به نام خدا

گزارش تمرین دوم درس NLP

دانشجو: محمد قربانی-97131099

دانشگاه صنعتی امیرکبیر – دانشکده مهندسی کامپیوتر

بهار 97-98

روش اول (word2vec و استفاده از میانگین ساده):

در این روش ابتدا یک پیش پردازش اولیه بر روی داده ترین انجام صورت گرفت و کلمات ایست که تنها سربار محاسباتی به سیستم اضافه می­کند حذف شدند. بعد از حذف کلمات ایست، مطالب را بر روی یک فایل ریخته و سپس بر روی آنها یک word2vec ساختیم. در ساخت مدل مقدار min\_count را برابر با 1 در نظر گرفتم که مشکل کلمات دیده نشده بوجود نیاید. سپس با استفاده از این مدل بردار کلمات هر داکیومنت را به­دست آوردیم و سپس این بردارها را برای هر داکیومنت جمع کردیم و تقسیم بر تعداد کلمات آن داکیومنت­ها کردیم و این بردار را به عنوان بازنمایش آن داکیومنت در نظر گرفتیم. بعد از ایجاد بردار بازنمایی هر داکیومن این بردارها را به الگوریتم کلاسترینگ خود دادیم تا مدل کلاسترینگ خود را ایجاد کنیم. بعد از ایجاد خوشه­بند، بردار هر داکیومنت داده ترین خود را به خوشه­بند دادیم تا آن را خوشه­بندی کند. سپس بیشترین تعداد موضوع در هر خوشه را به عنوان لیبل آن خوشه در نظر گرفتیم و سپس به محاسبه NMI، Accuracy، F-Measure و V-Measure پرداختیم.

|  |  |  |
| --- | --- | --- |
| داده تست | داده ترین | معیار |
| 0.58 | 0.56 | NMI |
| 0.72 | 0.72 | Accuracy |
| 0.74 | 0.73 | F-Measure |
| 0.58 | 0.56 | V-Measure |

NMI یکی دیگر از معیارهایی است که برای سنجش کیفیت خوشه­بندی استفاده می­شود. این معیار با در نظر گرفتن خوشه­بندی واقعی آیتم­ها و همچنین خوشه­بندی تولید شده توسط الگوریتم شباهت بین این دو مجموعه از خوشه­ها را بررسی می­کند. این معیار بررسی می­کند که این دو دسته از خوشه­ها تا چه اندازه به هم مشابه و تا چه اندازه نسبت به هم متفاوت هستند. این معیار مقداری بین 0 و 1 دارد که مقدار 1 بیانگر خوشه­بندی ایده آل می­باشد. مزیت معیار NMI این است که نسبت به تعداد خوشه­های در نظر گرفته شده نرمال می­شود.

نتایج حاصل از این روش نشان­ می­دهند که این روش دقت نسبتا خوبی داشته است.

روش دوم (word2vec و استفاده از میانگین وزن­دار):

در این روش بردار هر کلمه در وزن tfidf آن کلمه ضرب شده است. مقدار tfidf به صورت زیر محاسبه می­شود:

که در آن N بیانگر تعداد کل اسناد موجود در پیکره است و dfi بیانگر تعداد داکیومنت­هایی که کلمه مورد نظر را دارند است. استفاده از Idf موجب می­شود که مشکل کلمات ایست که سربار محاسباتی دارند و هیچ گونه اطلاعات مفیدی نمی­دهند حل شود. کلمات ایست حاصل idf کمی می­گیرند و این باعث می­شود که در محاسبات ارزش کمی بگیرند.

نتایج حاصل از این روش به صورت زیر می­باشد. در مقایسه با روش اول کمی کاهش اندازه معیارها را نشان می­دهد. دلیل این امر این است که ما در روش اول خود در ابتدا در یک پردازش اولیه کلمات ایست را از پیکره حذف نمودیم و سپس به آموزش مدل­های خود پرداختیم اما در این روش این کار به اثرات حاصل از فرمول idf سپردیم که نتایج نشان می­دهد که بهتر است حذف کلمات ایست را به عنوان بخشی از اقدامات پیش پردازشی خود انجام دهیم.

|  |  |  |
| --- | --- | --- |
| داده تست | داده ترین | معیار |
| 0.54 | 0.52 | NMI |
| 0.67 | 0.66 | Accuracy |
| 0.66 | 0.64 | F-Measure |
| 0.54 | 0.52 | V-Measure |

روش سوم (doc2vec):

در این روش برای بازنمایی هر سند از الگوریتم doc2vec استفاده شد. این روش در واقع یک الگوریتم بدون سرپرستی می­باشد که در آن برای جمله/پاراگراف/داکیومنت یک بردار ایجاد می­کند. این الگوریتم در واقع الهام گرفته از word2vec می­باشد که در آن برای هر کلمه یک بردار ایجاد می­کند. در این روش کلمات به صورت مستقل از هم در نظر گرفته نمی­شود و یک رابطه­ی سیکونسی بین کلمات در نظر گرفته می­شود.

|  |  |  |
| --- | --- | --- |
| داده تست | داده ترین | معیار |
| 0.18 | 0.17 | NMI |
| 0.50 | 0.51 | Accuracy |
| 0.47 | 0.46 | F-Measure |
| 0.18 | 0.17 | V-Measure |

این روش نسبت به دو روش قبل دارای نتایج ضعیف­تری می­باشد که این موضوع کمی عجیب به نظر می­رسد چرا که انتظار ما این بود که در این روش نسبت به دو روش قبل نتایج بهتری مشاهده کنیم. به همین دلیل به بررسی چندین باره روند کار پرداختیم اما متوجه دلایل نتایج ضعیف نشدیم.

روش چهارم (document-term):

در این روش برای بازنمایی هر سند از ماتریس سند-کلمه استفاده شد. اندازه­ی این ماتریس برابر با مقداری در حدود 8600\*65000 شد. برای بازنمایی از کل سند­های آموزش و ارزیابی استفاده شد. بعد از این که ماتریس حاصل محاسبه شد ما برای این که آن را تبدیل به یک ماتریس dense کنیم از الگوریتم svd استفاده کردیم. از آنجایی که این ماتریس دارای مقادیر صفر زیاد بود برخی از الگوریتم­ها در زمان مناسب پاسخ­گو نبوده و ما مجبور به استفاده از الگوریتم­های خاصی برای بدست آوردن ماتریس dense شدیم. الگوریتم مورد استفاده ما svds که از کتابخانه­ی scipy است بود.

بعد از بدست آوردن ماتریس dense با استفاده از قسمت مربوط به آموزش آن، مدل خوشه­بندی خود را آموزش دادیم. بعد از خوشه­بندی، طبق الگوی گفته شده هر خوشه را ورچسب­گذاری کرده و به محاسبه معیارهای خواسته شده پرداختیم. در ادامه به سراغ داده تست رفته و بازنمایی که برای هر کدام از این اسناد به دست آورده بودیم را به مدل خوشه­بند خود داده تا ورچسب آن را پیش­بینی کند. سپس به محاسبه معیارهای خواسته شده پرداختیم که نتایج حاصل آن به صورت زیر می­باشد:

|  |  |  |
| --- | --- | --- |
| داده تست | داده ترین | معیار |
| 0.04 | 0.04 | NMI |
| 0.27 | 0.28 | Accuracy |
| 0.42 | 0.42 | F-Measure |
| 0.04 | 0.04 | V-Measure |

این روش نسبت به روش­های قبل از نتایج ضعیف­تری برخوردار می­باشد که به نظر نویسنده یکی از دلایل آن می­تواند استفاده از الگوریتم svd نامناسب باشد.