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.