنایزر', 'ما', 'اهمیت', 'م', '.', 'داشت&دار', '.', 'ما', 'دانست&دان', 'که', 'در', 'تاریخ', '20', 'سپتامبر', '2004', '(', '29', 'شهریور', '،', '1383', ')', 'مقاله', '""', 'ویکی', 'پدیا', '""', 'در', '105', 'زبان', 'به', 'یک', 'میلیون', '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', 'یک', 'ایمیل', 'آزمایشی', 'از', 'من', 'sh@sbu', '.', 'ac', '.', 'ir', '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', 'یک', 'متن', 'تست', 'اس', '.', 'حداقل', 'به', 'من', 'اینطور', 'گفته\u200cشده\u200cاست', '.)', 'مجلهٔ', 'تایم', 'در', 'گزارش', 'سال', '2006', 'خود', '،', 'جیمی', 'ویلز', 'را', 'در', 'گروه', '100', 'فرد', 'تاثیرگذار', 'سال', 'اعلام', 'کرد', '.', 'همچنین', 'در', 'همین', 'سال', 'ویکی', 'پدیای', 'روسی', 'برندهٔ', 'جایزهٔ', 'ران', '(', 'روسی', ':', 'Рунета', 'Премия', ')', 'در', 'بخش', '«', 'دانش', 'و', 'آموزش', '»', 'شد', '.', 'این', 'جایزه', 'از', 'طرف', 'دولت', 'اعطا', 'شد&شو', '.', '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', 'داشت&دارد', 'که', 'وظایف', 'غیردانشنامه\u200cای', 'را', 'انجام', 'داد&ده']

صرفا از ریشه یاب آن استفاده کردیم ولی هم کار ریشه یابی هم بت یابی را برایمان انجام داد و میبینیم که به مراتب بهتر و تمیز تر ریشه و بن کلمات را پیدا کرده.

متاسفانه انجام ریشه یابی و بن یابی با دادما با مشکل مواجه شد و بعد از گذاشتن زمان طولانی برای حل مشکل چیزی درست نشد.

3- توکن بندی:

الف) ساده ترین و رایج ترین مدل توکن بندی بر اساس فاصله است یعنی همان جدا کردن بر اساس کلمات موجود در متن و هر کلمه یک توکن میشود.

ب) اول هضم را بررسی میکنیم:

['عنو', 'مقاله', ':', 'صفحه', 'اصل', '<p>متن', 'ویکی\u200cپدیا', 'فارسی</', '/', 'p>', 'من', 'john', 'هس', 'و', 'در', 'ب', '.', 'م', '.', '#هست', 'گیر', 'تخصص', 'دارم', '!', 'کل', 'عرب', 'مانند', 'اصلاح', 'کاف', 'و', 'یا', 'برا', 'توکنایزر', 'ما', 'اهم', 'دارند', '.', 'ما', 'می\u200cدان', 'که', 'در', 'تاریخ', '۲۰', 'سپتامبر', '۲۰۰۴', '(', '۲۹', 'شهریور', '،', '۱۳۸۳', ')', 'مقاله', '«', 'ویک', 'پدیا', '»', 'در', '۱۰۵', 'زب', 'به', 'یک\u200cمیلیون', 'رسید', '.', 'که', 'این', 'مقال', 'شامل', 'زمان', 'پیشین', 'نیستند', '.', 'در', 'ویکی\u200cپدیا', 'فارس', 'ممکن', 'اس', '(', 'گاهی', ')', 'فاصله', 'پرانتز', 'رعا', 'شد#شو', 'باشد', '،', 'یا', 'حت', 'ممکن', 'اس', 'درباره', 'محبوب', 'فارس', 'صحب', 'شده_باشد', '.', 'در', 'اینجا', 'یک', 'ایمیل', 'آزمایش', 'از', 'من', 'sh@sbu', '.', 'ac', '.', 'ir', 'قرار', 'دارد', '.', 'برا', 'اطلاع', 'ب', 'توانست#توان', 'به', 'وبسا', 'ویکی\u200cپیدا', 'فارس', 'به', 'آدرس', 'com', '.', 'wikipedia', '.', '://', 'http', 'سر', 'بزنید', '.', '(', 'داخل', 'پرانتز', 'بگویم', '،', 'این', 'یک', 'متن', 'تست', 'است', '.', 'حداقل', 'به', 'من', 'اینطور', 'گفته\u200cشده', 'است', '.', ')', 'مجله', 'تا', 'در', 'گزار', 'سال', '۲۰۰۶', 'خود', '،', 'جیم', 'ویلز', 'را', 'در', 'گروه', '۱۰۰', 'فرد', 'تأثیرگذار', 'سال', 'اعلا', 'کرد', '.', 'همچنین', 'در', 'همین', 'سال', 'ویک', 'پدیا', 'روس', 'برنده', 'جایزه', 'ران', '(', 'روسی', ':', 'Рунета', 'Премия', ')', 'در', 'بخ', '«', 'دان', 'و', 'آموزش', '»', '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', 'United', 'Nations', ':', 'چندین', 'پروژه', 'متن-آزاد', 'داشت#دار', 'که', 'وظایف', 'غیردانشنامه', 'را', 'انجا', 'داد#ده']

