گزارش پروژه بازیابی اطلاعات

فرهاد امان – ۹۹۳۱۰۰۶

ابتدا فایل json حاوی اسناد را با 'encoding='utf-8 باز میکنیم. و با استفاده از لایبرری json آن را load میکنیم.

سپس سه دیکشنری از این فایل جدا میکنیم. کلید هر سه دیکشنری index هر خبر و value هر item از دیکشنری ها به ترتیب content, url, title خبر میباشد.

سپس مقادیر موجود در دیکشنری content_dataset که حاوی متن اخبار است را به شکل زیر normalize میکنیم.

تابع normalizer در فایل جدایی پیاده سازی شده است و در این فایل import شده است که پیاده سازی آن در زیر شرح داده میشود.

همانطور که مشاهده میشود، این تابع در سه مرحله و با فراخوانی 3 فانکشن دیگر، متن ورودی را نرمال میکند. در زیر هرکدام از این توابع را توضیح میدهم.

در تابع replace_unicode_symbols کلماتی که ممکن است بصورت خاصی در خبر ها آمده باشند، مانند (" ") به صورت استاندارد خود تبدیل میشوند. لیست کلمات در تابع بالا مشخص است.

فانکشن apply_patterns_replacement نیز که در بدنه این فانشکن استفاده شده بصورت زیر پیاده سازی شده است.

این تابع با استفاده از re در پایتون که دارای قابلیت های کار با regex است، با pattern ای که در ورودی میگیرد، replacement را انجام میدهد.

فانکشن دیگری که در فرایند نرمال سازی استفاده شده است، adjust_spacing است.

این تابع با الگو های regex ای که دارد، افعالی مانند می شود را به می‌شود تبدیل میکند. همچنین نیم فاصله بین ها جمع با کلمات، تر، گری و ... دیگر مواردی که به جای فاصله به نیم فاصله نیاز دارند را اصلاح میکند. در نهایت با فانکشن apply_patterns_replacement که در بالا توضیح داده شد، جایگزینی رخ میدهد.

در نهایت با تابع زیر همه اعداد انگلیسی به فارسی تبدیل میشوند.

از دو کلاس زیر برای ساخت positional index ها استفاده خواهد شد که در زیر توضیح داده میشود.

این کلاس نماینده posting یک خبر در postings list یک کلمه است. این کلاس شامل تکرار آن کلمه در آن خیر، موقعیت هایی که آن کلمه در آن خبر آمده است و tf-idf آن کلمه در آن خبر است.

این کلاس نیز نشان دهنده postings list های یک کلمه است که شامل تعداد تکرار کلمه در کل خبر ها و لیستی از posting ها برای خبر های متفائت است که این کلمه در آن ها آمده است.

بعد از نرمال سازی متن خبر ها، با قطعه کد زیر توکنایز کردن متون و ساختن ایندکس ها در قطعه کد زیر شکل میگیرد.

ابتدا دیکشنری مربوط به شاخص مکانی ساخته میشود. در حلقه اول متن هر خبر با تابع tokenize توکنایز میشود. پیاده سازی این تابع بصورت زیر است:

قسمت اول این فانکشن با استفاده از ماژول punctuation ،re ها و تب ها و \n ها که نشانه خط جدید است را با space (فاصله) جایگزین میکند. سپس متن خبر را با استفاده از فاصله split میکنیم. یعنی هر کلمه با استفاده از فاصله از هم جدا میشوند و تبدیل به یک توکن میشوند. در نهایت لیستی از punctuation ها را آورده و آن ها را از لیست توکن ها حذف کرده ایم.

بعد از توکنایز کردن متون اخبار، توکن ها را به فانکشن process_verbs میدهیم.

