
بسم الله الرحمن الرحيم

دانشکده‌ی مهندسی برق و کامپیوتر

پایان‌نامه‌ی کارشناسی ارشد در رشته‌ی مهندسی کامپیوتر- هوش مصنوعی

ارائه‌ی یک مدل آماری برای خوشه‌بندی متون

به وسیله‌ی

حمید محمودی

استاد راهنما

دکتر اقبال منصوری

تیر ماه ۱۳۹۱

بـــه نـــام خـــدا

اظهار نامـــه

اینجانب حمید محمودی دانشجوی رشته ی مهندسی کامپیوتر گرایش هوش مصنوعی، دانشکده مهندسی برق و کامپیوتر اظهار می کنم که این پایان نامه حاصل پژوهش خودم بوده و در جاهایی از منابع دیگران استفاده کرده ام.نشانی دقیق و مشخصات کامل آنرا نوشته ام.همچنین اظهار می کنم که تحقیق و موضوع پایان نامه ام تکراری نیست و تعهد می نمایم که بدون مجوز دانشگاه دستاورد های آنرا منتشر ننموده و در اختیار غیر قرار ندهم.کلیه حقوق این اثر مطابق با آیین نامه مالکیت فردی و معنوی متعلق به دانشگاه شیراز است.

نام و نام خانوادگی: حمید محمودی

تاریخ و امضا: ۱۳۹۱/۵/۱۴

بنام خدا

ارائه‌ی یک مدل آماری برای خوشه‌بندی متون

به کوشش

حمید محمودی

پایان نامه

ارائه شده به تحصیلات تکمیلی دانشگاه به عنوان بخشی

از فعالیت‌های تحصیلی لازم برای اخذ درجه کارشناسی ارشد

در رشته

مهندسی کامپیوتر-هوش مصنوعی

از دانشگاه شیراز

شیراز

جمهوری اسلامی ایران

ارزیابی شده توسط کمیته پایان نامه با درجه: **عالی**

دکتر اقبال منصوری،استادیار بخش مهندسی و علوم کامپیوتر (رئیس کمیته) .....

دکتر فرشاد تاجری پور، استادیار بخش مهندسی و علوم کامپیوتر .....

دکتر محمدهادی صدرالدینی، دانشیار بخش مهندسی و علوم کامپیوتر .....

مرداد ۱۳۹۱

سپاسگزاری

برخود لازم می‌دانم از زحمات استاد ارجمند جناب آقای دکتر اقبال منصوری که همواره در تمام مدت انجام پایان‌نامه مرا از راهنمایی‌ها و مساعدت‌های بی‌دریغ‌شان بهره‌مند نمودند تشکر کنم. همچنین از استادان مشاور خود جناب آقای دکتر فرشاد تاجری‌پور و جناب آقای دکتر محمدهادی صدرالدینی که با نظرات و راهنمایی‌های مفیدشان مرا در پیشبرد این پایان‌نامه یاری نموده‌اند کمال تشکر را دارم.

در انتها از تمام عزیزانی که مرا در انجام این پروژه تحقیقاتی یاری یاری نمودند کمال تشکر و قدردانی را دارم.

