فصل اول

مقدمه

خوشه بندی یک تکنیک قدرتمند در تجزیه و تحلیل داده ها است که اشیاء مشابه را بر اساس ویژگی های آنها گروه بندی می کند. در خوشه بندی اسناد فارسی، هدف گروه بندی اسناد فارسی زبان بر اساس محتوای آنهاست. در اسناد فارسی روش های مختلفی برای خوشه بندی وجود دارد که هر کدام مزایا و معایبی دارند. در این گزارش چندین مورد از این روش ها را بررسی می کنیم و بهترین روش موجود در حال حاضر را ارائه می دهیم.

اما قبل از آشنایی با روش های خوشه بندی ( (clusteringباید با مرحله پیش پردازش متن آشنا شویم پیش پردازش داده های متن فارسی به دلیل ساختار منحصر به فرد آن با سایر زبان ها متفاوت است. به طور کلی، مرحله پیش پردازش شامل کارهایی مانند توکن سازی، ریشه یابی و حذف کلمه توقف است که در ادامه مختصر به هرکدام میپردازیم. مرحله پیش پردازش را ما با بهره گیری از ابزار تجزیه و تحلیل کتابخانه hazm در پایتون استفاده میشود.

در اینجا ابتدا به مراحل کلی پردازش زبان طبیعی میپردازیم که هر محققی پردازش زبان های طبیعی (NLP) را به روش هایی تقسیم میکند ولی اصل آن به 3 دسته تقسیم میشود.1.پیش پردازش 2.پردازش 3.ارزیابی

ارزیابی

پردازش

پیش پردازش

TF\_IDF

Clustering

ارزیابی روش های دو مرحله قبل

Normalization

Tokenization

Stemming

stop words

چالش های روش های خوشه بندی به غیر از زبان انگلیسی

به دلیل اینکه زبان های مختلف میتواند ساختارهای vocabularyوmorphologyو grammarو syntactic متفاوتی داشته باشد ممکن است نیاز به الگوریتم های خوشه بندی موجود یا توسعه الگوریتم های جدید باشد.

چالش های خوشه بندی متن های فارسی

برخی از چالش های خوشه بندی زبان فارسی عبارتند از:

Morphological complexity(ریخت شناسی پیچیده): زبان فارسی دارای Morphological complexity است، با مجموعه ای غنی از پسوندها و پسوندها که معنا و کارکرد کلمات را تغییر می دهد. این می تواند تشخیص کلمات مشابه و گروه بندی آنها را با هم دشوار کند.

Ambiguity(ابهام): زبان فارسی دارای کلمات متعدد با معانی متعدد است و ممکن یک کلمه چندین معنی متفاوت داشته باشد. این امر می تواند خوشه بندی دقیق اسناد بر اساس محتوای آنها را چالش بزرگی دچار کند.

Lack of standardized data (نبود داده های استاندارد شده): کمبود مجموعه داده های استاندارد و بزرگ برای خوشه بندی زبان فارسی وجود دارد که می تواند آموزش و ارزیابی موثر الگوریتم های خوشه بندی را دشوار کند.

Limited availability of tools and resources (دسترسی محدود به ابزارها و منابع): ابزارها و منابع محدودی برای پردازش زبان فارسی در مقایسه با سایر زبان ها وجود دارد که می تواند پیش پردازش و تجزیه و تحلیل متن را به طور موثر چالش برانگیز کند.

Dialectal variation (تنوع گویش): زبان فارسی دارای چندین گویش و تنوع منطقه‌ای است که می‌تواند بر نتایج خوشه‌بندی تأثیر بگذارد و اعمال الگوریتم‌های خوشه‌بندی را در مناطق و زمینه‌های مختلف چالش‌برانگیز کند.

نتیجه

شایان ذکر است که اهمیت خوشه بندی اسناد می تواند تا حد زیادی بر روش خوشه بندی مورد استفاده و ساختار حاصل تأثیر بگذارد. به عنوان مثال، اگر اسنادی که در حال خوشه بندی هستند بسیار مهم هستند و حاوی اطلاعات حیاتی هستند، ممکن است به یک روش خوشه بندی دقیق تر و پیچیده تر نیاز باشد تا اطمینان حاصل شود که خوشه های حاصل به خوبی تعریف شده و به راحتی قابل تفسیر هستند. از سوی دیگر، اگر اسناد اهمیت کمتری داشته باشند یا حاوی اطلاعات ساختاری کمتری باشند، ممکن است روش خوشه بندی ساده تری کافی باشد. بنابراین، هنگام انتخاب روش خوشه‌بندی، ماهیت و اهمیت اسناد خوشه‌بندی شده در نظر گرفته می‌شود تا اطمینان حاصل شود که ساختار حاصل هم دقیق و هم قابل درک است.

