

SID



بلاگ مرکز اطلاعات علمی



کارگاه‌های آموزشی



سرویس ترجمه تخصصی



فیلم‌های آموزشی

کارگاه‌ها و فیلم‌های آموزشی مرکز اطلاعات علمی

آشنایی با پایگاه‌های اطلاعات علمی بین‌المللی و ترندهای جستجو بین‌المللی و ترندهای جستجو

کاربرد نرم افزار SPSS در پژوهش

بروبوزال نویسی (علوم انسانی)

کاربرد نرم‌افزار End Note در استناددهی مقالات و متون علمی

صدور گواهینامه نمایه مقالات نویسندگان در SID

Constructing and Analyzing Movie Similarity Graph Based on Topical Analysis of Movie Subtitles

Dadfar Momeni ^{*1}, Hossein Rahmani², Mohammad Nazari³

¹ School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
dadfar.momeni@gmail.com

² Assistant Professor, School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
h_rahmani@iust.ac.ir

³ Master of Software Engineering, Iran University of Science and Technology, Tehran, Iran
nazari.mo1997@gmail.com

Abstract

Nowadays, considering the huge amount of DATA, to search through them, we ought to use methods for analyzing the DATA according to our needs. This challenge also exists in the entertainment and cinema industry to find movies and TV shows with the same topic aiming to recommend and minimize the search space for the audience. Therefore, methods are needed to efficiently recognize the movies with the same topic and present them to the users. Most of the existing services lean on user-based information, and usually, not on the original content of the movies. These services use DATA such as user ratings and comments or features like actors, directors, and the movie genre or a combination of both.

In this paper, we use low-level features of the movie subtitles, extracted using LDA, for thematic analysis of textual contents of the movies (subtitles). To do so, using the extracted features and Cosine similarity measure, we construct the similarity graph of movies. In this graph, each node represents a movie and each edge indicates the similarity between them. In the following, using clustering methods on movies graphs we were able to achieve a noticeable Thematic correlation between the movies.

Keywords: Data Mining, Topic Extraction, Graph Analysis, Movie, Subtitle

ساخت و تحلیل گراف شباهت فیلم‌ها براساس تحلیل موضوعی زیرنویس‌ها

دادفر مؤمنی*^۱، حسین رحمانی^۲، محمد نظری^۳

^۱ کارشناسی مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران
dadfar.momeni@gmail.com

^۲ استادیار دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران
h_rahmani@iust.ac.ir

^۳ کارشناسی ارشد مهندسی نرم افزار، دانشگاه علم و صنعت ایران، تهران
nazari.mo1997@gmail.com

چکیده

امروزه با توجه به حجم عظیم داده‌ها، برای جست‌وجو میان آن‌ها، ناگزیریم از روش‌هایی بهره بگیریم که بتوانیم اطلاعات را طبق نیاز خود پالایش کنیم. این چالش در صنعت سینما و سرگرمی نیز به منظور یافتن فیلم‌ها و سریال‌هایی با موضوعات مشابه و مرتبط در جهت پیشنهاد و کوچک کردن فضای جست‌وجو برای مخاطبان وجود دارد. بنابراین روش‌هایی لازم است که بتوانند به نحوی کارآمد فیلم‌های مرتبط و دارای موضوعات مشابه را تشخیص دهند و در اختیار کاربران بگذارند. اکثر سرویس‌های موجود در این زمینه، بر اطلاعات بدست‌آمده از کاربران تکیه می‌کنند و معمولاً محتوای اصلی فیلم، توسط آن‌ها به کار گرفته نمی‌شود. این سرویس‌ها از اطلاعاتی مانند سلیقه و نظرات کاربران، یا ویژگی‌هایی نظیر بازیگران، کارگردان و ژانر فیلم، یا ترکیبی از این دو استفاده می‌کنند. در این مقاله با استفاده از ویژگی‌های سطح پایین استخراج‌شده از زیرنویس‌ها به تحلیل موضوعی محتوای متنی فیلم‌ها (زیرنویس) پرداخته‌ایم. به این منظور با بهره‌گیری از ویژگی‌های استخراج‌شده به وسیله الگوریتم LDA و سنجش شباهت کسینوسی، اقدام به ساخت گراف شباهت فیلم‌ها نموده‌ایم. در این گراف هر گره معرف یک فیلم و هر یال بیانگر شباهت میان دو فیلم است. در ادامه با استفاده از روش‌های خوشه‌بندی بر روی گراف فیلم‌ها توانستیم در خوشه‌ها، هم‌بستگی موضوعی قابل توجهی میان فیلم‌ها بدست آوریم.

کلمات کلیدی

داده‌کاوی، استخراج موضوع، تحلیل گراف، فیلم، زیرنویس

استخراج شباهت فیلم‌ها و توصیه فیلم‌های مشابه به مخاطبان صنعت سینما، می‌تواند برای خدمات گوناگونی همچون وبسایت‌های پخش آنلاین ویدئو، ناشران فیلم‌ها، فروشگاه‌های محتوای چندرسانه‌ای و خدمات اطلاع‌رسانی بسیار مفید باشد.

۱- مقدمه

باتوجه به حجم روزافزون داده‌های ویدئویی، توسعه روش‌هایی برای خوشه‌بندی و جست‌وجو سریع میان آن‌ها اهمیت فراوانی پیدا کرده است.

هم‌بستگی میان موضوع فیلم‌ها و ویژگی‌های سطح پایینی که با بهره‌گیری از الگوریتم LDA [3] از متن زیرنویس آن‌ها استخراج شده، سنجیده شود.

یکی دیگر از کارهای اخیر که در زمینه دسته‌بندی فیلم‌ها با استفاده از زیرنویس انجام شده است، بهره‌گیری از روش‌های یادگیری ماشین بدون ناظر، بدون نیاز به برچسب‌گذاری داده‌ها در تمام طول فرایند است [7]. مقاله [7] از در کنار هم قرار دادن دو روش استخراج ویژگی و یادگیری بدون ناظر استفاده کرده است [8]. همچنین از روش کاهش بعد^۱ برای کاهش ابعاد در این پژوهش استفاده شده است. در بخش یادگیری بدون ناظر مشاهده شده است که الگوریتم‌های K-Means و Bisecting K-Means [2] از نظر کیفیت خوشه‌بندی، بهترین عملکرد را داشته‌اند.