چکیده

ارائه‌ی مدل آماری برای خوشه‌بندی متون

به کوشش
حمید محمودی

در این پایان‌نامه سعی در ارائه‌ی یک مدل آماری برای خوشه‌بندی متون داشته‌ایم. هر خوشه به منزله‌ی جزئی از یک مدل ترکیبی در نظر گرفته می‌شود که شامل پارامترهای عدد اولویت جزء، بردار میانگین و ماتریس کواریانس جزء می‌باشد. هدف از ارائه‌ی یک مدل آماری، در نظر گرفتن پخش‌های متفاوت برای مجموعه داده‌هایی که لزوماً پخشش آنها کروی نیست، می‌باشد. الگوریتمهای خوشه‌بندی مانند *K-Means* و مشتقات آن که با یک پارامتر-که معمولاً مراکز خوشه‌هاست- کار می‌کنند، سعی دارند خوشه‌هایی با پخشش کروی را ایجاد کنند که این در مورد همه‌ی مجموعه داده‌های دنیای واقعی صدق نمی‌کند. هدف دیگر این پایان‌نامه، به مقدار دهی اولیه‌ی پارامترهای مدل بر می‌گردد. بسیاری از کارهای انجام شده در زمینه‌ی خوشه‌بندی بدون نظارت متون، ساخت چندین مدل مختلف با مقداردهی‌های اولیه‌ی تصادفی بوده است و نهایتاً مدل برتر را بر اساس یک معیار خاص انتخاب می‌کردند. از آنجائیکه مقداردهی تصادفی در همه‌ی اجراها نتایج قابل اعتماد و منظمی ارائه نمی‌دهد، ما با ارائه‌ی یک رویه‌ی چند مرحله‌ای و بدون کمک از هرگونه ناظر خارجی و با استفاده از الگوریتمهای سلسله مراتبی که پیچیدگی محاسباتی آن‌را با انتخاب یک مجموعه کوچک از نمونه‌های برتر و همچنین کاهش فضای ابعاد، کاهش دادیم موفق شدیم برای هر خوشه بهترین متون مربوطه را بیابیم و بوسیله‌ی آنها پارامترهای مذکور را مقداردهی کنیم. نتایج آزمایشگاهی و نمودارهای مقایسه‌ای به صورت کاملاً واضح نشان می‌دهند که روش‌های پیشنهاد شده در این پایان‌نامه از عملکرد بالاتری نسبت به روشهای ارائه شده‌ی مشابه داشته‌اند.

واژگان کلیدی: خوشه‌بندی بدون نظارت متون، روش بیشینه کردن امید، خوشه‌بندی سلسله مراتبی.

فهرست مطالب

عنوان **صفحه**

عنوان **صفحه**

فصل دوم:انتخاب ویژگیها با استفاده از معیارهای آماری

فصل سوم: روشهای مختلف ارزیابی خوشه‌بندی

فصل چهارم: ارائه راه حل پیشنهادی برای مسئله‌ی خوشه‌بندی متون

فصل پنجم: نتایج آزمایشات

عنوان **صفحه**

فهرست جداول

عنوان **صفحه**

۸۳

فهرست اشکال

عنوان **صفحه**

عنوان **صفحه**

فصل اول

مقدمه

۱-۱- سیستم‌های بازیابی اطلاعات

با پیشرفت رایانه ها هم از جهت سخت افزار و هم از جهت نرم افزار خصوصا در دو دهه اخیر، تکنیکهای مدیریت داده ها و اطلاعات بیش از پیش ماشینی و خودکار شد. سیستم های مدیریت پایگاه داده مخصوص ذخیره و پردازش داده های ساختاریافته بود که مکانیزمهای جستجو برای بازیابی و پردازش اینگونه داده ها سرراست و آسان بود. به علت ماهیت ویژه و غیرساختاریافته داده های متنی و پیچیدگی مکانیزمهای جستجو و پردازش آنها، نوع خاصی از تکنولوژی پردازش داده تحت عنوان سیستمهای بازیابی متون بوجود آمد که تفسیر کلی آن عبارت بود ازتکنیکهای ذخیره و دستکاری حجمی از داده های متنی برای اهداف مختلف بازیابی اطلاعات.

اکثر سیستمهای بازیابی اطلاعات سنتی اطلاعات مربوط به فهرست داده ها را ذخیره و پردازش می کردند و اساس مکانیزم جستجوی آنها بر روی کلیدهای فهرست برای بازیابی منبع و آدرس محل ذخیره اطلاعات اصلی استوار بود. امروزه سیستم های خودکار بازیابی اطلاعات قادر هستند که هر دو موجودیت فهرست داده ها و کل داده های اصلی را پردازش و اداره کنند.

بازیابی اطلاعات به مطالعه تکنیک های جستجو و بازیابی اطلاعات برای داده های غیرساختیافته متون، اطلاعاتی که درون متون هستند و ابرداده هایی در مورد متون اطلاق می-شود[۱,۲]. واژه داده غیرساختیافته معمولابه داده هایی اشاره می کند که ساختار منظم و منسجم و واضحی نداشته باشند. در مقابل داده های ساختیافته مانند پایگاه داده رابطه ای که شامل جداولی از رکوردهای مرتب هستند، دارای ساختاری منظم و ارتباطات معنایی و قابل فهم هستند. هرچند داده های غیر ساختیافته ای مانند تصاویر گرافیکی، فایل های صوتی، فایل های صوتی تصویری و... نیز وجود دارند، اما کاوش های بازیابی اطلاعات روی بازیابی متون زبان طبیعی متمرکز است. به علت اینکه بسیاری از متون شامل سرصفحه های ساختار یافته -که شامل ابرداده هایی از قبیل عنوان، توضیحات کوتاه در مورد متن، موضوع متن،نویسنده، ناشر، تاریخ ها و... - و بدنه غیر ساختیافته هستند، گاهی از متون به عنوان داده های نیمه ساختیافته نیز یاد می شود.