Application of Self-Organizing Maps in Text Clustering: A Review

کاهش ابعاد یک مرحله ضروری در خوشه بندی متن است. چندین تکنیک وجود دارد

برای کاهش ابعاد بردار ویژگی با ابعاد بالا. PCA (جزء اصلی

روش تجزیه و تحلیل) یکی از تکنیک های کاهش ابعاد پرکاربرد است. با توجه به n ×

ماتریس سند-مرتب m، بردارهای ویژه k PCA با مرتبه m × m cova-

ماتریس ریانس برای کاهش ابعاد فضای کلمه استفاده می شود و در نهایت منجر به

یک بعد فضایی k-term، که بسیار کوچکتر از m است.

ارزیابی اهمیت کلمه یک مقاله علمی را به عنوان مثال در نظر بگیرید، نشان داده شده است که

حدود 65٪ تا 90٪ کلمات کلیدی علامت گذاری شده توسط نویسنده را می توان در محتوای اصلی در اصل پیدا کرد.

کاغذ اینال[18]. به این معنی که با ارزیابی اهمیت می توان کلمات کلیدی را استخراج کرد

از اسناد برای نمایش محتوای اصلی. اساسا استخراج کلمات کلیدی قابل مشاهده است

به عنوان یک مشکل یادگیری ماشین نظارت شده؛ این ایده برای اولین بار توسط Turney [19] ارائه شد.

Turney همچنین یک مطالعه مقایسه ای بر اساس الگوریتم های ژنتیک و الگوریتم استخراج کلمات کلیدی مبتنی بر درخت تصمیم انجام داد. عواملی که می توانند کلمه اهمیت را در

شامل فراوانی کلمه، مکان کلمه (عنوان، عنوان و غیره) می شود. تحقیقات زیادی نشان داد

که کلمات با بسامد بالا کلمات مهمتر هستند. برخی از کلمات کلیدی معمولی استخراج شده

سیستم عملیات در جدول 1 فهرست شده است.

Quantization Error (abbreviated as MQE) فرمولش در فایل موجود است.

شبکه SOM (نقشه خود سازماندهی) چندین ویژگی مهم را نشان می دهد که به اثربخشی آن به عنوان طبقه بندی کننده الگو کمک می کند. اولاً، از انتظارات ریاضی همه اسناد درون یک خوشه به عنوان مرکز خوشه استفاده می کند. این بدان معنی است که مرکز خوشه نشان دهنده میانگین یا نمایندگی مرکزی اسناد در آن خوشه است. این ویژگی به شبکه SOM اجازه می دهد تا ویژگی های کلی و ویژگی های به اشتراک گذاشته شده توسط اسناد درون خوشه را ثبت کند.

ثانیاً، شبکه SOM داده های ورودی را به صورت خوشه ای سازماندهی می کند و در عین حال نظم توپولوژیکی را حفظ می کند. این بدان معنی است که اسناد یا الگوهای مشابه با هم گروه بندی می شوند و موقعیت ها و روابط نسبی بین خوشه ها حفظ می شوند. این ترتیب توپولوژیکی امکان بازیابی و مقایسه موثر الگوهای درون شبکه را فراهم می کند.

هنگامی که یک شبکه SOM به طور کامل آموزش داده می شود، می توان آن را به عنوان یک طبقه بندی کننده الگوی قدرتمند مشاهده کرد. با وارد کردن یک سند یا الگو، نورون‌های لایه خروجی که کلاس الگوی خاص را نشان می‌دهند، بالاترین پاسخ را نشان می‌دهند. به عبارت دیگر، نورون‌های مربوط به خوشه‌ای که به بهترین شکل با الگوی ورودی مطابقت دارد، به شدت فعال می‌شوند. این پاسخ نشانه روشنی از کلاس الگوی که سند ورودی به آن تعلق دارد را ارائه می دهد.

