



دانشکده مهندسی کامپیوتر

بازیابی پیشرفته اطلاعات

مدرس: دکتر بیگی

شماره گروه: ۵

تهیه کنندگان: نیما جمالی - سپهر فعلی - سینا کاظمی

گزارش فاز سوم پروژه

فهرست مطالب

| | |
|----|--|
| ۲ | خوشه‌بندی |
| ۲ | پیش پردازش اخبار |
| ۲ | فضای TF-IDF |
| ۵ | فضای Word2Vec |
| ۸ | جمع‌بندی مشاهدات |
| ۱۰ | پیاده‌سازی خزنده، واکنشی اطلاعات مقالات و PAGERANK |
| ۱۰ | نحوه‌ی تقسیم وظایف |
| ۱۰ | نیما جمالی |
| ۱۰ | سپهر فعلی |
| ۱۰ | سینا کاظمی |
| ۱۱ | مراجع |

خوشه‌بندی

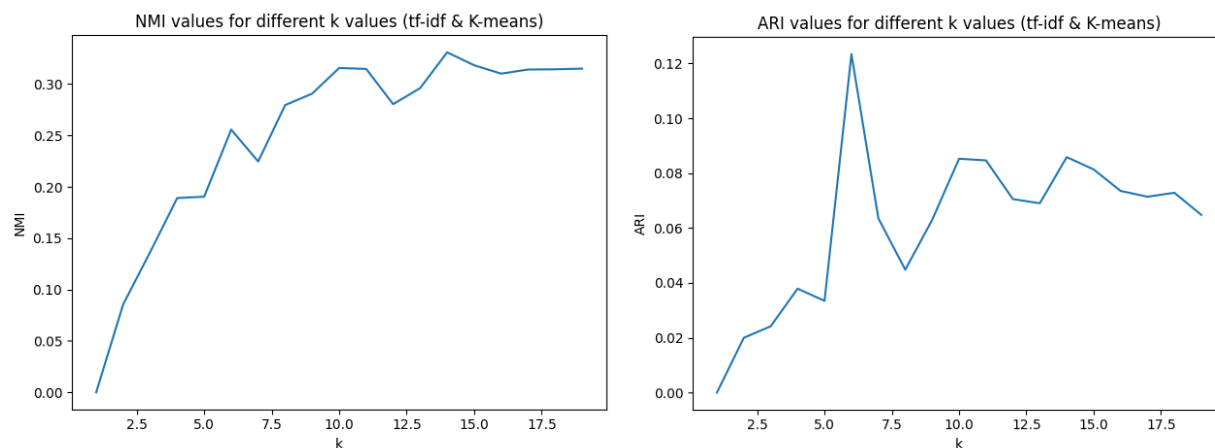
در این بخش، پس از پیش‌پردازش کلمات، دو فضای برداری TF-IDF و Word2Vec را ایجاد می‌کنیم و سپس خوشه‌بندی‌های K-means، GMM و سلسله مراتبی جمع‌کننده را انجام می‌دهیم.