۲-۱- خوشه بندی متون

خوشه بندی یکی از پرکاربردترین کارها در حوزه داده کاوی است، که به منظورتعیین توزیع-های نمونه‌هایی که از آن توزیع تبعیت می کنند و همچنین کشف گروههایی که تا حد ممکن از هم متفاوت و تا حد ممکن نمونه های درون گروهی مشابه باشند استفاده می‌شود[۳,۴].

خوشه بندی معمولاً به دسته بندی مجموعه ای از متون در خوشه های مختلف بر اساس میزان شباهتی که بین آنها وجود دارد اطلاق می شود. به طور ایده آل تا جایی که ممکن است خوشه ها از هم جدا و شباهت درون خوشه ای حداکثر می باشد. اما معمولا همپوشانی خوشه-ها، وجود داده های نامرتبط با داده های دیگر و وجود ویژگیهای نویزی که اثر منفی روی جداکنندگی داده ها از هم دارند، سه چالش اصلی در خوشه بندی متون و دیگر حوزه های بازیابی اطلاعات هستند.

۳-۱- اهمیت خوشه‌بندی متون

در ابتدا خوشه‌بندی متون برای بهبود دقت و یادآوری یک سیستم بازیابی اطلاعات و یک راه موثر برای یافتن نزدیکترین همسایه متون مورد استفاده قرار می‌گرفت. امروزه از خوشه‌بندی برای تنظیم و سازماندهی نتایج موتورهای جستجوگر برای پاسخ به پرس‌وجوهای کاربران مورد استفاده واقع می‌شود. یکی دیگر از کاربردهای مهم خوشه بندی طبقه‌بندی خودکار مجموعه-های بزرگی از متون به صورت سلسله مراتبی می‌باشد.

یکی از مهمترین انگیزه‌های استفاده از خوشه‌بندی، تعیین وآشکارکردن ساختار ذاتی و پنهان یک مجموعه داده است. کاربران انسانی به علت تفاوت در سلیقه و طرزتفکرات مختلف از کشف ساختار ذاتی و درونی مجموعه داده‌ای بزرگ متون ناتوان هستند. به عنوان مثال فرض کنید مجموعه‌ای از متون مربوط به بازیهای المپیک را در اختیار داریم. یک متخصص انسانی برای خوشه‌بندی ممکن است متون را بر اساس نام بازیهای مختلف (فوتبال، هاکی، شنا،...)، متخصص دیگری بر اساس نام کشورهای شرکت‌کننده (انگلیس، استرالیا،...) و دیگری بر اساس نوع بازیها (زمستانی یا غیر زمستانی، با توپ یا بدون توپ، قدرتی یا سرعتی،...) خوشه‌بندی کنندو در نهایت چندین خوشه‌بندی متفاوت داشته باشیم و به علت اینکه هیچ یک نتوانسته‌اند بر اساس ساختار درونی خوشه‌بندی کنند نمی‌توانیم هیچ یک را بر دیگری برتری دهیم.

یکی دیگر از عوامل مهم استفاده از خوشه‌بندی، گسترش روزافزون اطلاعات خصوصا در حوزه‌ی وب و اینترنت می باشد که خوشه‌بندی و طبقه‌بندی اطلاعات را جزء جداناپذیر

موتورهای جستجوگر مدرن امروزی کرده است. موتورهای جستجوگر امروزی برای یافتن بهترین نتایج مربوط به یک پرس‌وجو به جای جستجو در میان میلیونها صفحه وب که از نقطه نظر زمانی برای یک کاربر معمولی قابل قبول نیست، کوشش می کند تا در میان خوشه های مختلف که تعدادشان به مراتب کمتر از تعداد صفحات وب هستند جستجو کنند.

کاربردهای خوشه‌بندی

کاهش داده[1]: آنالیز خوشه‌بندی می تواند در فشرده‌سازی و یکجا کردن اطلاعات موجود در داده‌ها به کار رود. در بسیاری از مواقع مجموعه داده‌ای در اختیار، آنقدر وسیع وگسترده است که نیاز به طبقه بندی و فشردگی تا حد ممکن داده ها دارد.