در یک پژوهش دیگر، به دسته‌بندی چندبرچسبی فیلم‌ها با بهره‌گیری از روش‌های بانظارت یادگیری ماشین برای دسته‌بندی فیلم‌ها در ژانرهای مرتبط پرداخته شده است. نوآوری این روش در استفاده از دسته‌بندی‌های بهینه‌شده و ترکیب آن‌ها است. در مقاله [9] از ترکیب SVM^۱ و DT^۱ استفاده شده است. در مقاله [9] مشاهده می‌شود که از استفاده از دسته‌بند ترکیبی فارغ از دیتاست و بردار ویژگی استفاده شده، دقت بهتری را ارائه می‌دهد.

۱-۱-۲- سیستم‌های توصیه‌گر

در مقاله [10] سعی می‌شود با استفاده از یک مدل چندگانه براساس محتوای متنی فیلم‌ها (زیرنویس)، همچنین استفاده از کانال‌های صوتی و تصویری، یک ارائه^۲ از فیلم استخراج شود. سپس با کنار هم قرار گرفتن ویژگی‌های استخراج شده از سه محتوای متنی، صوتی و تصویری فیلم‌ها، به همراه داده‌هایی مانند کارگردان و بازیگران فیلم‌ها اقدام به استخراج شباهت فیلم‌ها استفاده شده است. بررسی‌ها نشان می‌دهد که استفاده از ویژگی‌های متنی، صوتی و تصویری عملکرد سیستم توصیه‌گر را تا ۵۰٪ نسبت به حالتی که تنها از مشخصات کلی فیلم‌ها مانند کارگردان و بازیگران استفاده شده است، بهبود می‌بخشد.

در [11] یک سیستم توصیه‌گر معرفی می‌شود که در آن تنها از محتوای فیلم‌ها (زیرنویس‌ها) برای استخراج ویژگی استفاده می‌شود. در این پژوهش از روش‌های بازیابی اطلاعات^۳، پردازش زبان طبیعی و مطالعه سبکی^۴ بهره‌گرفته شده است تا میزان هم‌بستگی میان شباهت ویژگی‌های استخراج شده با شباهت موضوعی فیلم‌ها سنجیده شود.

خالصه‌سازی ویدئوها رویکرد مناسبی برای درک محتوای آن‌ها و جستجو بهینه برای یافتن محتوای موردنظر است. در مقاله [12] تلاش شده است، بستری برای تولید خودکار پیش‌پرده^۵ فیلم‌ها با توجه به زیرنویس آن‌ها ایجاد شود. در این پژوهش [12] از ویژگی‌های استخراج شده از متن زیرنویس فیلم‌ها، برای دسته‌بندی آن‌ها در ژانرهای مرتبط استفاده شده است که در آن دقت دسته‌بندی فیلم‌ها به ۸۹٪ رسیده است.

فیلم‌های ساخت Hollywood در کشور اندونزی مخاطبان زیادی دارند. برای جلوگیری از این که محتوای نامناسب برای کودکان، حتی پس از سانسور فیلم‌ها، مورد تماشای آن‌ها قرار گیرد، در مقاله [13] سعی شده است با خوشه‌بندی فیلم‌ها با استفاده از زیرنویس ترجمه شده آن‌ها به زبان اندونزیایی، امتیاز کاربران و ژانر آن‌ها، روشی ایجاد شود، تا به‌عنوان منبعی

اکثر سرویس‌های موجود در این زمینه، از اطلاعاتی مانند سلیقه و نظرات کاربران، یا ویژگی‌هایی نظیر بازیگران، کارگردان و ژانر فیلم، یا ترکیبی از این دو استفاده می‌کنند و معمولاً محتوای اصلی فیلم، توسط آن‌ها به کارگرفته نمی‌شود. این موضوع سبب می‌شود فیلم‌هایی که داده ثبت شده زیادی توسط کاربران برای آن‌ها وجود ندارد، کمتر دیده شوند.

کارهای گوناگونی به‌منظور یافتن شباهت موضوعی فیلم‌ها و در جهت توسعه سیستم‌های توصیه‌گر با تکیه بر محتوای خود فیلم‌ها انجام شده است. به طور کلی تلاش‌های انجام شده در این زمینه از روش‌های گوناگونی مانند استخراج موضوع از متن^۱ [1]، پردازش زبان طبیعی، یادگیری با نظارت و یادگیری بدون نظارت برای استخراج ویژگی از محتوای متنی، صوتی، تصویری فیلم‌ها استفاده می‌کنند. سپس با دسته‌بندی با کمک یادگیری ماشین یا بهره‌گیری از انواع روش‌های خوشه‌بندی [2]، براساس ویژگی‌های استخراج شده، سعی بر یافتن هم‌بستگی موضوعی میان فیلم‌ها، می‌کنند.

یکی از رویکردهای مفید و پرکاربرد در زمینه داده‌کاوی و کشف شباهت‌ها، مدل‌سازی مسئله در قالب گراف است. ساخت گراف می‌تواند در بصری‌سازی^۲ و ارائه قابل درک داده‌ها بسیار کمک‌کننده باشد. مدل‌سازی فیلم‌ها در قالب گراف امکان مطالعه ارتباطات، کشف شباهت‌های میان فیلم‌ها و بهره‌گیری از تحلیل‌های گرافی را فراهم می‌سازد.

در این مقاله از الگوریتم LDA [3] برای استخراج ویژگی‌های سطح پایین از زیرنویس فیلم‌ها بهره‌گرفته‌ایم. در ادامه با بهره‌گیری از ویژگی‌های استخراج شده و سنجش شباهت کسینوسی^۴ [4] میان بردارهای ویژگی فیلم‌ها، اقدام به ساخت گراف شباهت فیلم‌ها نموده‌ایم. سپس سعی می‌کنیم با استفاده از روش‌های تحلیل گراف^۴ و خوشه‌بندی^۵ بر روی گراف فیلم‌ها، هم‌بستگی^۶ موضوعی میان فیلم‌ها را بسنجیم.