میبینیم که به خوبی کلمات جدا شده و حتی علائم نگارشی هم جدا شده.
حال سراغ پارسیوار میرویم:

['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', '(', '29', 'شهریور', '،', '1383', ')', 'مقاله', '""', 'ویکی', 'پدیا', '""', 'در', '105', 'زبان', 'به', 'یک', 'میلیون', 'رسید', '.', 'که', 'این', 'مقالات', 'شامل', 'زمان', 'پیشین', 'نیستند', '.', 'در', 'ویکی\u200cپدیا', 'فارسی', 'ممکن', 'اس', '(', 'ناهی', ')', 'فاصله', 'پرانتز', 'رعایت\u200cنشده\u200cباشد', '،', 'یا', 'حتی', 'ممکن', 'اس', 'درباره', 'محبوب', 'ترین\u200cهای', 'فارسی', 'صحبت\u200cشده\u200cباشد', '.', 'در', 'اینجا', 'یک', 'ایمیل', 'آزمایشی', 'از', 'من', 'ir', 'ac', '.', 'sh@sbu', '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', 'یک', 'متن', 'تست', 'اس', '.', 'حداقل', 'به', 'من', 'اینطور', 'گفته\u200cشده\u200cاست', '.)', 'مجلهٔ', 'تایم', 'در', 'گزارش', 'سال', '2006', 'خود', '،', 'جیمی', 'ویلز', 'را', 'در', 'گروه', '100', 'فرد', 'تاثیرگذار', 'سال', 'اعلام', 'کرد', '.', 'همچنین', 'در', 'همین', 'سال', 'ویکی', 'پدیای', 'روسی', 'برندهٔ', 'جایزهٔ', 'ران', '(', 'روسی', ':', 'Рунета', 'Премия', ')', 'در', 'بخش', '«', 'دانش', 'و', 'آموزش', '»', 'شد', '.', 'این', 'جایزه', 'از', 'طرف', 'دولت', 'اعطا', 'شد&شو', '.', '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', 'داشت&دارد', 'که', 'وظایف', 'غیردانشنامه\u200cای', 'را', 'انجام', 'داد&ده']

و میبینیم که دقیقا همان محتوای استم شده است و به درستی توکنایز شده. و برای کتابخانه و بررسی تشخیص نیم فاصله:

word_tokenizer.tokenize_words("کتاب‌خانه")

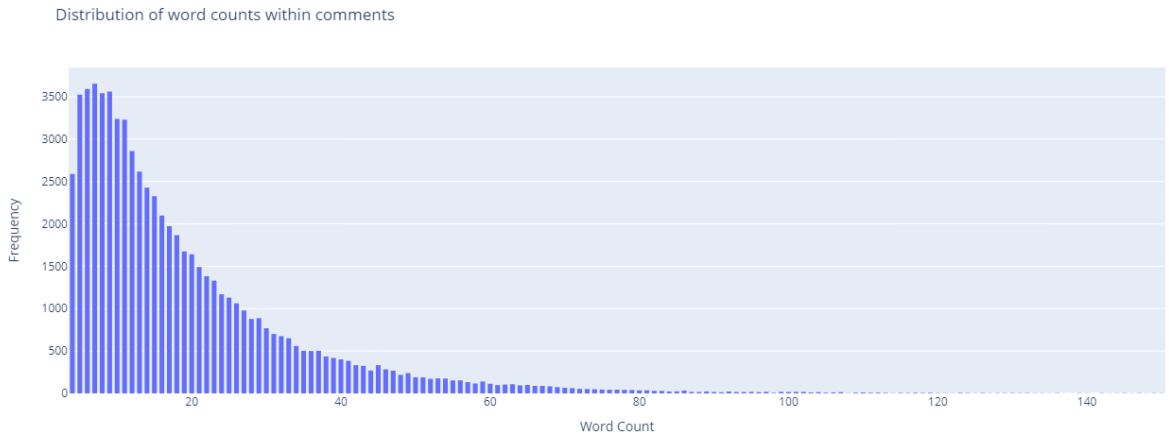
['کتاب\u200cخانه']

میبینیم که کلمه دارای نیم فاصله را یکی در نظر گرفته.
متاسفانه دادما هم در اینجا عمل نکرد.

