به نام خدا

گزارش پروژه پایانی درس مدل‌های گرافی احتمالاتی

(پیاده‌سازی Box 19.A – خوشه‌بندی داده‌های MSNBC)

فرزانه فخریان – ۹۶۷۲۵۰۰۸

استاد درس: جناب آقای دکتر رحمانی

بهمن ماه ۱۳۹۷

کتاب‌خانه bnpy

Bnpy کتابخانه‌ای برای پایتون ۲/۷ می‌باشد که به منظور خوشه‌بندی شبکه‌ی بیزی از آن می‌توان بهره برد. که دارای مدهای مختلفی برای آموزش داده‌های کامل[۱] و غیر کامل[۲] است. به طور مثال در bnpy می‌توان از Mixture models ها، Hidden Markov models و الگوریتم‌های چون Expectation-maximization بر روی داده‌ها بهره برد. در این‌جا از این کتابخانه برای به‌دست آوردن ساختار موجود در داده‌ها یا به عبارتی به‌دست آوردن هم‌بستگی[۳] بین ویژگی‌های[۴] موجود در مجموعه داده به کمک الگوریتم EM، استفاده می‌شود.

برای استفاده از این کتابخانه باید مجموعه داده موجود که به فرمت seq می‌باشد به فرمت csv تبدیل شود. به همین دلیل در ابتدای کار مجموعه داده را به فرمت csv تبدیل می‌کنیم.

مجموعه داده

در مجموعه داده MSNBC هر سطر به عنوان یک کاربر[۵] در نظر گرفته شده است. به این صورت که در هر سطر دنباله‌ای[۶] از اعداد ارائه شده است که هر عدد نشان دهنده دسته‌بندی خاصی از صفحات است که کاربر به آن مراجعه کرده است. این صفحات به ۱۷ دسته به صورت زیر تقسیم می‌شوند:

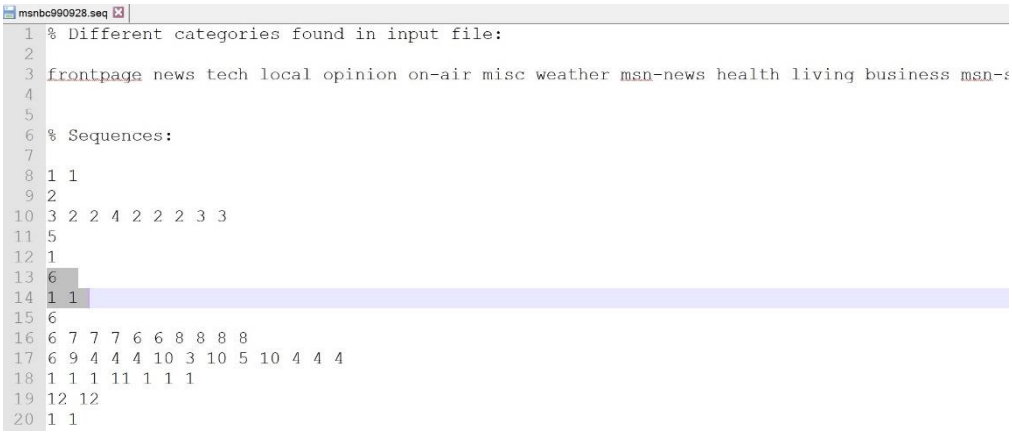
Frontpage, news, tech, local, opinion, on-air, misc, weather, msn-news, health, living, business, msn-sports, sports, summary, bbs, travel.