تولید فرضیه[2]: این مورد برای استنتاج یا کشف فرضیه‌هایی که پنهان در داده‌ها هستندکاربرد دارد. برای مثال پایگاه داده‌ی خرده‌فروشی را درنظر بگیرید که دو گروه از مشتریها را بر اساس سن افراد و زمان خرید دارد. می‌توانیم به عنوان مثال یک گمانه کلی در این مورد داشته باشیم که افراد جوان در عصر یا افراد مسن در صبح به خرید می روند.

پیش‌بینی بر اساس خصوصیات خوشه‌ها: خوشه‌بندی مجموعه‌ای از داده‌ها را به گروههای متفاوتی بر اساس ویژگیهایشان تقسیم می‌کند. داده‌هایی که پس از خوشه‌بندی به دست می‌آید و در فرایند خوشه‌بندی دخالت نداشته‌اند، می‌توانند بر اساس شباهتی که با نزدیکترین خوشه را دارند به آن خوشه تخصیص یابند. به عنوان مثال فرض کنید داده‌هایی در مورد بیمارانی داریم که با یک الگوریتم خوشه‌بندی بر اساس واکنشی که به داروهای مختلفی از خود بروز می‌دهند، خوشه‌بندی می شوند. پس برای بیمار جدیدی، و بر اساس نزدیکترین خوشه‌ای که به آن تخصیص می‌یابد می‌توان معالجه‌ای که روی بیماران آن خوشه انجام می-شود، برای او نیز تجویز کرد.

تجارت: در تجارت، خوشه‌بندی می‌تواند به بازاریها در کشف گروههای معنی داری از مشتری‌هایشان بر اساس خریدهایی که می‌کنند کمک کند.

زیست‌شناسی: در زیست‌شناسی می‌توان از خوشه‌بندی در طبقه‌بندی و دسته‌بندی ژنها بر اساس عاملیت[3] آنها در ساختارهای موروثی در جمعیت‌ها استفاده کرد.

دیگر کاربردهای خوشه‌بندی می‌توان به وب‌کاوی، متن‌کاوی، پردازش‌تصویر، تصاویر ماهواره‌ای، تجهیزات پزشکی، سیستم اطلاعاتی جغرافیایی[4] (*GIS*) اشاره کرد.

---

[1] Data reduction
[2] Hypothesis Generation
[3] Functionality
[4] Geographical Information System

۴-۱- انگیزه‌ی انجام این پایان‌نامه

بسیاری از الگوریتمهای ارائه شده در حوزه‌ی خوشه‌بندی متون اهمیتی به پخشش داده ها نمی دهند. الگوریتمهای مشهوری مانند *K-Means* و مشتقات آن مانند *K-Means* کروی، *K-Means* فازی،... پخشش داده‌ها را به صورت کروی یا ابرکروی در نظر می‌گیرند. علت این موضوع به متریک‌های معیار فاصله یا عدم شباهت اقلیدسی یا کسینوسی برمی گردد. ما در این پایان‌نامه قصد داریم تا با ارائه یک مدل احتمالی گوسی از توزیع خوشه‌ها در فضای بزرگ متن-عبارت، ضمن در نظر گرفتن پخشش داده‌ها در ابعاد مختلف بتوانیم مجموعه‌هایی با توزیع غیرکروی را مدل کنیم.

یکی دیگر از اهداف ما، توجه به مقداردهی اولیه برای پارامترهای خوشه‌ها قبل از عمل خوشه‌بندی است. مقداردهی اولیه در *K-Means* شامل تعیین تنها پارامتر، یعنی مراکز اولیه خوشه‌های موجود در مجموعه است. انتخاب تصادفی پارامترها منجر به غیرقابل اعتماد و نوسانی بودن نتایج در اجراهای مختلف الگوریتم می شود.

ما در این پایان‌نامه یک روش جدید انتخاب ویژگی را ارائه نموده ایم که ضمن اینکه به میزان زیادی ابعاد مسئله را کاهش می‌دهد، می تواند بهترین و ارزشمندترین ویژگی ها (ابعاد) را نیز انتخاب کنید.

