گزارش مرحلهٔ اول فاز اول پروژهٔ درس بازیابی اطلاعات – پیش‌پردازش اسناد

نگار موقتیان، ۹۸۳۱۰۶۲

با ذکر مثال شرح دهید که در گام پیش‌پردازش چه عملیاتی انجام داده‌اید. همچنین دلیل انجام هر پردازش را ذکر کنید.

اولین عملی که بر روی متن هر یک از اسناد صورت می‌گیرد نرمال‌سازی متن سند است. در این مرحله تمام کلمات یکسان به فرم مشابهی در می‌آیند و یک شکل نوشتاری یکسان پیدا می‌کنند. برای مثال فاصله‌ها و نیم‌فاصله‌های میان واژه‌های هر کلمه نرمال‌سازی می‌شود و دو فعل «می‌گوید» و «می گوید» هر دو در نهایت به شکل «می‌گوید» نوشته خواهند شد. و یا کلماتی مانند «تهران» و «طهران» که دو شکل نوشتاری از یک کلمهٔ یکسان هستند به فرم نرمال «تهران» در خواهند آمد. یا اینکه تنوین کلمات حذف خواهد شد تا تمایزی میان «حتماً» و «حتما» وجود نداشته باشد. این کار باعث می‌شود حجم dictionary کم شده و کاربر به نتایج بهتر و جامع‌تری برسد.

دو متن زیر مربوط به بخشی از سند اول، پیش و پس از نرمال‌سازی هستند:

« به گزارش خبرگزاری فارس، کنفدراسیون فوتبال آسیا (AFC) در نامه ای رسمی به فدراسیون فوتبال ایران و باشگاه گیتی پسند زمان قرعه کشی جام باشگاه های فوتسال آسیا را رسماً اعلام کرد. بر این اساس ۲۵ فروردین ماه ۱۴۰۱ مراسم قرعه کشی جام باشگاه های فوتسال آسیا در مالزی برگزار می شود...»

« به گزارش خبرگزاری فارس، کنفدراسیون فوتبال آسیا (AFC) در نامه‌ای رسمی به فدراسیون فوتبال ایران و باشگاه گیتی پسند زمان قرعه کشی جام باشگاه‌های فوتسال آسیا را رسما اعلام کرد. بر این اساس ۲۵ فروردین ماه ۱۴۰۱ مراسم قرعه کشی جام باشگاه‌های فوتسال آسیا در مالزی برگزار می‌شود...»

در مرحلهٔ بعد متن نرمال‌سازی شده tokenized می‌شود. در این مرحله تمام کلمات و تمام علائم نگارشی از یکدیگر جدا خواهند شد.

برای مثال جملهٔ اول متن بالا به توکن‌های زیر شکسته می‌شود:

['به', 'گزارش', 'خبرگزاری', 'فارس', '،', 'کنفدراسیون', 'فوتبال', 'آسیا', '(', 'AFC', ')', 'در', 'نامه‌ای', 'رسمی', 'به', 'فدراسیون', 'فوتبال', 'ایران', 'و', 'باشگاه', 'گیتی', 'پسند', 'زمان', 'قرعه', 'کشی', 'جام', 'باشگاه‌های', 'فوتسال', 'آسیا', 'را', 'رسما', 'اعلام', 'کرد', '.']

در مرحلهٔ بعد هر یک از توکن‌های پیدا شده ریشه‌یابی می‌شوند. در این مرحله حروف اضافهٔ بعد از کلمات حذف شده (مانند «ای» یا نشانهٔ جمع) و کلمات به ریشهٔ خود تبدیل می‌شوند. این کار باعث کم شدن حجم dictionary خواهد شد و باعث می‌شود که کاربر در مواردی به نتیجهٔ مطلوب‌تری برسد (برای مثال زمانی که کاربر به دنبال کلمهٔ «نامه» است انتظار دارد متنی که حاوی کلمهٔ «نامه‌ای» هست نیز برای او نمایش داده شود).

برای مثال توکن‌های مرحلهٔ قبل به توکن‌های زیر تبدیل می‌شوند:

['به', 'گزارش', 'خبرگزاری', 'فارس', '،', 'کنفدراسیون', 'فوتبال', 'آسیا', '(', 'AFC', ')', 'در', 'نامه', 'رسم', 'به', 'فدراسیون', 'فوتبال', 'ایران', 'و', 'باشگاه', 'گیتی', 'پسند', 'زمان', 'قرعه', 'کشید#کش', 'جام', 'باشگاه', 'فوتسال', 'آسیا', 'را', 'رسما', 'اعلام', 'کرد#کن', '.']

در مرحلهٔ آخر stopword ها و علائم نگارشی از میان توکن‌ها حذف می‌شوند. این کار باعث می‌شود حجم index ها به شدت کاهش یابد. این کلمات، کلمات بسیار پرتکراری هستند که تاثیر چندانی نیز در نتیجهٔ جستجوی کاربر ندارند، لذا ذخیره‌سازی آن‌ها بیهوده است.

برای مثال توکن‌های مرحلهٔ قبل به توکن‌های زیر تبدیل می‌شوند:

['گزارش', 'خبرگزاری', 'فارس', 'کنفدراسیون', 'فوتبال', 'آسیا', 'AFC', 'نامه', 'رسم', 'فدراسیون', 'فوتبال', 'ایران', 'باشگاه', 'گیتی', 'پسند', 'زمان', 'قرعه', 'کشید#کش', 'جام', 'باشگاه', 'فوتسال', 'آسیا', 'رسما', 'اعلام']

همانطور که مشاهده می‌شود در این مرحله تعداد توکن‌ها از ۳۴ به ۲۴ کاهش می‌یابد.