پیش‌پردازش اخبار

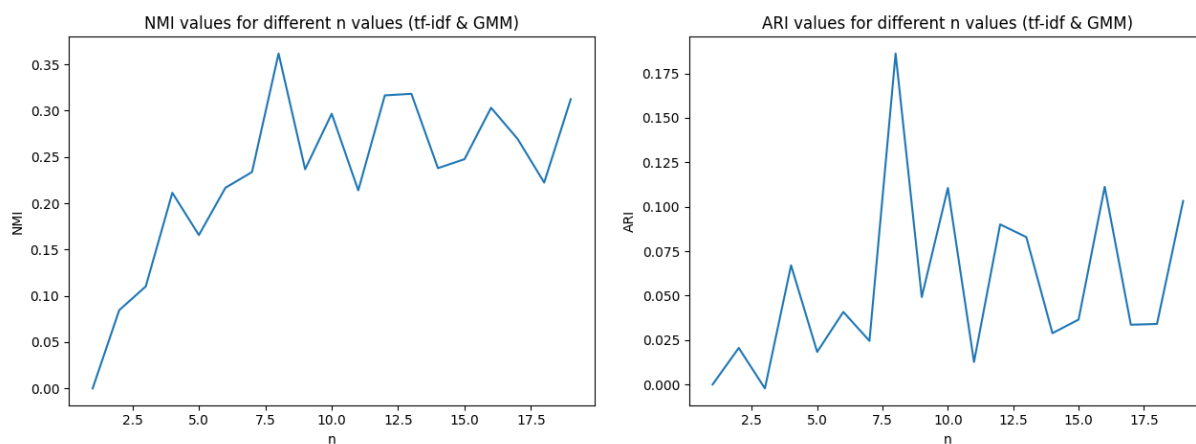
ابتدا کلاسی به نام News می‌سازیم که هر خبر را به آبجکتی از این کلاس تبدیل می‌کنیم و تمامی اخبار را به صورت یک آرایه از این کلاس ذخیره می‌کنیم. سپس در تابع initialize_data برای هر خبر، قسمت موضوع و خلاصه‌ی آن را جدا می‌کنیم و به تابعی که در فاز اول پروژه برای پیش‌پردازش متون فارسی استفاده شد، پاس می‌دهیم. از آنجا که TfidfVectorizer مستندات را به صورت یک رشته و Word2Vec مستندات را به صورت آرایه‌ای از کلمات جدا شده ورودی می‌گیرند، ورودی مناسب هر کدام در این تابع محاسبه می‌شوند.

فضای TF-IDF

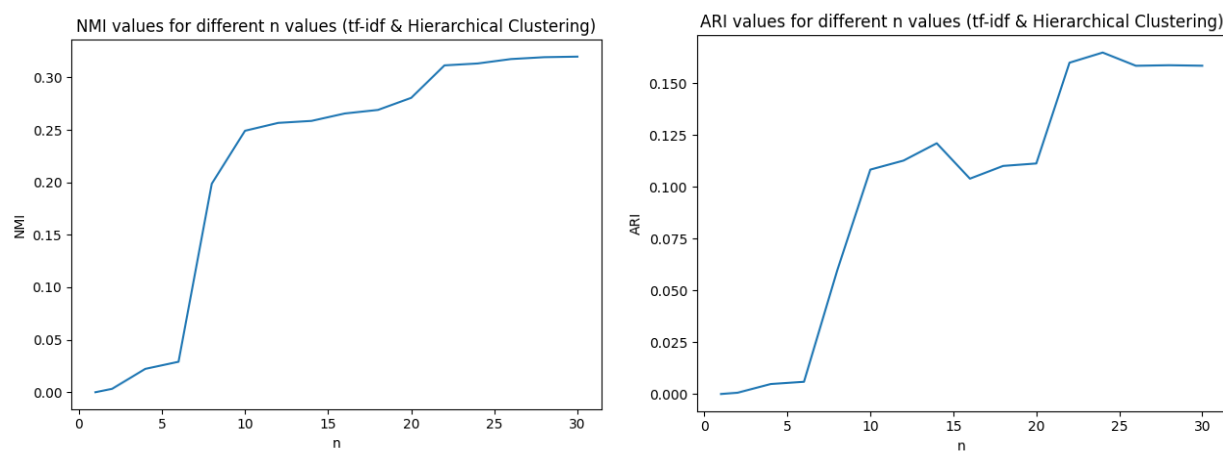
تابع tf_idf_initializer خروجی تابع قسمت قبل را به عنوان ورودی می‌گیرد و با استفاده از کتابخانه‌ی آماده‌ی پایتون، فضای برداری مد نظر را ایجاد می‌کند. توابع موردنیاز برای خوشه‌بندی‌ها نیز با استفاده از کتابخانه‌ی sklearn پیاده شده‌اند. دو معیار ارزیابی ARI و NMI برای ارزیابی خوشه‌بندی‌های مختلف استفاده شده‌اند. برای هر یک از دسته‌بندها یک تابع با نام tf_idf_[cluster's name] ایجاد شده است که یک متغیر boolean ورودی می‌گیرد و در صورتی که مقدار آن True باشد، بهترین پارامتر را محاسبه می‌کند. بهترین پارامتر را پارامتری در نظر گرفتیم که بیشترین مقدار ARI را تولید می‌کند. سپس حاصل این دسته‌بندی را ذخیره می‌کند، و در صورتی که boolean گفته شده برابر با False باشد، با توجه به اینکه قبلاً بهترین پارامتر به دست آمده و حاصل ذخیره شده است، حاصل خوشه‌بندی را لود می‌کند و همان طور که در صورت پروژه گفته شده، در یک فایل csv به همراه لینک خبر ذخیره می‌کند. این فایل‌ها به پیوست پروژه ارسال شده‌اند. البته باید اشاره کرد که چون ابعاد ماتریس GMM برای پردازش توسط خوشه‌بندی GMM بزرگ بود، با استفاده از PCA کاهش بعد اعمال می‌شود و سپس پردازش صورت می‌گیرد. همچنین با توجه به این که خروجی GMM هر بار متفاوت است، بهترین نتیجه بعد از چند بار اجرا ذخیره شده است. در خوشه‌بندی سلسله مراتبی هم average linkage عملکرد بهتری نسبت به سایر موارد داشت. علاوه بر این تابعی به نام visualize وجود دارد که نتایج هر خوشه‌بندی را در دو بعد مجسم‌سازی می‌کند، که نتایج آن در تصاویر بعدی قابل مشاهده‌اند.