ساختار این پایان‌نامه بر این اساس است ؛ در ادامه این فصل، به بررسی مفاهیم اولیه پرداخته شده می‌پردازیم. در فصل دوم به پیاده سازی و بررسی روش های موجود در این حوزه پرداخته می‌شود و نقاط قوت و ضعف هر روش را بیان می کنیم. در فصل سوم به روش‌های مختلف ارزیابی و اعتبارسنجی نتایج خوشه بندی ارائه شده توسط محققین مختلف خواهیم پرداخت. ارائه روش پیشنهادی و امکان بهبود کارایی آن نسبت به کارهای انجام شده همراه با جزئیات پیاده‌سازی ، در فصل چهارم مطرح می‌شود. در فصل پنجم نتایج تجربی و نتایج آزمایشگاهی را که برآمد کار این پایان‌نامه است، با تفصیل مورد ارزیابی و تحلیل قرار داده می-شود. در پایان، نتیجه‌گیری کلی از کار انجام شده را توضیح می‌دهیم و چندین پیشنهاد را برای کارهای آینده معرفی می‌کنیم.

۵-۱- پیش‌پردازش متون

قبل از اعمال پردازشهایی از قبیل انتخاب ویژگی و عمل خوشه بندی نیاز است که پردازشهای مقدماتی روی متون صورت گیرد تا تفسیری قابل فهم برای ماشین تولید شود. خروجی این مرحله اهمیت زیادی برای افزایش صحت نتایج و کاهش مرتبه زمانی دارد.

۱-۵-۱- نشانه‌گذاری[1]

در اولین مرحله پیش پردازش، می بایست متن را به صورت عباراتی جدا از هم تقسیم کرد. هر عبارت ممکن است فقط یک کلمه یا چندین کلمه که یک اصطلاح را تشکیل می دهند باشد، ولی ما همه عبارات را تک کلمه ای در نظر می گیریم و فرض می کنیم عبارت ها از هم مستقل هستند و ترتیب، اثری در آنها ندارد.

۲-۵-۱- حذف کلمات بی‌اثر[2]

تعداد این کلمات در زبان انگلیسی معمولاً کمتر از ۶۰۰ عدد است و تعداد آنها برای لیست های مختلف متغیر است. در این رساله یکی از لیست های استاندارد که شامل ۵۷۱ کلمه بی اثر[3](اضافه) است استفاده شده است. معمولا هر متنی تعدادی کلمات موجود در این لیست را داراست که باید حذف شوند. این کلمات اثر جداکنندگی ندارند و اغلب به عنوان نویز در نظر گرفته می شوند. مثالی از این کلمات عبارتند از (“a”, “the” , “if”,…).

۳-۵-۱- ریشه‌یابی[4]

در زبان انگلیسی، معمولا کلمات هم خانواده در چندین حرف در انتهای کلمات باهم اختلاف دارند. همه آنها یک ریشه دارند و به علت اینکه مطابقت دو کلمه به صورت مطابقت نظیر به نظیر تک تک حروف انجام می شوند، لازم است تا کلمات هم خانواده به یک شکل ریشه‌شان کاهش یابند. رویه ی استانداردی که برای حذف پسوند و پیشوند در این رساله استفاده شده است، Porter-Stemmer نام دارد[۶]. این مرحله پیش پردازش باعث می شود کلمات به تعداد قابل توجهی کاهش یابد و صحت نتایج را نیز افزایش دهد.

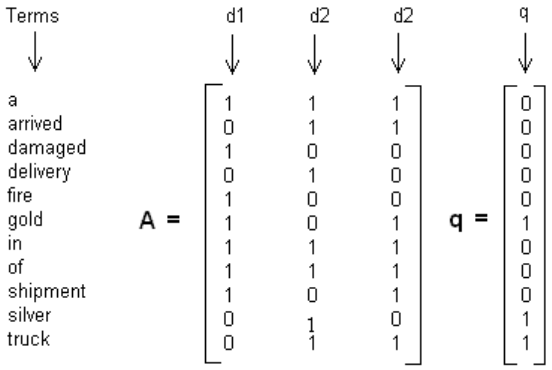
[1] Tokenization
[2] Stop-Word
[3] ftp://cs.cornell.edu/pub/smart/english.stop
[4] Stemming

۶-۱- مدل فضابرداری

مدل فضابرداری یک ابزار استانداردی در طول سه دهه اخیر برای سیستمهای بازیابی اطلاعات شده است [۱,۲,۵]. یکی از محاسن این روش امکان رتبه بندی مرتبط[1] برای متون با پرس وجوی کاربران است. سیستم بازیابی اطلاعات فضا برداری[۳,۷] متون را به صورت بردارهایی از عبارات در نظر می گیرد. مجموعه متون شامل یک ماتریس $m \times n$ از m ترم و n متن است. ماتریس مذکور را ماتریس عبارت-متن می نامند که هر ستون نماینده یک متن از مجموعه و هر سطر نماینده یک عبارت است.

