بازیابی پیشرفته‌ی اطلاعات نیم‌سال اول ۱۳۹۹ـ۱۳۹۸

گزارش فاز سوم پروژه

مدرس: دکتر بیگی اعضای گروه: احسان سلطان‌آقایی، وحید بالازاده مرشت

مشارکت وحید بالازاده

درصد مشارکت: ۵۰ درصد.

کارهای انجام‌گرفته:

- بخش دوم: در این بخش با استفاده از کتابخانه‌ی BeautifulSoup به پارس کردن صفحات html پرداخته شد. ضمنا با استفاده از کلاس Cache در فایل crawler.py خروجی‌ها به صورت pickle ذخیره شدند تا دوباره خزش انجام نشود. حاصل همه‌ی مقالات در فایل article.json قرار دارد.
- بخش سوم: در این بخش ابتدا با استفاده از کتابخانه‌ی networkx یک گراف روابط ساخته شده و سپس روی این گراف الگوریتم page rank از کتابخانه‌ی networkx اجرا شد. مهم‌ترین مقاله طبق این معیار مقاله‌ای با عنوان Normalized Cuts and Image Segmentation بود.

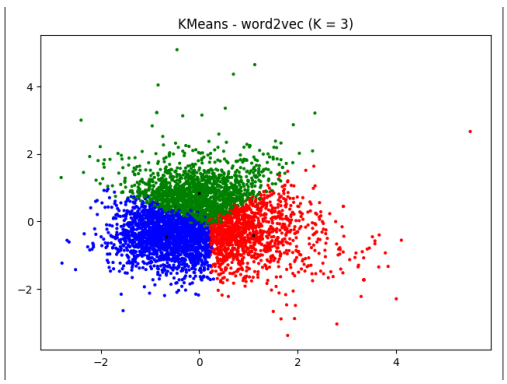
مشارکت احسان سلطان‌آقایی

درصد مشارکت: ۵۰ درصد.

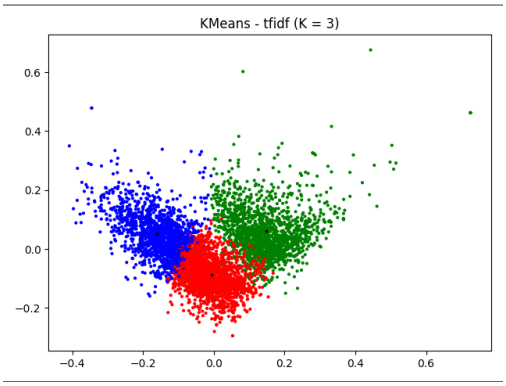
کارهای انجام‌گرفته:

بخش اول: پس از بررسی تعداد خوشه‌های متفاوت، ۳ خوشه در نظر گرفته شد و خوشه‌بندی‌های انجام‌شده به صورت زیر هستند:

- مدل k-means : در شکل‌های زیر نتیجه خوشه‌بندی به این روش با استفاده از تبدیل‌کننده‌های به فضای برداری مختلف قابل مشاهده است:

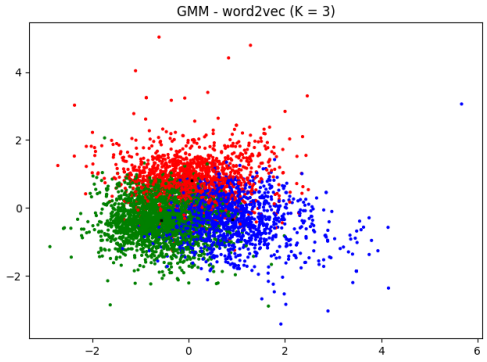


KMeans with Word2Vec (ب)

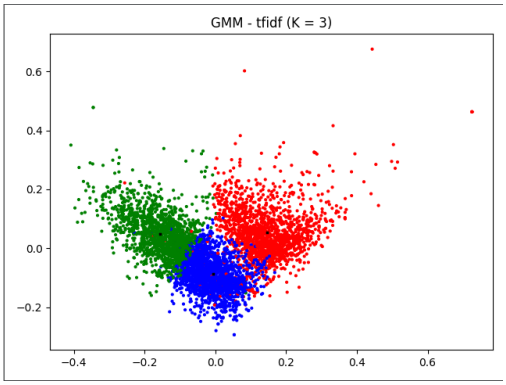


KMeans with TfIdf (آ)

- مدل Guassian Mixture Model : در شکل‌های زیر نتیجه خوشه‌بندی به این روش با استفاده از تبدیل‌کننده‌های به فضای برداری مختلف قابل مشاهده است:



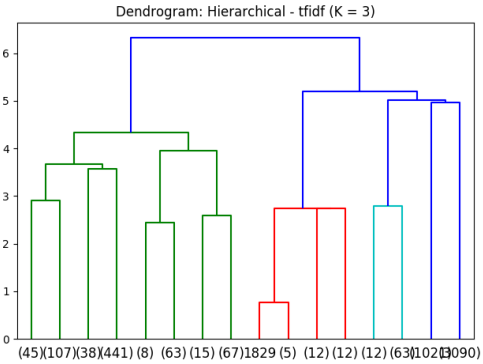
GMM with Word2Vec (ب)



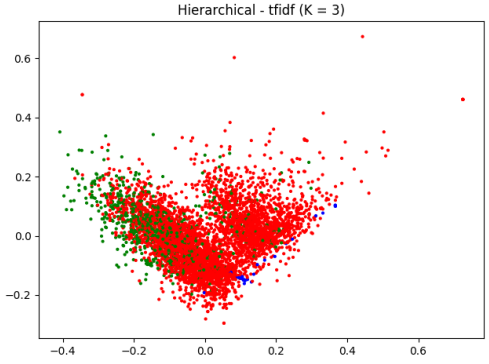
GMM with TfIdf (آ)

برای این مدل، دو نوع کواریانس full و diag بررسی شدند و در نهایت نوع full مورد استفاده قرار گرفت.

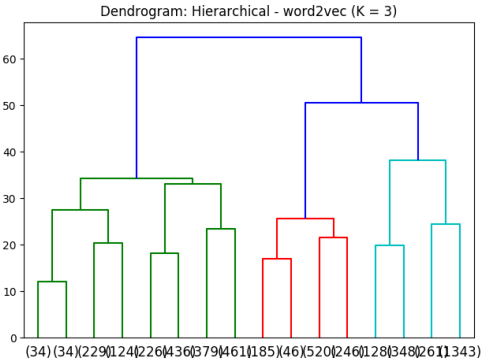
- مدل Hierarchical Clustering : در شکل‌های زیر نتیجه خوشه‌بندی به این روش با استفاده از تبدیل‌کننده‌های به فضای برداری مختلف قابل مشاهده است:



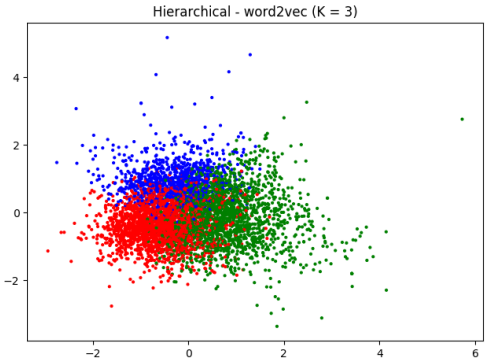
Dendrogram with TfIdf (ب)



Hierarchical with TfIdf (آ)



Dendrogram with Word2Vec (د)



Hierarchical with Word2Vec (ج)