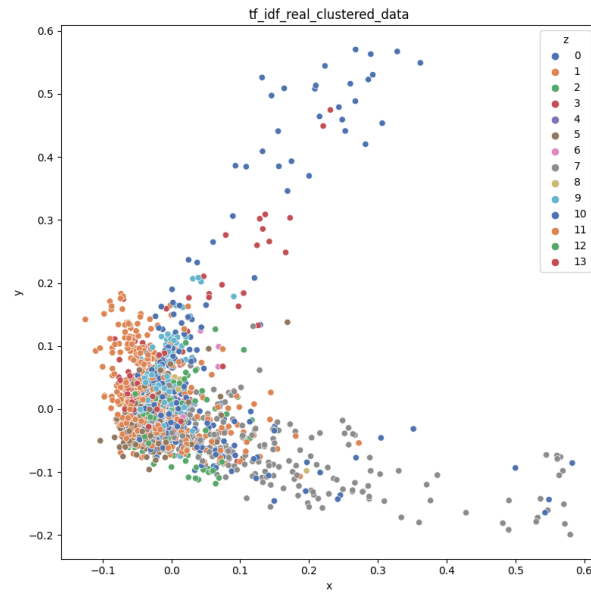
تصویر ۱: نمودار ARI و NMI برای پارامترهای مختلف، تحت خوشه‌بندی K-means



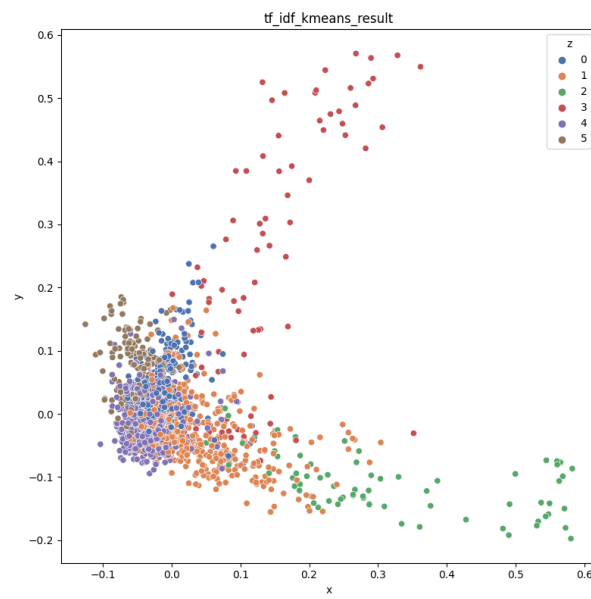
تصویر ۲: نمودار ARI و NMI برای پارامترهای مختلف، تحت خوشه‌بندی GMM