۱-۶-۱- مدل دودویی

در مدل برداری دودویی، هر متن به صورت بردار منطقی از تمام عبارات موجود در مجموعه است و مقدار هر ورودی یک است، اگر عبارت متناظر با آن ورودی در متن وجود داشته باشد و در غیر این صورت مقدار صفر خواهد داشت.



شکل ۱: مدل برداری بولیین از سه متن و یک پرس وجوی کاربر

در شکل ۱ مدل دودویی از سه بردار متن d_1وd_2وd_3 اندیس شده در ماتریس A ، و پرس وجوی q اندیس شده در بردار q ، را مشاهده می کنید. هر چهار بردار در فضای ۱۱ بعدی از ۱۱ عبارت اندیس شده اند. عیب این روش در این است که تعداد دفعات تکرار عبارات را در نظر نمی گیرد و فقط بود و نبود عبارت مهم است. معمولا برای رتبه بندی ارتباط پرس و جوی کاربر با متون از ضرب نقطه ای بین پرس و جو و هر یک از متون استفاده می شود.

[1] Relevance Ranking

۲-۶-۱- مدل وزن‌دهی عبارت

در این مدل، متون را به صورت مخزنی از عبارات[1] در نظر می گیرند. به این معنا که ترتیب عبارات در این روش اهمیتی ندارد. بر خلاف مدل بولیین، این مدل تعداد دفعات تکرار هر عبارت در متن را در نظر می گیرد و هرسلول ورودی از ماتریس، دفعات تکرار عبارت اندیس شده سطر متناظر سلول در متن اندیس شده ستون متناظر را نشان می دهد. یک متن d_i با بردار $x_i = \{f_1, f_2, \dots, f_m\}$ که یک فضای m بعدی از ترمها است نمایش داده می شود که در آن f_t دفعات تکرار عبارت t ام در متن iام است. یکی از مهمترین محاسن این نوع مدل نمایش، ارائه ساختاری است که عملیات جبری را آسان و قابل درک می کند. وزن هر عبارت در هر متن تابعی است از دفعات تکرار عبارت در متن *(tf)*[2] و تعداد متونی از مجموعه *(df)*[3] که عبارت در آن حداقل یک بار دیده شده است. در ادامه به تعدادی از مشهورترین این توابع اشاره می کنیم.

$$\begin{bmatrix} & D_1 & D_2 & \dots & D_n \\ T_1 & w_{11} & w_{12} & \dots & w_{1n} \\ T_2 & w_{21} & w_{22} & \dots & w_{2n} \\ : & \dots & \dots & \dots & \dots \\ T_m & w_{m1} & w_{m2} & \dots & w_{mn} \end{bmatrix}$$

شکل ۲: مدل وزن دهی عبارت

در شکل ۲ نمایی کلی از مدل وزن دهی عبارت برای n متن و m ترم را نشان می دهد. وزن w معمولا توسط یکی از روش های زیر محاسبه می شود.

وزن‌دهی *Tf-Idf*

تکرار عبارت ها *(tf)* وعکس تکرار متون *(df)*، که اولی باعث افزایش اثر تکرار عبارت و دومی کاهش اثر عبارت هایی است که در اسناد زیادی تکرار شده اند. فرض بر این است عبارت هایی که در متون زیادی ظاهر شده اند خاصیت جداکنندگی بالایی ندارند و عمدتاً باعث کاهش صحت نتایج می شوند. این فرضیه در مورد مجموعه داده های متوازن که تعداد نمونه های همه خوشه ها تقریبا نزدیک به هم است خوب عمل می کند، اما برای مجموعه داده های نامتوازن ممکن است وزن برخی عبارت های مهم با قدرت جداکنندگی نسبتا بالا را در خوشه های بزرگ به علت کوچک بودن عکس تکرار متون را کاهش دهد.

---

[1] Bag of Words
[2] Term Frequency
[3] Document Frequency