۱-۱-۱- کارهای مرتبط

در این قسمت به مرور چند نمونه از کارهایی که در زمینه استخراج موضوع از متن و تشخیص شباهت موضوعی متن‌ها و فیلم‌ها صورت گرفته است پرداخته می‌شود. کارهای انجام شده در دو دسته کشف شباهت متن‌ها و توسعه سیستم‌های توصیه‌گر^۷ بررسی می‌شوند.

۱-۱-۱-۱- کشف شباهت میان متن‌ها

در [5] سعی شده با استفاده از روش‌های استخراج موضوع، کشف شود که در متن‌های درام فرانسوی مربوط به دوره کلاسیک و عصر روشنگری^۸، زیرژانرهای ژانر درام تا چه حد دارای موضوعات عمده و الگوهای موضوعی مربوط به متن هستند. به این شکل که تا چه حد استفاده از الگوریتم‌های خوشه‌بندی براساس امتیازات بدست‌آمده از تشخیص موضوعات، با تفاوت‌های مرسوم زیرژانرها منطبق است.

یکی از کارهای انجام شده در زمینه کشف شباهت میان فیلم‌ها، استفاده از ویژگی‌های سطح پایین استخراج شده از متن زیرنویس فیلم‌ها با بهره‌گیری از روش‌های پردازش زبان طبیعی و استخراج موضوع است. در [6] سعی شده

زیرنویس ارائه می‌شود. اکثر زیرنویس‌ها شامل توضیحات اضافه‌ای هستند که بعضی از صداها را مهم فیلم مانند صدای همهمه، زمزمه، تشویق، عبور خودروها و غیره را توصیف می‌کنند. این توضیحات معمولاً در داخل علائم [] و () آورده می‌شوند. باتوجه به این که توضیحات اضافی جزء متن اصلی زیرنویس نیستند، از زیرنویس‌ها حذف می‌شوند.

در فرایند استخراج ویژگی از زیرنویس‌ها، نتایج تحت تأثیر اسامی خاص مانند شخصیت‌های فیلم، اسم شرکت‌ها و شهرها و غیره قرار می‌گیرد. اسامی خاص در یک زیرنویس به دفعات تکرار می‌شوند و به ندرت در زیرنویس فیلم‌های دیگر دیده می‌شوند، به همین سبب گاهی جزو کلمات کلیدی و تعیین‌کننده تشخیص داده می‌شوند. این در حالی است که این اسامی بار مفهومی مفیدی برای کشف شباهت معنایی میان فیلم‌ها ندارند؛ از این رو با استفاده از ابزار NER، ابتدا اسامی خاص از متن زیرنویس‌ها تشخیص داده و سپس حذف شده‌اند.

پس از پاک‌سازی اولیه، می‌بایست فرایند تکواژه‌سازی^{۳۳} روی متن زیرنویس‌ها انجام شود تا برای پردازش‌های بعدی آماده شوند. از میان رویکردهای گوناگونی که برای تکواژه‌سازی متن وجود دارد، روش lemmatization که هر کلمه را به ریشه آن تبدیل می‌کند، نتایج بهتری را در پی داشت.

بخش عمده‌ای از بار مفهومی جملات مربوط به اسم‌ها، افعال، صفت‌ها و قیده‌های جمله است. از این رو می‌توان ضمن حفظ کلمات پراهمیت و تعیین‌کننده، اندازه پیکره متنی را با حذف دیگر کلمات کاهش داد. به همین منظور پس از تکواژه‌سازی متن، ابتدا کلمات ایست حذف شده‌اند، سپس با بهره‌گیری از ریشه‌یابی، فقط کلماتی که یکی از نقش‌های اسم، فعل، صفت یا قید را دارند، در پیکره متنی نگه‌داشته شده‌اند و از باقی کلمات چشم‌پوشی شده است.

۲-۳-۲- حذف داده‌های پرت

در فرایند داده‌کاوی در بسیاری از موارد، حذف داده‌های پرت قبل از انجام پردازش‌های بعدی انجام می‌شود. در این مقاله نیز برای بهبود دقت در اجرای الگوریتم‌ها و پردازش‌های آینده، داده‌های پرت حذف شده‌اند. بدین منظور زیرنویس‌هایی که تعداد کلمات آن‌ها از آستانه^{۳۴} مشخصی کمتر و بیشتر است از پایگاه‌داده کنارگذاشته شده‌اند. برای شناسایی داده‌های پرت و تعیین آستانه برای حداقل و حداکثر کلمات در زیرنویس‌ها از روش تشخیص داده‌های پرت به کمک منحنی نرمال استفاده شده است. برای ادامه کار از بازه $(\mu - 2\sigma, \mu + 2\sigma)$ که در آن ۹۵٪ داده‌ها انتخاب می‌شوند، استفاده شده است. شکل (۱) گزینش داده‌ها بر اساس تعداد کلمات زیرنویس‌ها را نشان می‌دهد. براین اساس از میان ۴۲۸۴ فیلم موجود در پایگاه‌داده، ۴۱۱۴ نمونه انتخاب می‌شوند و ۱۷۰ نمونه به عنوان داده پرت کنار گذاشته می‌شوند.

۲-۳-۳- کاوش پایگاه‌داده^{۳۵}

مرحله کاوش پایگاه‌داده یکی از مراحل مهم در فرایند استخراج دانش است. در این مرحله با استفاده از تحلیل‌های آماری و جست‌وجو در داده‌ها می‌توان

برای تشخیص فیلم‌های نامناسب برای بچه‌ها مورد استفاده قرار گیرد. در این پژوهش [13] از روش‌های متن‌کاوی^{۳۶} [14] برای استخراج کلماتی که بیشتر در سه دسته کلمات نامناسب، کلمات جنسی و کلمات ترسناک دیده می‌شوند استفاده شده است.