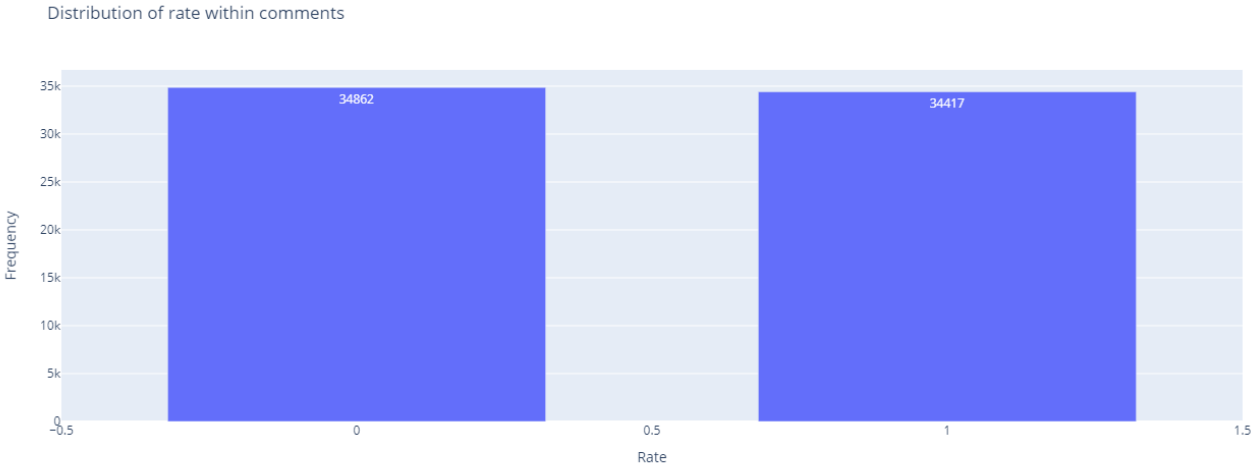
تمرین دوم:

1- مدل زبانی چیست با استفاده از تعریف n-grams. ساده‌ترین مدل‌های زبانی که دنباله کلمات احتمالاتی را تولید می‌کنند مدل‌های n-grams هستند. ما برای پیدا کردن کلمه بعدی در یک متن ناقص به کلمه‌های قبلی احتیاج داریم. و برای هر مثالی برای یافتن کلمه ی بعدی در جمله مان یک احتمال جداگانه به تمام کلمات موجود داده می‌شود. n-gram مجموعه ای از n مورد متوالی در یک داکیومنت متن است که ممکن است شامل کلمات، اعداد، نمادها و علائم نگارشی باشد.

2- اولین کار بعد از لود کردن دیتاست تمیز کردن دیتاس که چون در تمرین قبلی کامل توضیح داده شده بود اینجا موارد لازم فقط اعمال شده و چیزهای اضافه تر بررسی نشده. کار اضافه تری که انجام شد شمردن کلمات بعد از تمیز کردن داده ها بود که متوجه شدیم 99.71% داده ها شامل 3 تا 150 کلمه هستن پس تصمیم گرفتیم که کامنت هایی که تعداد کلماتش تو این بازه قرار نمیگیره رو حذف کنیم. حال نگاهی به تعداد کلمه ها میندازیم و متوجه میشیم کامنتای زیر 60 حرف بخش کوچکیو شامل میشن و اکثر کامنت ها بین 3 تا 20 حرف داشتن.




حالا بررسی میکنیم چه تعداد مثبت و چه تعداد منفی هستن و دیده میشه که حدودا یه اندازه هستن و نیازی نیست که تغییری تو تعداد بدیم.




حالا ادامه تمیز کردن رو انجام میدیم که شامل پاک کردن اعداد و حروف انگلیسی و علامت های اضافی و نگارشیه و از نرمال کننده هضم استفاده میکنیم

متن اولیه: مثل همیشه عاااالی هستی پرپروک جاااان مچکریم ازت

متن ترمال شده: مثل همیشه عالی هستی پرپروک جان مچکریم ازت

حالا دوباره تعداد کلمات رو برسی میکنیم در خط 13 همه حروف انگلیسی بود و الان خالی شده پس چنین خط هایی را پاک میکنیم و به مرحله بعد میرویم.

در این مرحله داده های آموزشی و تست و ولیدیشن را جدا میکنیم. 80% برای آموزش و 10% از 20% باقی مانده هم برای هرکدام از دو گروه دیگر استفاده میشود.

در مرحله بعد نوبت آماده سازی مدل زبانی است. از توکنایزر استفاده میکنیم و میبینیم که چه کرده:

```
Comment: گوشت قیمه بسیار سفت و غیر قابل خوردن بود
Tokens: گوشت قیمه بسیار سفت و غیر قابل خوردن بود
[Token IDs: [5835, 36666, 3177, 12777, 1379, 3268, 3496, 6146, 2834
```

میبینیم به هر کلمه مقداری داده  است و به خوبی کار کرده.
حال نوبت به طرح مدل میرسد در ابتدا از مدل از پیش آماده پارس برت استفاده میکنیم و بعد از دراپ اوت و  سپس لایه کلسیفایر را اضافه میکنیم. بهینه سازی که از آن  استفاده میکنیم آدام است و سپس توابع لازم برای آموزش را تعریف و اضافه میکنیم.
و در آخر منتظر یادگیری مدل میمانیم که در اینجا با وجود جی پی یو ای که کگل به ما داد و حدود 30 دقیقه زمان برد و با تعداد epoch 3 به دقت 86% رسیدیم که باز  هم حدود همان تمرین هفته گذشته شد اما این بار با استفاده از مدل های زبانی.
3- مراحل پاکسازی عین قبل انجام شد و در ادامه ار امبدینگی که برت در اختیار گذاشت استفاده کردیم و یک رندم فارست را آموزش دادیم اما متاسفانه دقت بسیار پایین و 68% شد.