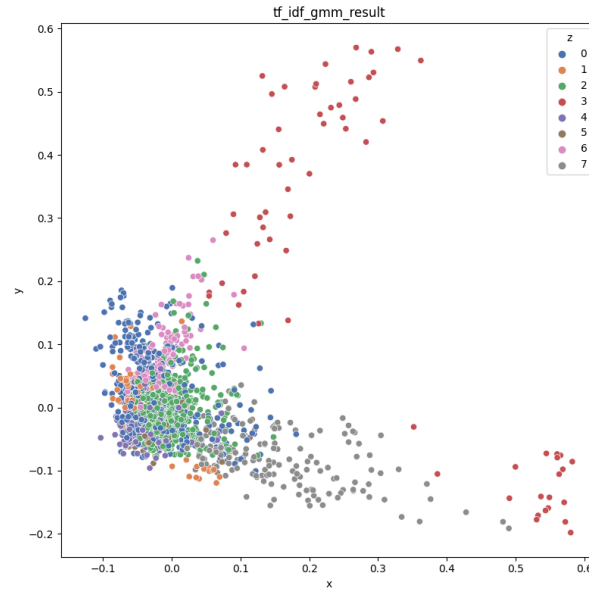
تصویر ۳: نمودار ARI و NMI برای پارامترهای مختلف، تحت خوشه‌بندی Hierarchical



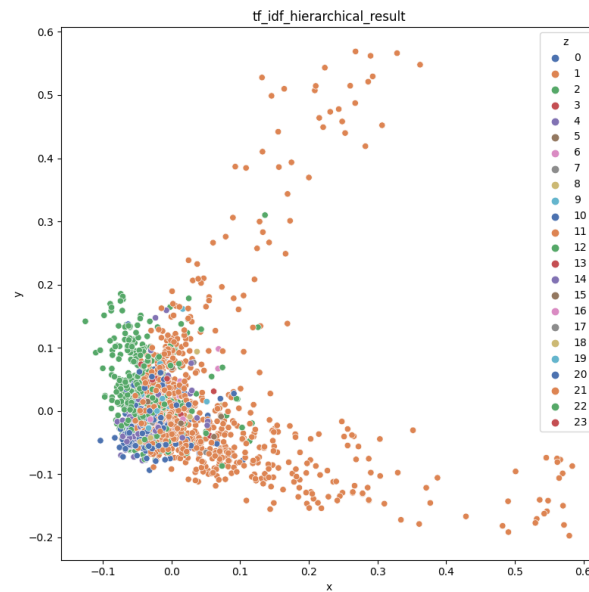
تصویر ۴: خوشه‌بندی اصلی اخبار (۱۴ دسته)



تصویر ۵: خوشه‌بندی اخبار با K-means (۶ دسته)



تصویر ۶: خوشه‌بندی اخبار با GMM (۸ دسته)