۲- مجموعه داده

داده‌های گردآوری شده برای این پژوهش به دو بخش کلی تقسیم می‌شوند. بخش اول پیکره متنی است که از زیرنویس فیلم‌ها تشکیل شده است. بخش دوم مربوط به اطلاعاتی از فیلم‌ها مانند: عنوان، سال انتشار، ژانرها، میانگین امتیاز کاربران، تعداد آرای کاربران، نام نویسندگان و کارگردانان است.

۲-۱- گردآوری داده‌ها

برای گردآوری متن زیرنویس‌ها، که در این مقاله داده‌های اصلی برای استخراج شباهت موضوعی فیلم‌ها هستند، از روش استخراج فایل^{۳۷} از وبسایت‌ها بهره گرفته شده است. در مرحله استخراج، ۵۳۵۵ نمونه زیرنویس بدست آمده است.

برای بدست آوردن اطلاعات فیلم‌ها نیز، از پایگاه‌داده‌های وبسایت IMDB [15] استفاده شده است. این وبسایت بخشی از اطلاعات خود را در قالب چند پایگاه‌داده جداگانه شامل پایگاه‌داده اطلاعات کلی فیلم‌ها، پایگاه‌داده رأی‌های کاربران و پایگاه‌داده کارگردانان و نویسندگان در دسترس عموم قرار داده است.

۲-۲- یکپارچه‌سازی^{۳۸} و کاهش داده‌ها^{۳۹}

در مرحله یکپارچه‌سازی زیرنویس‌های استخراج شده، با پایگاه‌داده اطلاعات فیلم‌ها، ۴۲۸۴ نمونه به صورت یک‌به‌یک با داده‌های موجود در پایگاه‌داده فیلم‌ها نظیر شده‌اند. از میان نمونه‌های اولیه استخراج شده، ۱۰۷۱ زیرنویس به علت تفاوت در عنوان فیلم‌ها در دو پایگاه‌داده، مطابقت داده نشدند، که از پیکره داده‌ها کنار گذاشته شده‌اند.

۲-۳- پیش‌پردازش

در اولین قدم برای پیش‌پردازش متن زیرنویس‌ها به پاک‌سازی آن‌ها پرداخته می‌شود. بدین منظور، طی چندگام متن زیرنویس‌ها پالایش می‌شود.

۲-۳-۱- پالایش متن زیرنویس‌ها

به منظور پاک‌سازی متن زیرنویس‌ها، ابتدا فراداده‌های^{۴۰} که در فایل‌های زیرنویس‌ها در قالب برچسب‌های HTML^{۴۱} وجود دارد، حذف می‌شوند. این برچسب‌ها معمولاً شامل اطلاعاتی در مورد تهیه‌کننده زیرنویس، وبسایت ارائه‌دهنده، مشخصات فیلم و جزئیاتی در مورد شیوه نمایش زیرنویس از جمله نوع و رنگ قلم هستند.

یکی از اهداف مهم زیرنویس‌ها، فراهم‌سازی امکان تماشای محتوای ویدئویی برای افراد ناشنوا است. از این رو حتی برای فیلم‌های صامت نیز

شایان ذکر است ۹۰.۳٪ فیلم‌ها از در بازه سال ۲۰۱۰ تا ۲۰۲۱، ۳.۵٪ در بازه ۲۰۰۰ تا ۲۰۰۹ و ۶.۱٪ در بازه ۱۹۱۸ تا ۱۹۹۹ قرار دارند.

• **کلمات پراهمیت در هر ژانر**

برای استخراج کلمات مهم هر ژانر از معیار $tf-idf^{25}$ بهره گرفته شده است [16]. بدین منظور پس از پیش‌پردازش و تکواژه‌سازی متن زیرنویس‌ها، هر ژانر به‌عنوان یک سند در نظر گرفته شده است و کلمات موجود در زیرنویس هر فیلم که در آن ژانر حضور دارد، در سند نماینده آن ژانر تجمیع شده است. در جدول (۱) پراهمیت‌ترین کلمات برای هر ژانر براساس معیار $tf-idf$ ذکر شده است.

جدول ۱: پراهمیت‌ترین کلمات برای هر ژانر براساس امتیاز $tf-idf$

Genre	Most Important Words
Action	agent, target, cover, chief, save, weapon, army, officer, report, security, fight, force, mission
Adventure	destroy, key, sword, dangerous, hero, save, island, peace, planet, fly, magic, ship, attack, secret, space
Animation	batman, dragon, monster, princess, superman, hero, wolf, evil, sky, super, planet, magic, awesome
Biography	art, English, film, support, talk, peace, brother, coach, decide, government, holy, lead, letter, respect
Comedy	idiot, birthday, coffee, cute, bro, college, favorite, kiss, lovely, uncle, awesome, class, glad, kidding
Crime	cop, detective, jail, killer, lawyer, missing, murder, bank, crime, drug, evidence, chief, cash, court
Documentary	action, animal, build, campaign, cancer, character, community, company, society, science, culture
Drama	hospital, mama, Christ, evening, lie, Christmas, love, poor, song, dance, office, husband, marry, couple
Family	cat, fairy, grandma, prince, Santa, witch, grandpa, merry, tree, summer, lovely, uncle, sweet, couple
Fantasy	sea, soul, magic, protect, ship, master, Christmas, dark, queen, fear, future, earth, lord, king, kill
History	Lincoln, madam, minister, English, American, freedom, narrator, nation, political, vote, battle, majesty
Horror	mommy, mum, camera, evil, holy, hospital, quiet, chartist, dark, wake, police, lord, sound, feeling
Music	concert, crew, fan, guitar, light, performance, pop, records, singer, audience, album, song, studio
Musical	angry, arm, beach, bless, blind, broke, cheer, darkness, doubt, gift, happiness, heaven, impossible
Mystery	accident, murder, missing, detective, mum, camera, evidence, police, hospital, questions, office, lie office, lie office, lie office, lie
Romance	hotel, birthday, marriage, letter, cute, wedding, marry, darling, kiss, lovely, collage, love
Sci-Fi	alien, contact, signal, station, suit, machine, sound, weapon, planet, security, protect, mission, space
Sport	announcer, arm, baseball, boxing, Brazil, champ, breathe, championship, chess, cup, field, fight
Thriller	cops, officer, security, questions, clean, Christ, goddamn, boss, street, gun, captain, police, shoot
War	bridge, command, cross, duty, force, German, lieutenant, officer, Russian, tank, troop, truck, wound
Western	blondie, bounty, clinch, deputy, desert, east, farmer, fair, folk, fellow, gun, hat, horse, Indian, pistol

اطلاعات و شناخت خوبی از پایگاه‌داده بدست آورد، که می‌تواند در ادامه فرایند پژوهش بسیار مفید باشد. در ادامه چند نمونه از بررسی‌های انجام‌شده ذکر می‌شود.