به طور کلی، توانایی شبکه SOM برای محاسبه مراکز خوشه‌ای بر اساس انتظارات ریاضی و حفظ نظم توپولوژیکی در میان خوشه‌ها، آن را قادر می‌سازد تا به عنوان یک طبقه‌بندی کننده الگوی مؤثر عمل کند. می‌تواند با تعیین نورون‌هایی در لایه خروجی که بالاترین پاسخ را نشان می‌دهند، کلاس الگوی اسناد ورودی را با دقت شناسایی کند، بنابراین قابلیت‌های طبقه‌بندی آن را نشان می‌دهد. The running process of the SOM network can be divided into two stages: training and map‐ ping. In the training phase, the samples were input randomly. For a particular input pattern, there will be a winning node in the output layer, which produces the greatest response. At the beginning of the training phase, which node in the output layer will generate the maxi‐ mum response is uncertain. When the category of the input pattern is changed, the winning node of the two-dimensional plane will also change. Due to the lateral mutual excitatory ef‐ fects, Nodes around the winning node have a greater response, so all the nodes of the win‐ ning node and its neighborhood will both perform different levels of adjustment

SOM وزن گره های لایه خروجی را با تعداد زیادی نمونه آموزشی تنظیم می کند.

و در نهایت هر گره در لایه خروجی به یک کلاس الگوی خاص حساس است. وقتی که

ویژگی های کلاس از دو خوشه نزدیک هستند، گره ها به نمایندگی از این دو خوشه

همچنین در موقعیت نزدیک هستند.

کل بخش **2.3. SOM And Its Application For Text Clustering روش کلی این کلاستر است.**

**مقایسه SOM با سایر روش‌های خوشه‌بندی متن: سلسه مراتبی و ذکر فرمول** intra-cluster entropy

K means ،

**به طور کلی ثابت شده است که SOM مناسب ترین روش خوشه بندی اسناد است. می تواند**

**نگاشت اسناد روی نمودار دو بعدی برای نشان دادن رابطه بین تفاوت**

**اسناد و مدارک SOM می تواند متن را به صورت تصویری تر و بصری بهتر به تصویر بکشد. با کیفیت بالا**

**فضای ذهنی را می توان به فضای دوبعدی تبدیل کرد و شباهت آن**

**بین داده های ورودی در فضای چند بعدی به خوبی در فضای دو بعدی حفظ می شود.**

**فضای گسسته ذهنی، درجه شباهت بین فضایی با ابعاد بالا**

**داده ها همچنین می توانند به مجاورت مکان فضای نمایش تبدیل شوند که**

**می تواند نظم توپولوژیکی را حفظ کند. SOM همچنین دارای مزایای زیر است:** 1) noise im‐ munity; 2) visualization; 3) parallel processing

Evaluating Self-Organizing Map Quality Measures as Convergence Criteria

Quantization error برای تعداد خوشه ها

**نقشه خودسازماندهی**

**SOM یک شبکه عصبی مصنوعی است که توسط Teuvo Kohonen [1] توسعه یافته است. این**

**الگوریتم از یک رویه بدون نظارت 1 و تکراری برای مدلسازی فضای ورودی استفاده می کند**

**یک شبکه ثابت از گره ها. یک نسخه سطح بالا از الگوریتم در شکل 1 نشان داده شده است.**

**الگوریتم با شبکه ای از نورون ها (یا نقشه) مقداردهی اولیه می شود. هر نورون دارای یک**

**بردار وزن با همان ابعاد فضای ورودی. حلقه while بیرونی**

**تکرارهای آموزشی را تشکیل می دهد که معیارهای توقف انجام نشده به آن اشاره دارد**

**تعداد تکرارهایی که الگوریتم باید اجرا کند. حلقه for داخلی به روز رسانی می شود**

**نقشه با استفاده از تمام نمونه ها (مشاهدات) در داده های آموزشی. چند وجود دارد**

**مراحل اولیه برای هر یک از نقاط داده در ورودی. ابتدا نورون (گره) در**

**نقشه ای که نزدیک ترین نقطه به نقطه است که به عنوان \****بهترین-واحد تطبیق (BMU) شناخته می شود.**

**یافت. سپس نورون به روز می شود تا به نقطه نزدیکتر شود. در نهایت، نورون ها**

**\"نزدیک" BMU (همسایگی آن) نیز به روز می شود تا به نقطه نزدیک تر باشد.**