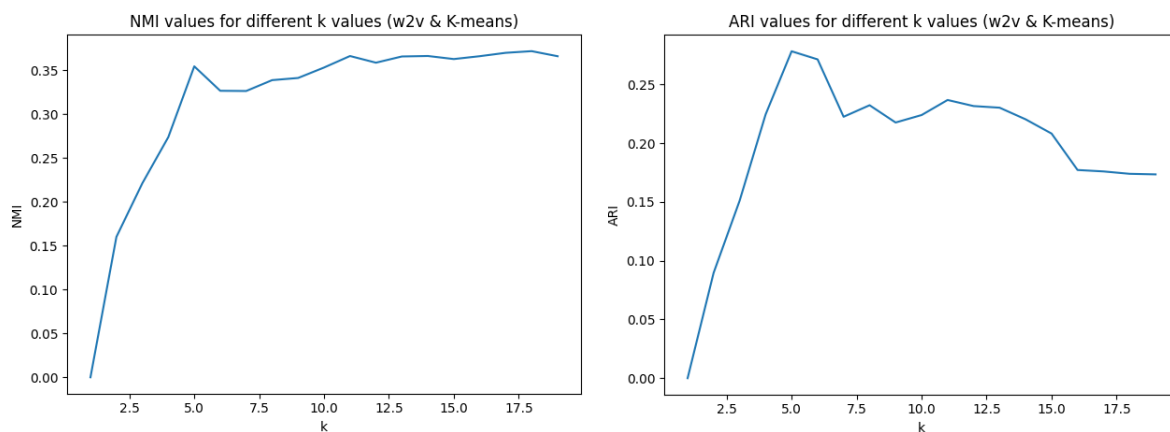


تصویر ۷: خوشه‌بندی سلسله مراتبی اخبار (۲۴ دسته)

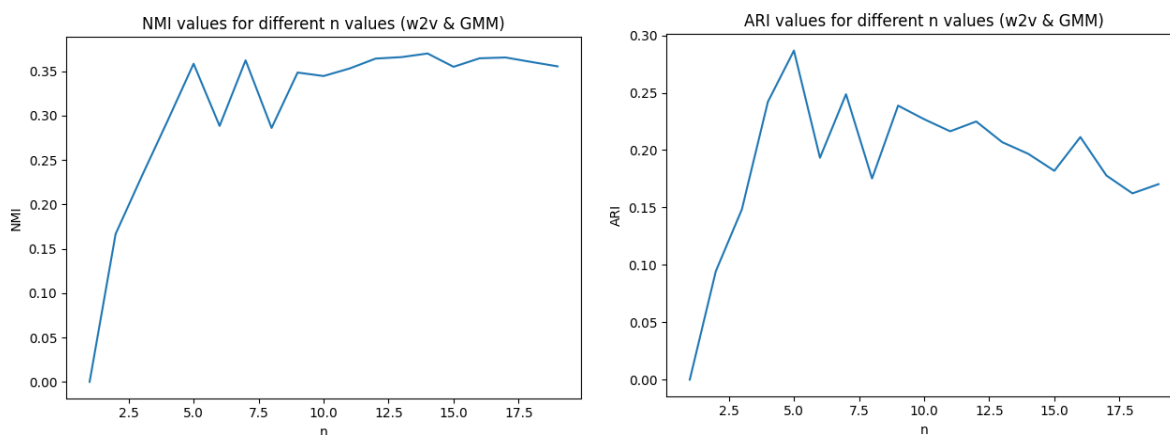
فضای Word2Vec

این بخش نیز مشابه بخش قبل پیاده شده است و تعداد iterationها برابر با ۱۰۰ اختیار شده است. مانند بخش قبل از توابع مشابه برای به دست آوردن بهترین پارامترها استفاده می‌کنیم. البته شایان ذکر است که دیگر برای GMM نیازی