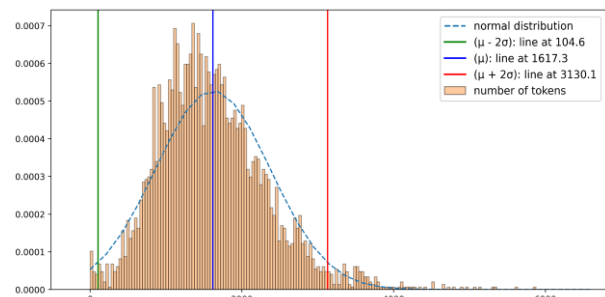
• **بررسی آماری براساس ژانر فیلم‌ها**

با توجه به وب‌سایت IMDB در کل ۲۷ ژانر مختلف برای فیلم‌ها وجود دارد. فیلم‌هایی که در پایگاه‌داده وجود دارند فقط مربوط به یک ژانر نیستند و تعداد ژانرهای هر فیلم بین ۱ تا ۳ ژانر متغیر است. از میان ۴۲۸۴ نمونه حاضر، ۷۶۲ نمونه فقط در یک ژانر، ۱۱۰۳ نمونه دو ژانر و ۲۴۱۹ نمونه در ۳ ژانر مختلف حضور دارند. شایان ذکر است، ۵ ژانر News, Film-Noir, Game-Show, Reality-TV و Talk-Show که تعداد فیلم‌های آن‌ها کمتر از ۱۰ عدد بود، کنار گذاشته شده‌اند. همچنین ژانر Short نیز به این علت که به‌عنوان نوع فیلم در نظر گرفته شده است، در این قسمت کنار گذاشته شده است.

شکل (۲) توزیع فیلم‌ها در ژانرهای مختلف را نمایش می‌دهد. گفتنی است که ژانرهای درام، کمدی، اکشن به‌ترتیب بیشترین نمونه‌ها را به خود اختصاص داده‌اند.

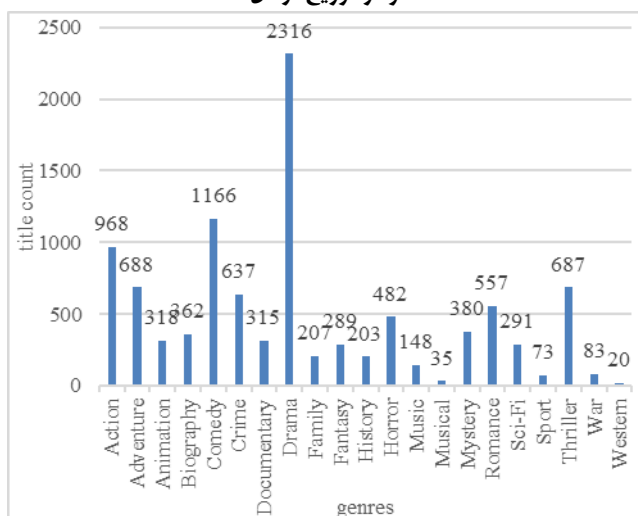
• **بررسی آماری براساس سال انتشار فیلم‌ها**

فیلم‌های پایگاه‌داده از نظر سال انتشار، در بازه ۱۹۱۸ تا ۲۰۲۱ جای می‌گیرند. سال ۲۰۱۹ با ۶۴۸ عنوان بیشترین تعداد فیلم را به خود اختصاص داده است.



شکل ۱: نمودار میله‌ای تعداد کلمات در متن زیرنویس‌ها به‌همراه

نمودار توزیع نرمال



شکل ۲: نمودار تعداد فیلم‌های موجود در هر ژانر

در شکل (۳) تغییرات شمار رأس‌های گراف برحسب تغییر مقدار آستانه شباهت مشاهده می‌شود. باتوجه‌به این‌که با آستانه شباهت ۹۰٪ تغییر چشمگیری در شمار رأس‌ها مشاهده می‌شود، برای ساخت گراف شباهت از این مقدار استفاده شده است. براین اساس شمار رأس‌های گراف شباهت به ۳۶۵۸ و شمار یال‌ها به ۵۶۰۷۱ می‌رسد. شکل (۴) نمای کلی گراف ساخته شده را نمایش می‌دهد.

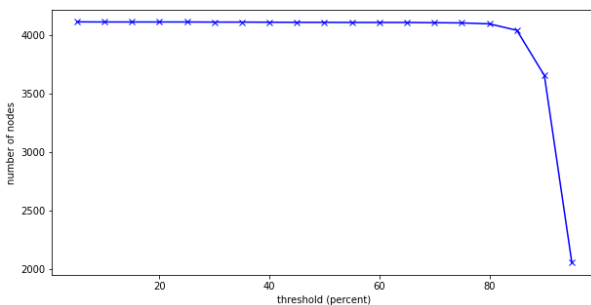
۳-۳- تحلیل گراف شباهت

پس از ساخت گراف شباهت، می‌توان گراف را با رویکردهای گوناگونی بررسی کرد. در ادامه برخی از تحلیل‌هایی انجام شده بر روی گراف بیان می‌شود.