این تابع باعث میشود افعالی که بصورت نخواهم آمد در توکن ها به عنوان دو توکن جدای نخواهم و آمد تبدیل شده اند، به یک توکن واحد بصورت نخواهم_آمد ذخیره شوند تا کلیت فعل حفظ شود(بخش امتیازی ذکر شده در پروژه). به این صورت که توکن ها را از آخر به اول پیمایش میکنیم و اگر با یکی از کلماتی که در لیست ابتدای فانکشن به عنوان before_verbs تعریف شده اند، مواجه شویم، این توکن و توکن قبل از آن را به عنوان یک توکن واحد در نظر بگیریم و با _ به هم وصل کنیم.

حلقه داخلی تر را بررسی میکنیم:

در این حلقه توکن های موجود در یک خبر بررسی میشود.

از enumerate استفاده میکنیم تا index هر توکن را نیز داشته باشیم. در if اول چک میکنیم که که اگر این توکن جزو کلمات vocabulary بود، یعنی قبلا این کلمه را دیده بودیم و برای آن postings list ساخته ایم، frequency آن کلمه را یک عدد افزایش میدهیم، سپس چک میکنیم که آیا ایندکس خبر مربوط به این توکن جزو postings list آن کلمه میباشد یا نه. اگر این خبر از قبل جزو postings list آن کلمه نبود، یک posting برای ذخیره مکان های وقوع آن کلمه در خبر

میسازیم و سپس frequency آن کلمه در آن خبر را نیز یک واحد افزایش میدهیم. در نهایت محل وقوع آن کلمه در آن خبر را در positional index ذخیره میکنیم.

اما اگر این کلمه از قبل در vocabulary ما موجود نبود، یک posting list خالی برای آن میسازیم. سپس frequency آن کلمه را یک واحد افزایش میدهیم. سپس posting مربوط به ایندکس آن خبر را میسازیم و frequency آن کلمه در آن خبر را یک واحد افزایش میدهیم. سپس محل وقوع آن کلمه در خبر را دخیره کرده و در نهایت posting ساخته شده را در postings list آن کلمه قرار میدهیم.

در قطعه کد زیر پس از ساخته شدن positional index ها کلمات موجود در vocabulary را ریشه یابی میکنیم.

برای ریشه یابی کلمات از کتابخانه hazm استفاده میکنیم. روی کلمات vocabulary پیمایش میکنیم و آن ها را ریشه یابی میکنیم. اگر کلمه ریشه یابی شده از قبل در vocabulary وجود نداشت، کلمه ریشه یابی شده بجای کلمه اصلی قرار میگیرد و postings list تغییری نمیکند. اما امکان دارد چند کلمه بعد از ریشه یابی با هم برابر شوند. در این حالت این چند کلمه را یک postings list در نظر میگیریم و posting آن کلمه ها را توسط merge_posting_lists ادغام میکنیم که پیاده سازی آن در زیر توضیح داده شده است.

در قطعه کد که کار ادغام را انجام میدهد مشاهده میشود که برای به دست آوردن frequency کلی از جمع frequency دو posting list استفاده میکنیم. برای خبر هایی که در هر دو posting list وجود دارند و جزو اشتراک هر دو هستند، position ها را با هم ترتیب کرده و مرتب میکنیم تا ترتیب حفظ شود. اگر هم خبری صرفا در یکی از posting list ها آمده بود، position های مربوط به آن خبر را از آن posting list میگیریم.

ابتدا کلمات موجود در positional index را براساس frequency آن ها مرتب میکنیم. سپس 50 کلمه اول را از positional index ها حذف میکنیم.

لیست کلمات حذف شده:

و', 'در', 'به', 'از', 'این', 'که', 'با', 'را', 'است', '2', 'برای', 'کرد', 'تیم', 'هم', '1', 'ما', 'یک', 'شد', 'ان', 'بر', 'تا', 'کشور', 'باید', 'وی', 'بازی', 'بود', 'شده', 'خود', 'فارس', 'مجلس', 'اسلامی', 'گفت', 'گزارش', 'پیام', 'ایران', 'مردم', 'خبرگزاری', 'انتهای', 'اما', 'دولت', 'شود', 'دارد', 'ملی', 'سال', 'داشت', 'اینکه', 'قرار', 'دو', '4', 'رئیس'

در این قسمت از کد، tf-idf هر کلمه در هر خبر را به دست می‌آوریم. یک دیکشنری doc_length هم در نظر گرفتیم که مربع هر tf-idf ای که حساب میکنیم را جمع کرده و در آن میریزیم که بعدا برای نرمال کردن طول خبر ها استفاده شود.