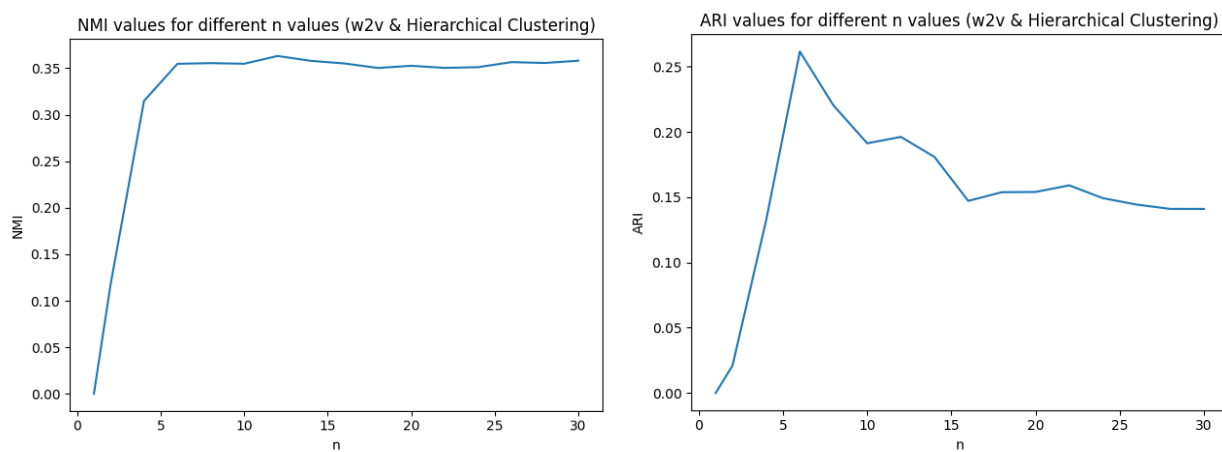
به کاهش بعد نداریم و همچنین ward linkage بهترین نتیجه را برای الگوریتم خوشه‌بندی hierarchical به دست می‌دهد.



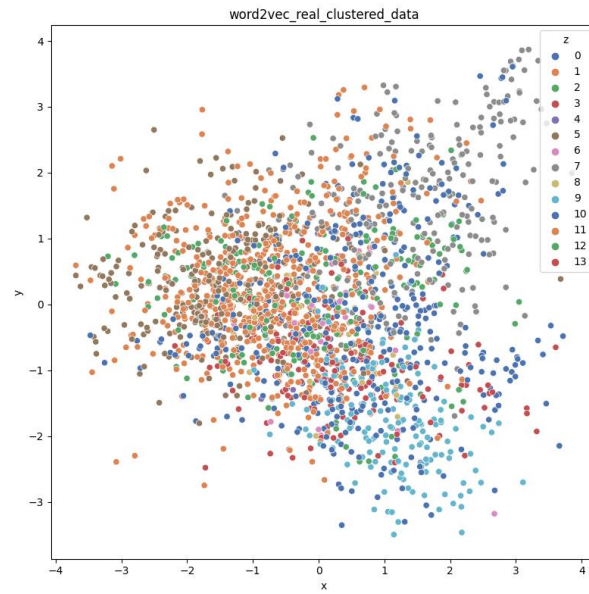
تصویر ۸: نمودار NMI و ARI برای پارامترهای مختلف، تحت خوشه‌بندی K-means



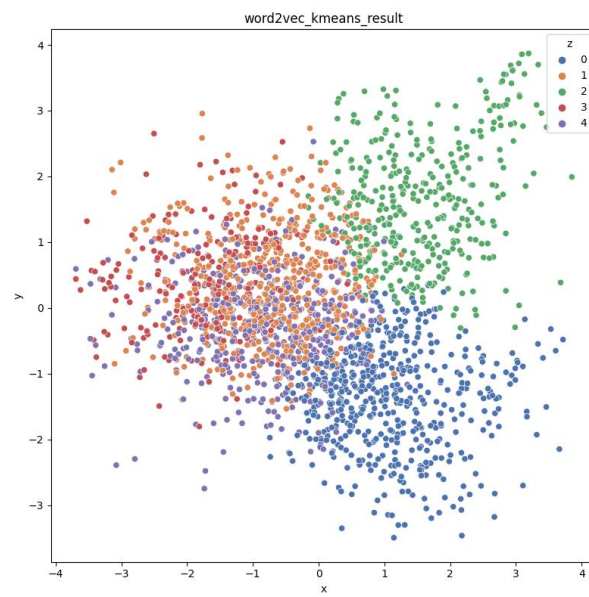
تصویر ۹: نمودار NMI و ARI برای پارامترهای مختلف، تحت خوشه‌بندی GMM