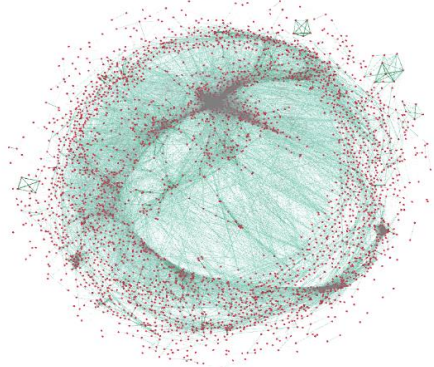
۳-۳-۱- بررسی گراف از نظر درجات رأس‌ها

یکی از نخستین بررسی‌هایی که می‌توان در تحلیل گراف‌ها انجام داد، بررسی گراف از نظر درجه رأس‌ها است. پس از محاسبه درجه رأس‌ها، نمودار توزیع درجه رأس‌ها مطابق شکل (۵) بدست آمده است. میانگین درجه رأس‌ها برابر ۳۰۶۵۷ است.

همانطور که با مقایسه شکل (۲) و شکل (۶) مشاهده می‌شود، با وجود بیشتر بودن شمار فیلم‌هایی که در ژانر درام هستند، به ترتیب میانگین درجه رأس‌های ژانرهای کمدی و عاشقانه بیشتر از ژانر درام می‌باشد. این مسئله نشان می‌دهد که موضوعات کمدی و عاشقانه، امکان ترکیب شدن با طیف وسیعی از موضوعات مختلف، از جمله موضوعات سیاسی، خانوادگی، ورزشی، هیجانی، اکشن، درام و غیره را دارند. به همین سبب است که دو ژانر کمدی و عاشقانه بیشتر به همراه انواع دیگر ژانرها، مشاهده می‌شوند.



شکل ۳: نمودار تغییرات شمار رأس‌ها برحسب درصد آستانه شباهت



شکل ۴: نمای کلی گراف ساخته شده

۳- ساخت و تحلیل گراف

به منظور مدل‌سازی فیلم‌ها و شباهت میان آن‌ها در قالب گراف، ابتدا به استخراج ویژگی از متن زیرنویس‌ها می‌پردازیم. سپس از سنجه شباهت کسینوسی [4] برای محاسبه شباهت میان بردار ویژگی فیلم‌ها و ساخت ماتریس شباهت بهره می‌گیریم. پس از آن بر اساس ماتریس شباهت، گراف شباهت فیلم‌ها ساخته می‌شود. در ادامه به تشریح هر یک از این مراحل و درنهایت به تحلیل گراف ساخته شده می‌پردازیم.

۳-۱- استخراج ویژگی

برای استخراج ویژگی از پیکره متنی، از الگوریتم LDA [3] استفاده شده است. در این الگوریتم، هر سند به شکل توزیع احتمالی از موضوعات مدل می‌شود. تعداد موضوعات می‌بایست از پیش، به عنوان ورودی تعیین شود. برای یافتن تعداد موضوعات مناسب از معیار coherence بهره گرفته شده است که باتوجه به آن، در نهایت تعداد موضوعات برابر ۱۵۰ انتخاب شده است. خروجی LDA [3]، برای هر فیلم، یک بردار ویژگی است که در واقع توزیع احتمالی از موضوعات است.

با استفاده از سنجه شباهت کسینوسی [4]، شباهت بردار هر یک از سندها با بردار سندهای دیگر محاسبه می‌شود و نتایج در قالب یک ماتریس با طول و عرض برابر با تعداد زیرنویس‌ها (۴۱۱۴ × ۴۱۱۴) بدست می‌آید. در این ماتریس شباهت میان فیلم‌ها به صورت دویبدو در هر درایه به شکل مقداری بین ۰ و ۱ بیان شده است.

۳-۲- ساخت گراف شباهت

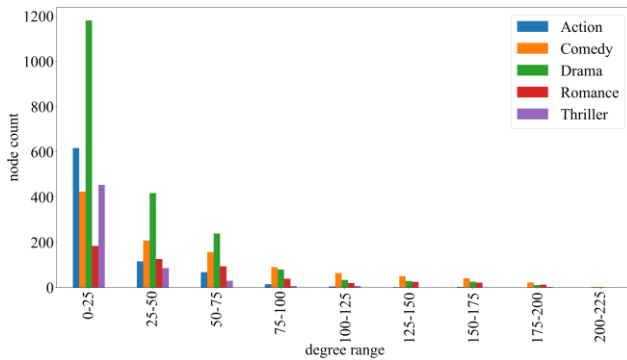
برای این‌که بتوان، رابطه بین زیرنویس فیلم‌ها را در قالب گراف تحلیل کرد، زیرنویس فیلم‌ها در قالب رأس‌های گراف و اطلاعات هر فیلم (مانند ژانرها، سال ساخت و غیره) به عنوان ویژگی‌های رأس‌ها در نظر گرفته شده است. شباهت کسینوسی [4] میان بردار ویژگی فیلم‌ها نیز در قالب وزن یال‌ها مدل شده است.

باتوجه‌به این‌که اگر یال میان هر جفت از فیلم‌ها در گراف وجود داشته باشد تحلیل گراف مشکل می‌شود. برای کاهش شمار یال‌ها در گراف تنها یال‌هایی که وزن آن‌ها (شباهت کسینوسی میان دو رأس) از آستانه مشخصی بیشتر است، انتخاب شده‌اند.

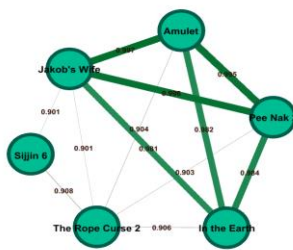
۳-۲-۱- تعیین آستانه برای میزان شباهت کسینوسی میان

زیرنویس‌ها

انتخاب آستانه شباهت مناسب، براساس تعداد رأس‌های گراف انجام شده است. در این راستا مشاهده می‌شود، با افزایش آستانه شباهت، شمار زیرنویس‌هایی که شباهت آن‌ها با هیچ زیرنویس دیگری بیش از آستانه تعیین شده نیست نیز افزایش می‌یابد. در نتیجه شمار رأس‌های گراف نیز کاهش می‌یابد.