در این حلقه از تابع tf-idf استفاده کردیم که در زیر توضیح داده شده است:

این فانکشن دو ورودی کلمه و شماره خبر را میگیرد. f_t_in_d تعداد تکرار این کلمه در این خبر است. n تعداد کل خبر هاست. n_t تعداد تعداد خبر هایی است که این کلمه در آن ها آمده است. در نهایت در خط آخر با فرمول tf-idf مقدار آن به دست آمده است.

در این کد از دیکشنری که بالاتر حاوی ایندکس خبر ها و جمع مربعات امتیاز آن ها است استفاده شده است. به این صورت که tf-idf هر خبر برای هر کلمه را تقسیم بر جذر مجموع مربعات tf-idf های آن خبر میکنیم. با این کار طولانی تر بودن خبر باعث امتیاز بالاتر آن نمیشود.

ابتدا postings list های مربوط به هر کلمه را براساس امتیاز tf-idf آن ها مرتب میکنیم. سپس یک inverted index جدید با همان کلمات و 20 خبر با امتیاز بالاتر tf-idf برای هر کلمه میسازیم.

در نهایت با دوتابع زیر inverted index های ساخته شده را با استفاده از pickle در فایل ذخیره میکنیم.

در فایلی دیگر به نام search_engine.py از فایل های inverted index ساخته شده استفاده میکنیم و به کوئری های کاربر پاسخ میدهیم.

در این قسمت از کد مانند فایل قبلی فایل json اخبار را میخوانیم و در سه دیکشنری content_dataset و url_dataset و title_dataset ذخیره میکنیم.

در این قسمت از کد، inverted index های ذخیره شده در فایل را توسط load ،pickle میکنیم.

در این تابع، ابتدا کوئری دریافتی را با استفاده از فانکشنی که در قبل توصیح داده شده و برای متن اسناد استفاده شده بود، نرمال میکنیم. سپس با فانکشنی که متن اخبار را توکنایز کردیم، کوئری را هم توکنایز میکنیم. در آخر نیز با استفاده از کتابخانه hazm توکن های به دست آمده را ریشه یابی میکنیم.

در این تابع برای هر توکن موجود در کوئری، یک دور تمام خبر ها را پیمایش میکنیم و امتیاز tf-idf ان کلمه در آن خبر را برای آن خبر جمع میکنیم و در نهایت هر خبر یک امتیاز نهایی برای کل کلمات کوئری دارد.

در این تابع هم خبر ها را بر اساس امتیاز به دست آمده در مرحله قبل مرتب میکنیم و k خبر با امتیاز برتر را برمیگردانیم.

در این قطعه کد کوئری ورودی را از کاربر میگیریم. بعد با متدی که در بالا نوضیح توکن های کوئری استخراج میشود. و با متد های ذکر شده، امتیاز خبر ها محاسبه شده و index ده خبر برتر برمیگردد. در اینجا حق انتخاب داریم که جستجو بین positional index عادی رخ دهد یا برای

افزایش سرعت بین champion list رخ دهد. در نهایت عنوان و امتیاز و لینک خبر های برگشتی چاپ میشود.