که اعداد ذکر شده در هر سطر (از ۱ تا ۱۷) هر کدام مربوط به دسته[۷] خاصی می‌باشد. به طور مثال در شکل زیر بخشی از این مجموعه داده، نشان داده شده است. در شکل زیر سطر اول نشان شده، به معنی یک بار

[1] Complete data
[2] Incomplete data
[3] covariance
[4] attribute
[5] user
[6] sequence
[7] category

مراجعه کاربر به دسته on-air و سطر دوم به معنی دو بار مراجعه کاربر به دسته frontpage می‌باشد. این مجموعه داده شامل ۹۸۹۸۱۸ دنباله می‌باشد. که از لینک زیر به دست آورده شد.

https://archive.ics.uci.edu/ml/datasets/msnbc.com+anonymous+web+data



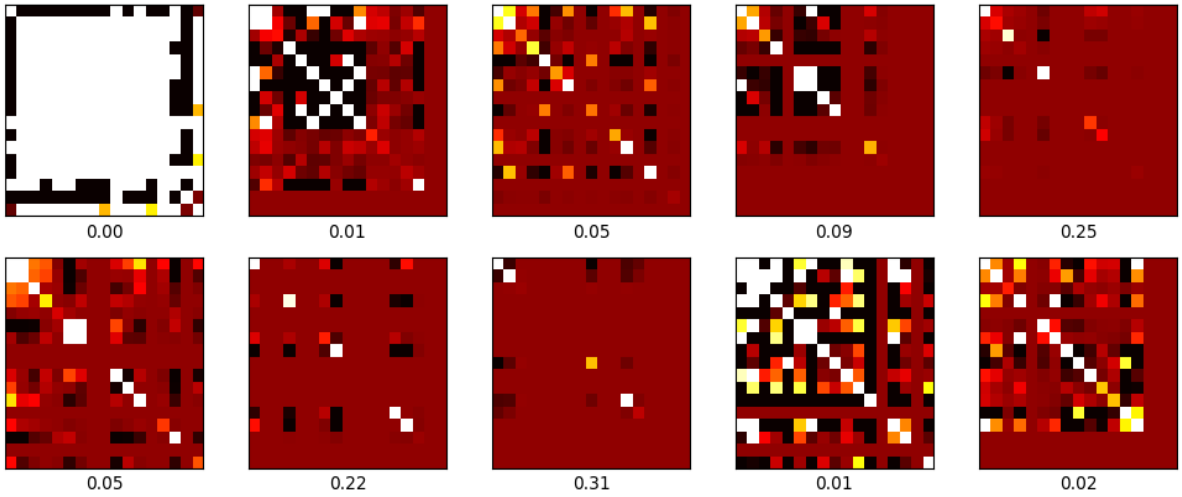
برای تبدیل این مجموعه داده به فرمت csv برای هر دسته یک ستون و برای هر کاربر مانند داده اصلی یک سطر در نظر گرفته شده و سپس تعداد هر دسته در هر سطر شمارش شده و برای دسته‌هایی که در یک سطر موجود نیستند، صفر در نظر گرفته می‌شود. که فایل این کد به همراه این گزارش موجود است اما داده به دست آمده به دلیل حجم بالاتر از حد مورد قبول سایت LMS در لینک گیت‌هاب این پروژه بارگذاری شده است.

سپس از این داده برای یافتن هم‌بستگی attribute ها، برای یافتن ساختار مناسب برای خوشه‌بندی استفاده می‌شود. نحوه نصب و پیش‌نیازهای مورد نیاز برای کتابخانه bnpy در لینک زیر موجود است.