شکل ۷: تغییرات تعداد فیلمها در ۵ ژانر برحسب درجه رأسها



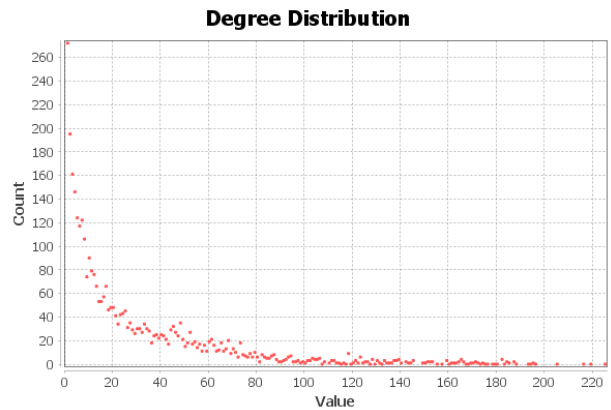
شکل ۸: زیرگراف کوچک ۱

اطلاعات اولیه فیلمها نظیر ژانرها، نام کارگردانان و بازیگران در کنار ویژگیهای استخراج شده استفاده کرد. یک رویکرد، در این زمینه انتخاب رأسها براساس ویژگیهای اولیه فیلمها و سپس خوشه‌بندی رأسهای انتخاب شده است. به این ترتیب برای کشف شباهت‌های موضوعی، می‌توان از دیدگاه جزئی‌تر در میان فیلم‌هایی کاوش کنیم که از نظر یک ویژگی اولیه با یکدیگر وجه اشتراک دارند.

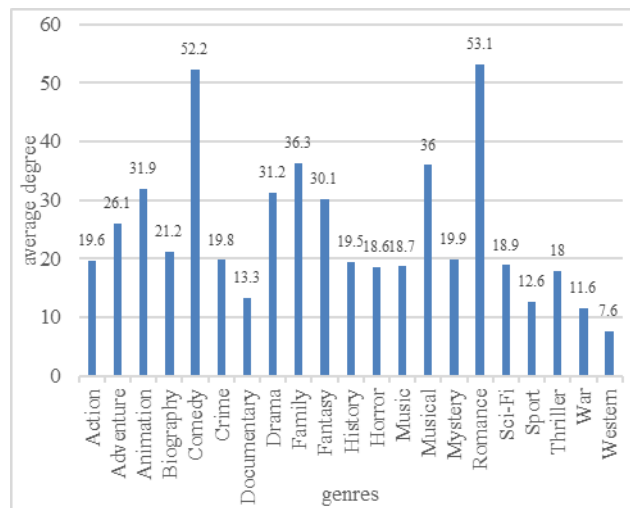
• خوشه‌بندی رأسهای دارای ژانر زندگی‌نامه

به‌عنوان نمونه، با بررسی خوشه‌بندی رأسهایی که دارای ژانر زندگی‌نامه هستند، هم‌بستگی موضوعی جالبی میان رأسها در خوشه‌ها مشاهده می‌شود. بدین منظور ابتدا رأسهایی که دارای ژانر زندگی‌نامه هستند، انتخاب می‌شوند، سپس الگوریتم خوشه‌بندی [17] روی آنها اجرا می‌شود. سنجه Resolution [18] برای خوشه‌بندی انجام شده برابر ۰.۶۶۴ اندازه‌گیری شده است، که براین اساس، می‌توان گفت خوشه‌بندی از نظر تفکیک‌پذیری کیفیت مناسبی دارد.

در شکل (۹) گراف رأسهای دارای ژانر زندگی‌نامه پس از خوشه‌بندی نمایش داده شده است. در شکل (۹) خوشه‌های عمده با رنگ‌های سبز، بنفش، نارنجی، آبی‌روشن، زرد، قرمز و آبی تیره مشخص شده‌اند. با بررسی رأسهای هر خوشه، مشاهده می‌شود که به‌طور کلی فیلم‌های هر خوشه، مربوط به زندگی‌نامه شخصیت‌هایی هستند، که از نظر حرفه، دوره تاریخی یا جایگاه اجتماعی شبیه یکدیگرند. برای مثال در جدول (۲) به موضوعات اصلی در خوشه‌های عمده رأسهای دارای ژانر زندگی‌نامه اشاره شده است.



شکل ۵: نمودار توزیع درجات رأسها



شکل ۶: نمودار میانگین درجه رأسهای مربوط به هر ژانر

در شکل (۷) نیز مشاهده می‌شود که فیلم‌هایی که دارای ژانرهای اکشن و هیجانی هستند، بیشتر در رأسهایی که درجات پایین‌تر از ۲۵ دارند ظاهر شده‌اند. همین‌طور در شکل (۷) مشاهده می‌شود، در رأسهایی که درجه بیش از ۷۵ دارند، تعداد فیلم‌هایی که دارای ژانر کمدی هستند از تعداد فیلم‌هایی که دارای ژانر درام هستند، پیشی می‌گیرد.

۲-۳-۳- بررسی چند نمونه از زیرگراف‌های کوچک

در تحلیل گرافها، معمولاً با بررسی اجتماع^{۲۸} رأسها و زیرگرافها می‌توان ویژگی‌های مشترک و قابل توجهی در میان رأسها یافت. از این جهت بررسی زیرگرافها می‌تواند به کشف شباهت‌های موضوعی میان فیلمها کمک کند. به‌عنوان نمونه، در شکل (۸) یکی از زیرگرافها مشاهده می‌شود. با بررسی مشخصات فیلم‌های این زیرگراف، مشاهده می‌شود که تمام فیلم‌های این زیرگراف متعلق به ژانر ترسناک می‌باشند.