در اینجا زوند ساخت ایندکس برای یک خبر را میبینیم:

متن اصلی خبر(با doc_id = 4092):

به گزارش خبرگزاری فارس، سردار آزمون ستاره تیم ملی کشورمان و عضو باشگاه زنیت به دلیل درخشش بی نظیرش در لیگ روسیه و نزدیک بودن زمان پایان قراردادش با باشگاه روسی مشتریان زیادی پیدا کرده است. سیماک سرمربی زنیت روز گذشته در مصاحبه با رسانه های روسیه اعلام کرد مسئله آینده آزمون آسان است این بازیکن قصد تمدید قراردادش را ندارد و در تابستان به باشگاه جدیدی می رود. در همین رابطه سایت «hitc» به تحلیل حرف های سرمربی زنیت پرداخت و این حرف ها را فرصتی استثنایی برای دو باشگاه اورتون و نیوکاسل دانست. این رسانه انگلیسی، نوشت: سردار آزمون ستاره ایرانی زنیت در لیگ قهرمانان اروپا در یک بازی جذاب به کابوس هواداران چلسی تبدیل شد. این ستاره ایرانی یک گل فوق العاده در این بازی به ثمر رساند و اگر واکنش های فوق العاده کپه آ نبود می توانست چندین بار دیگر دروزاه شاگردان توخل را باز کند. ستاره ایرانی از دو سال گذشته بنا به اعلام مدیر برنامه اش از باشگاه اورتون با ریاست فرهاد مشیری ایرانی پیشنهاد داشته و یکی از گزینه های تقویت خط حمله این تیم است. از طرف دیگر نیوکاسل که در آستانه سقوط قرار دارد برای رهایی از این وضعیت بد می خواهد با مدیریت جدید سعودی ها سردار آزمون را جذب کند تا خط حمله اش جانی دوباره بگیرد. قرارداد سردار آزمون تابستان پیش رو با زنیت به پایان می رسد و این بازیکن 17 میلیون پوندی به صورت رایگان به تیمی دیگر خواهد رفت. نیوکاسل و اورتون فرصتی طلایی برای جذب این بازیکن به صورت رایگان را دارند. البته مهاجم ملی پوش ایرانی از سوی باشگاه های یوونتوس، لیون و بایر لورکوزن نیز مورد توجه است و باید دید این بازیکن ایرانی 26 ساله چه تصمیمی برای آینده اش در تابستان می گیرد. آزمون از سال 2019 برای زنیت بازی می کند. ستاره ایرانی در مدت کمتر از 3 سال 3 بار قهرمانی لیگ برتر روسیه را تجربه کرد و یک بار نیز جایزه آقای گلی لیگ را از آن کرد. انتهای پیام/

سپس متن خبر را normalize میکنیم: این کار شامل تبدیل کلمات خاص مثل محمد **به محمد است. همچنین اعداد انگلیسی به فارسی تبدیل میشوند. علاوه بر آن فاصله گذاری کلمات هم تصحیح می‌شوند و برای مثال کتاب های تبدیل به کتاب‌های می‌شوند.**

متن نرمال شده:

به گزارش خبرگزاری فارس، سردار ازمون ستاره تیم ملی کشورمان و عضو باشگاه زنیت به دلیل درخشش بی نظیرش در لیگ روسیه و نزدیک بودن زمان پایان قراردادش با باشگاه روسی مشتریان زیادی پیدا کرده است. سیماک سرمربی زنیت روز گذشته در مصاحبه با رسانه‌های روسیه اعلام کرد مسئله اینده ازمون اسان است این بازیکن قصد تمدید قراردادش را ندارد و در تابستان به باشگاه جدیدی می‌رود. در همین رابطه سایت «hitc» به تحلیل حرف‌های سرمربی زنیت پرداخت و این حرف‌ها را فرصتی استثنایی برای دو باشگاه اورتون و نیوکاسل دانست. این رسانه انگلیسی، نوشت: سردار ازمون ستاره ایرانی زنیت در لیگ قهرمانان اروپا در یک بازی جذاب به کابوس هواداران چلسی تبدیل شد. این ستاره ایرانی یک گل فوق العاده در این بازی به ثمر رساند و اگر واکنش‌های فوق العاده کپه ا نبود می‌توانست چندین بار دیگر دروزاه شاگردان توخل را باز کند. ستاره ایرانی از دو سال گذشته بنا به اعلام مدیر برنامه‌اش از باشگاه اورتون با ریاست فرهاد مشیری ایرانی پیشنهاد داشته و یکی از گزینه‌های تقویت خط حمله این تیم است. از طرف دیگر نیوکاسل که در استانه سقوط قرار دارد برای رهایی از این وضعیت بد می‌خواهد با مدیریت جدید سعودی‌ها سردار ازمون را جذب کند تا خط حمله‌اش جانی دوباره بگیرد. قرارداد سردار ازمون تابستان پیش رو با زنیت به پایان می‌رسد و این بازیکن ۱۷ میلیون پوندی به صورت رایگان به تیمی دیگر خواهد رفت. نیوکاسل و اورتون فرصتی طلایی برای جذب این بازیکن به صورت رایگان را دارند. البته مهاجم ملی پوش ایرانی از سوی باشگاه‌های یوونتوس، لیون و بایر لورکوزن نیز مورد توجه است و باید دید این بازیکن ایرانی ۲۶ ساله چه تصمیمی برای اینده‌اش در

تابستان می‌گیرد. ازمون از سال ۲۰۱۹ برای زنیت بازی می‌کند. ستاره ایرانی در مدت کمتر از ۳ سال ۳ بار قهرمانی لیگ برتر روسیه را تجربه کرد و یک بار نیز جایزه اقای گلی لیگ را از ان کرد. انتهای پیام/

بعد از نرمال سازی مشاهد می‌شود که (آ) ها در خبر تبدیل به (ا) میشوند و برای مثال کلمه آقای تبدیل به کلمه اقای می‌شود.

همچنین اعداد انگلیسی به فارسی تبدیل میشوند.(در خبر نرمال نشده به دلیل فارسی بودن فونت در word اعداد فارسی هستند ولی در نتیجه کد، اعداد در حالت نرمال نشده، انگلیسی چاپ می‌شوند.)

فعل می گیرد تبدیل به می‌گیرد شده است. (هایلایت شده)

باشگاه های تبدیل به باشگاه‌های شده است.(هایلایت شده)

توکن سازی: در فرایند ساخت توکن ها ایمیل ها و آیدی ها و لینک ها و و اعداد اعشاری و صحیح شناسایی می‌شوند تا به درستی توکنایز شوند. به علاوه tab و \n به فاصله تبدیل می‌شوند. سپس کلمات با فاصله بین آن ها به توکن ها تبدیل می‌شوند. و در نهایت punctuation ها از بین توکن ها حذف می‌شوند.

توکن های استخراج شده از خبر اول:

['به', 'گزارش', 'خبرگزاری', 'فارس', 'سردار', 'ازمون', 'ستاره', 'تیم', 'ملی', 'کشورمان', 'و', 'عضو', 'باشگاه', 'زنیت', 'به', 'دلیل', 'درخشش', 'بی', 'نظیرش', 'در', 'لیگ', 'روسیه', 'و', 'نزدیک', 'بودن', 'زمان', 'پایان', 'قراردادش', 'با', 'باشگاه', 'روسی', 'مشتریان', 'زیادی', 'پیدا', 'کرده', 'است', 'سیماک', 'سرمربی', 'زنیت', 'روز', 'گذشته', 'در', 'مصاحبه', 'با', 'رسانه‌های', 'روسیه', 'اعلام', 'کرد', 'مسئله', 'اینده', 'ازمون', 'اسان', 'است', 'این', 'بازیکن', 'قصد', 'تمدید', 'قراردادش', 'را', 'ندارد', 'و', 'در', 'تابستان', 'به', 'باشگاه', 'جدیدی', 'می‌رود', 'در', 'همین', 'رابطه', 'سایت', 'hitc', 'به', 'تحلیل', 'حرف‌های', 'سرمربی', 'زنیت', 'پرداخت', 'و', 'این', 'حرف‌ها', 'را', 'فرصتی', 'استثنایی', 'برای', 'دو', 'باشگاه', 'اورتون', 'و', 'نیوکاسل', 'دانست', 'این', 'رسانه', 'انگلیسی', 'نوشت', 'سردار', 'ازمون', 'ستاره', 'ایرانی', 'زنیت', 'در', 'لیگ', 'قهرمانان', 'اروپا', 'در', 'یک', 'بازی', 'جذاب', 'به', 'کابوس', 'هواداران', 'چلسی', 'تبدیل', 'شد', 'این', 'ستاره', 'ایرانی', 'یک', 'گل', 'فوق', 'العاده', 'در', 'این', 'بازی', 'به', 'ثمر', 'رساند', 'و', 'اگر', 'واکنش‌های', 'فوق', 'العاده', 'کپه', 'ا', 'نبود', 'می‌توانست', 'چندین', 'بار', 'دیگر', 'دروزاه', 'شاگردان', 'توخل', 'را', 'باز', 'کند', 'ستاره', 'ایرانی', 'از', 'دو', 'سال', 'گذشته', 'بنا', 'به', 'اعلام', 'مدیر', 'برنامه‌اش', 'از', 'باشگاه', 'اورتون', 'با', 'ریاست', 'فرهاد', 'مشیری', 'ایرانی', 'پیشنهاد', 'داشته', 'و', 'یکی', 'از', 'گزینه‌های', 'تقویت', 'خط', 'حمله', 'این', 'تیم', 'است', 'از', 'طرف', 'دیگر', 'نیوکاسل', 'که', 'در', 'استانه', 'سقوط', 'قرار', 'دارد', 'برای', 'رهایی', 'از', 'این', 'وضعیت', 'بد', 'می‌خواهد', 'با', 'مدیریت', 'جدید', 'سعودی‌ها', 'سردار', 'ازمون', 'را', 'جذب', 'کند', 'تا', 'خط', 'حمله‌اش', 'جانی', 'دوباره', 'بگیرد', 'قرارداد', 'سردار', 'ازمون', 'تابستان', 'پیش', 'رو', 'با', 'زنیت', 'به', 'پایان', 'می‌رسد', 'و', 'این', 'بازیکن', '2', 'میلیون', 'پوندی', 'به', 'صورت', 'رایگان', 'به', 'تیمی', 'دیگر', 'خواهد', 'رفت', 'نیوکاسل', 'و', 'اورتون', 'فرصتی', 'طلایی', 'برای', 'جذب', 'این', 'بازیکن', 'به', 'صورت', 'رایگان', 'را', 'دارند', 'البته', 'مهاجم', 'ملی', 'پوش', 'ایرانی', 'از', 'سوی', 'باشگاه‌های', 'یوونتوس', 'لیون', 'و', 'بایر', 'لورکوزن', 'نیز', 'مورد', 'توجه', 'است', 'و', 'باید', 'دید', 'این', 'بازیکن', 'ایرانی', '2', 'ساله', 'چه', 'تصمیمی', 'برای', 'اینده‌اش', 'در', 'تابستان', 'می‌گیرد', 'ازمون', 'از', 'سال', '4', 'برای', 'زنیت', 'بازی', 'می‌کند', 'ستاره', 'ایرانی', 'در', 'مدت', 'کمتر', 'از', '1', 'سال', '1', 'بار', 'قهرمانی', 'لیگ', 'برتر', 'روسیه', 'را', 'تجربه', 'کرد', 'و', 'یک', 'بار', 'نیز', 'جایزه', 'اقای', 'گلی', 'لیگ', 'را', 'از', 'ان', 'کرد', 'انتهای', 'پیام']

توکن های خبر بعد از حذف کلمات پرتکرار:

['سردار', 'ازمون', 'ستاره', 'کشورمان', 'عضو', 'باشگاه', 'زنیت', 'دلیل', 'درخشش', 'بی', 'نظیرش', 'لیگ', 'روسیه', 'نزدیک', 'بودن', 'زمان', 'پایان', 'قراردادش', 'باشگاه', 'روسی', 'مشتریان', 'زیادی', 'پیدا', 'کرده', 'سیماک', 'سرمربی', 'زنیت', 'روز', 'گذشته', 'مصاحبه', 'رسانه‌های', 'روسیه', 'اعلام', 'مسئله', 'اینده', 'ازمون', 'اسان', 'بازیکن', 'قصد', 'تمدید', 'قراردادش', 'ندارد', 'تابستان', 'باشگاه', 'جدیدی', 'می‌رود', 'همین', 'رابطه', 'سایت', 'hitc', 'تحلیل', 'حرف‌های', 'سرمربی', 'زنیت', 'پرداخت', 'حرف‌ها', 'فرصتی', 'استثنایی', 'باشگاه', 'اورتون', 'نیوکاسل', 'دانست', 'رسانه', 'انگلیسی', 'نوشت', 'سردار', 'ازمون', 'ستاره', 'ایرانی', 'زنیت', 'لیگ', 'قهرمانان', 'اروپا', 'جذاب', 'کابوس', 'هواداران', 'چلسی', 'تبدیل', 'ستاره', 'ایرانی', 'گل', 'فوق', 'العاده', 'ثمر', 'رساند', 'اگر', 'واکنش‌های', 'فوق', 'العاده', 'کپه', 'ا', 'نبود', 'می‌توانست', 'چندین', 'بار', 'دیگر', 'دروزاه', 'شاگردان', 'توخل', 'باز', 'کند', 'ستاره', 'ایرانی', 'گذشته', 'بنا', 'اعلام', 'مدیر', 'برنامه‌اش', 'باشگاه', 'اورتون', 'ریاست', 'فرهاد', 'مشیری', 'ایرانی', 'پیشنهاد', 'داشته', 'یکی', 'گزینه‌های', 'تقویت', 'خط', 'حمله', 'طرف', 'دیگر', 'نیوکاسل', 'استانه', 'سقوط', 'رهایی', 'وضعیت', 'بد', 'می‌خواهد', 'مدیریت', 'جدید', 'سعودی‌ها', 'سردار', 'ازمون', 'جذب', 'کند', 'خط', 'حمله‌اش', 'جانی', 'دوباره', 'بگیرد', 'قرارداد', 'سردار', 'ازمون', 'تابستان', 'پیش', 'رو', 'زنیت', 'پایان', 'می‌رسد', 'بازیکن', 'میلیون', 'پوندی', 'صورت', 'رایگان', 'تیمی', 'دیگر', 'خواهد', 'رفت', 'نیوکاسل', 'اورتون', 'فرصتی', 'طلایی', 'جذب', 'بازیکن', 'صورت', 'رایگان', 'دارند', 'البته', 'مهاجم', 'پوش', 'ایرانی', 'سوی', 'باشگاه‌های', 'یوونتوس', 'لیون', 'بایر', 'لورکوزن', 'نیز', 'مورد', 'توجه', 'دید', 'بازیکن', 'ایرانی', 'ساله', 'چه', 'تصمیمی', 'اینده‌اش', 'تابستان', 'می‌گیرد', 'ازمون', 'زنیت', 'می‌کند', 'ستاره', 'ایرانی', 'مدت', 'کمتر', 'بار', 'قهرمانی', 'لیگ', 'برتر', 'روسیه', 'تجربه', 'بار', 'نیز', 'جایزه', 'اقای', 'گلی', 'لیگ']

مشاهده می‌شود که کلماتی مانند (و)، (کرد)، (از) حذف شده‌اند. همچنین کلمات (انتهای) و (پیام) که در آخر هر خبر وجود دارد از توکن ها حذف گردید.