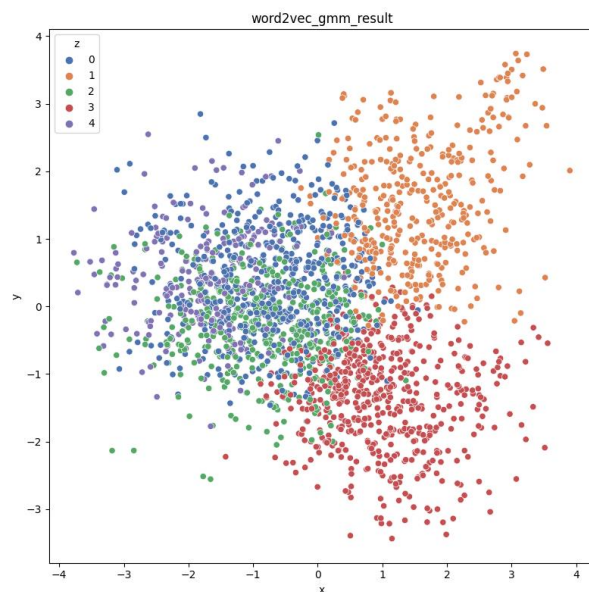
تصویر ۱۰: نمودار NMI و ARI برای پارامترهای مختلف، تحت خوشه‌بندی Hierarchical



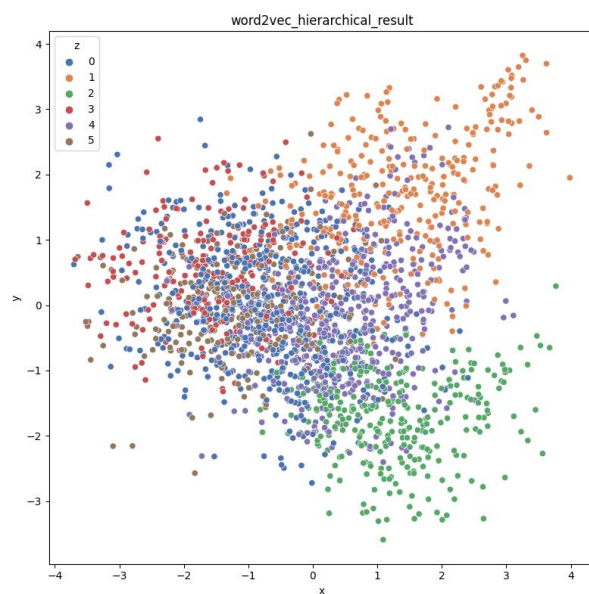
تصویر ۱۱: خوشه‌بندی اصلی اخبار (۱۴ دسته)



تصویر ۱۲: خوشه‌بندی اخبار با K-means (۵ دسته)



تصویر ۱۳: خوشه‌بندی اخبار با GMM (۵ دسته)



تصویر ۱۴: خوشه‌بندی سلسله مراتبی اخبار (۶ دسته)

جمع‌بندی مشاهدات













همان‌طور که در نمودارهای قسمت‌های قبل مشخص است، عملکرد Word2Vec بهتر از TF-IDF بوده و معیارهای ارزیابی آن بالاترند. در میان دسته‌بندها، ARI دسته‌بند GMM بهتر به نظر می‌رسد، ولی باید توجه داشت که این نتیجه بعد از چند بار اجرا به دست آمده است، در حالی که خوشه‌بندی سلسله مراتبی با average linkage در فضای برداری

tf-idf همیشه ARI برابر با ۱۶ درصد دارد که از میانگین اجراهای GMM بهتر است. همچنین از آنجا که دسته‌بندی اصلی تصاویر احتمالا خیلی دقیق نبوده، معیارها نسبتا پایین هستند اما در مجموع عملکرد فضای Word2Vec بهتر بوده است. در جدول زیر بهترین پارامترها برای هر خوشه‌بندی و دو معیار ارزیابی آنها به‌ازای آن پارامترها آمده است.