۳-۳-۳- خوشه‌بندی گراف برحسب ویژگی‌های اولیه

فیلمها

برای یافتن فیلم‌هایی که با یکدیگر شباهت موضوعی دارند، می‌توان از

- "SubRosa: Determining Movie Similarities based on Subtitles," 2021. [Online].
- [12] M. Hesham, B. Hani, N. Fouad and E. Amer, "Smart trailer: Automatic generation of movie trailer using only subtitles," 2018. [Online].
- [13] A. D. Alfarizy, Indahwati and B. Sartono, "Clustering box office movie with Partition Around Medoids (PAM) Algorithm based on Text Mining of Indonesian subtitle," 3 2017. [Online]. Available: <https://doi.org/10.1088/1755-1315/58/1/012032>.
- [14] R. Feldman, J. Sanger and others, The text mining handbook: advanced approaches in analyzing unstructured data, Cambridge university press, 2007.
- [15] IMDb, "IMDb Datasets," 16 05 2021. [Online]. Available: <https://www.imdb.com/interfaces/>.
- [16] P. Bafna, D. Pramod and A. Vaidya, "Document clustering: TF-IDF approach," 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pp. 61-66, 2016.
- [17] V. D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, "Fast unfolding of communities in large networks," 2008. [Online].
- [18] R. Lambiotte, J.-C. Delvenne and M. Barahona, "Laplacian dynamics and multiscale modular structure in networks," 2008. [Online].

زیر نویس ها

topic extraction	۱
visualization	۲
cosine similarity	۳
graph analysis	۴
clustering	۵
correlation	۶
recommendation system	۷
age of enlightenment	۸
dimension reduction	۹
support-vector machine	۱۰
decision tree	۱۱
representation	۱۲
information retrieval	۱۳
stylometry	۱۴
trailer	۱۵
text mining	۱۶
scraping	۱۷
data integration	۱۸
data reduction	۱۹
metadata	۲۰
html tags	۲۱
tokenization	۲۲
threshold	۲۳
data exploration	۲۴
term frequency – inverse document frequency	۲۵
attribute	۲۶
graphlet	۲۷

در نهایت با تحلیل گراف ساخته شده، به بررسی ارتباط میان فیلم‌ها به منظور کشف هم‌بستگی و شباهت موضوعی میان آن‌ها می‌پردازیم. در این بخش با بررسی اجتماع رأس‌ها و زیرگراف‌های کوچک هم‌بستگی، موضوعی چشمگیری میان رأس‌های آن‌ها مشاهده شد. همچنین با گزینش فیلم‌ها براساس اطلاعات اولیه آن‌ها و سپس خوشه‌بندی آن‌ها، شاهد قرابت موضوعی قابل توجهی در خوشه‌ها بودیم.

توسعه روش‌هایی برای استخراج شباهت فیلم‌ها و توصیه فیلم‌ها براساس محتوای آن‌ها اهمیت فراوانی پیدا کرده است. یکی از اقداماتی که می‌تواند در زمینه توسعه سیستم‌های خوشه‌بندی و توصیه‌گر فیلم‌ها انجام شود، مدل‌سازی محتوای ویدئوها در قالب گراف است، که به شکل خودکار امکان تحلیل و استخراج شباهت میان فیلم‌ها را فراهم سازد. استخراج خودکار شباهت موضوعی فیلم‌ها و توسعه سیستم‌های توصیه‌گرمی‌تواند برای خدمات گوناگونی همچون وب‌سایت‌های پخش آنلاین ویدئو، ناشران فیلم‌ها و فروشگاه‌های محتوای چندرسانه‌ای مفید باشد.

مراجع

- [1] R. Ibrahim, A. Elbagoury, M. S. Kamel and F. Karray, "Tools and Approaches for Topic Detection from Twitter Streams: Survey," 3 2018. [Online]. Available: <https://doi.org/10.1007/s10115-017-1081-x>.
- [2] D. Xu and Y. Tian, "A Comprehensive Survey of Clustering Algorithms," 2015. [Online]. Available: <https://doi.org/10.1007/s40745-015-0040-1>.
- [3] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation," 3 2003. [Online].
- [4] I. S. Dhillon, Y. Guan and J. Kogan, "Refining clusters in high-dimensional text data," in Proceedings of the workshop on clustering high dimensional data and its applications at the second SIAM international conference on data mining, 2002.
- [5] C. Schöch, "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama," 2021. [Online].
- [6] K. Bougiatiotis and T. Giannakopoulos, "Content Representation and Similarity of Movies Based on Topic Extraction from Subtitles," 2016. [Online]. Available: <https://doi.org/10.1145/2903220.2903235>.
- [7] M. M. Hasan, S. T. Dip, T. M. Kamruzzaman, S. Akter and I. Salehin, "Movie Subtitle Document Classification Using Unsupervised Machine Learning Approach," 2021. [Online].
- [8] L. Liu, J. Kang, J. Yu and Z. Wang, "A comparative study on unsupervised feature selection methods for text clustering," 2005. [Online].
- [9] M. M. Hasan, S. T. Dip, T. Rahman, M. S. Akter and I. Salehin, "Multilabel Movie Genre Classification from Movie Subtitle: Parameter Optimized Hybrid Classifier," 2021. [Online].
- [10] K. Bougiatiotis and T. Giannakopoulos, Enhanced movie content similarity based on textual, auditory and visual information, vol. 96, 2018, pp. 86-102.
- [11] M. A. N. D. T. J. Luhmann Jan AND Burghardt,

JCWR2022

May 11-12, 2022; Tehran, Iran

8th International Conference on
Web Research

community ^{٢٨}

SID



بلاگ مرکز اطلاعات علمی



کارگاه‌های آموزشی



سرویس ترجمه تخصصی



فیلم‌های آموزشی

کارگاه‌ها و فیلم‌های آموزشی مرکز اطلاعات علمی

آشنایی با پایگاه‌های اطلاعات علمی بین‌المللی و ترندهای جستجو بین‌المللی و ترندهای جستجو

کاربرد نرم افزار SPSS در پژوهش

بروبوزال نویسی (علوم انسانی)

کاربرد نرم‌افزار End Note در استناددهی مقالات و متون علمی

صدور گواهینامه نمایه مقالات نویسندگان در SID