| | K-means | | | GMM | | | Hierarchical | | |
|--------|---------|--------|--------|--------|--------|--------|--------------|--------|--------|
| | Params | ARI | NMI | Params | ARI | NMI | Params | ARI | NMI |
| TF-IDF | k = 6 | 0.1234 | 0.2557 | n = 8 | 0.1862 | 0.3618 | n = 24 | 0.165 | 0.3133 |
| W2V | k = 5 | 0.2777 | 0.351 | n = 5 | 0.2867 | 0.3583 | n = 6 | 0.2616 | 0.3547 |

جدول ۱: خلاصه‌ی نتایج مشاهده شده برای بهترین پارامترها

خروجی کل این بخش نیز در پوشه‌ی phase3_outputs ذخیره شده است.

| | | | |
|--|-------------------|-----------------------|--------|
|  tf-idf-agglomerative | 2/10/2021 5:52 PM | Microsoft Excel Co... | 300 KB |
|  tf-idf-agglomerative | 2/8/2021 6:30 PM | KMP - MPEG Movi... | 18 KB |
|  tf-idf-gmm | 2/10/2021 5:44 PM | Microsoft Excel Co... | 299 KB |
|  tf-idf-gmm | 2/6/2021 8:49 PM | KMP - MPEG Movi... | 18 KB |
|  tf-idf-kmeans | 2/10/2021 5:40 PM | Microsoft Excel Co... | 299 KB |
|  tf-idf-kmeans | 2/8/2021 5:04 PM | KMP - MPEG Movi... | 9 KB |
|  w2v-agglomerative | 2/10/2021 6:27 PM | Microsoft Excel Co... | 299 KB |
|  w2v-agglomerative | 2/8/2021 6:11 PM | KMP - MPEG Movi... | 18 KB |
|  w2v-gmm | 2/10/2021 6:24 PM | Microsoft Excel Co... | 299 KB |
|  w2v-gmm | 2/8/2021 5:34 PM | KMP - MPEG Movi... | 18 KB |
|  w2v-kmeans | 2/10/2021 6:19 PM | Microsoft Excel Co... | 299 KB |
|  w2v-kmeans | 2/8/2021 6:40 PM | KMP - MPEG Movi... | 9 KB |

تصویر ۱۵: خروجی‌های CSV خوشه‌بندی‌ها

پیاده‌سازی خزنده، واکشی اطلاعات مقالات و PageRank

این بخش از مستند در فایل ژوپیتر [crawl.ipynb](#) که ضمیمه شده‌است، آمده است.

نحوه‌ی تقسیم وظایف

وظایف اختصاص یافته به هر فرد به شرح زیر بود:

نیما جمالی

- ۱- پیاده‌سازی فضاهاى TF-IDF و Word2Vec
- ۲- یافتن بهترین پارامترها
- ۳- پیاده‌سازی الگوریتم‌های خوشه‌بندی و معیارهای ارزیابی
- ۴- رسم نمودارها بر حسب پارامترها

سپهر فعلی

- ۱- پیش‌پردازش متون فارسی
- ۲- مجسم‌سازی داده‌ها
- ۳- ذخیره فایل‌ها به فرمت CSV
- ۴- مرتب‌سازی بر اساس PageRank

سینا کاظمی

- ۱- پیاده‌سازی خزشگر
- ۲- ذخیره‌سازی فایل json
- ۳- بررسی عدم تکرار در مقالات
- ۴- پیاده‌سازی تابع PageRank

مراجع

- [1] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
- [2] <https://radimrehurek.com/gensim/models/word2vec.html>
- [3] <https://en.m.wikipedia.org/wiki/PageRank>