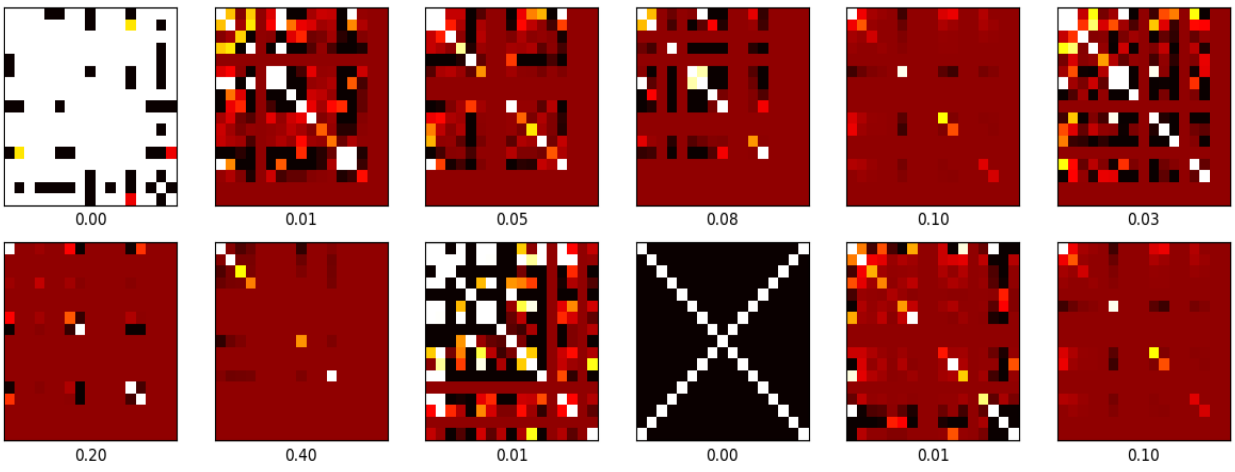
https://bnpy.readthedocs.io/en/latest

به کمک داده به دست آمده به فرمت csv در فایل plotCovMatrix.py به رسم ماتریس‌های کواریانس با تعداد تکرار بالا پرداخته تا الگوریتم به یک ساختار مناسب برای خوشه‌بندی همگرا شود. دراین تعداد داده که در حدود ۱ میلیون می‌باشد از ۵۰۰ بار تکرار استفاده شده است. که با توجه به الگوریتم و مدل انتخابی برای داده، زمان و تعداد تکرار مورد نیاز برای همگرا شدن متفاوت خواهد بود. که در این اجرا از الگوریتم EM استفاده می‌شود. (برای اجرا این بخش از سرور آزمایشگاه داده‌کاوی استفاده شده است زیرا به علت بالا بودن تعداد تکرارهای مورد نیاز و تعداد زیاد داده‌ها بر روی کامپیوتر شخصی زمان زیادی برای اجرا مورد نیاز بود و همچنین اجرای آن بر روی کامپیوترهای شخصی از سرعت پایینی برخوردار است.)

در این بخش با تعداد خوشه‌های مختلف (مقدار K در کد) به یافتن ماتریس‌های کواریانس پرداخته می‌شود. ماتریس‌های کواریانس برای هر خو شه، وابستگی attributeهای مختلف در خو شه‌ها را نشان می‌دهد. به طوری که اگر یک ویژگی در یک خوشه یافت نشود مقدار آن برابر صفر خواهد بود. (رنگ قرمز در شکل) در غیر این صورت مقداری بزرگتر از صفر خواهد داشت. (رنگ متمایل به سفید) در زیر چند نمونه از نتایج به‌دست آمده گزارش شده است و سپس در ادامه به تحلیل آن‌ها پرداخته خواهد شد. (نتایج کامل در گیت‌هاب پروژه گزارش شده است.)



۱۰ خوشه- ۱۰ ماتریس ۱۷*۱۷



۱۲ خوشه – ۱۲ ماتریس ۱۷*۱۷

همانطوری که در شـــکل‌های بالا مشـــاهده می‌کنید از این ماتریس‌های کواریانس می‌توان نتایجی را برای خوشه‌بندی این شبکه بیزی به‌دست آورد. این نتایج به شرح زیر است:

۱- هرچه تعداد خوشـــه‌ها را افزایش دهیم تعداد خوشـــه‌های که مقدار داده بســـیار کمی در آن‌ها قرار می‌گیرد زیادتر می‌شود در واقع تعداد خوشه‌هایی که نمی توانند نماینده خوبی برای نمایش درست این داده باشند افزایش می‌یابد.

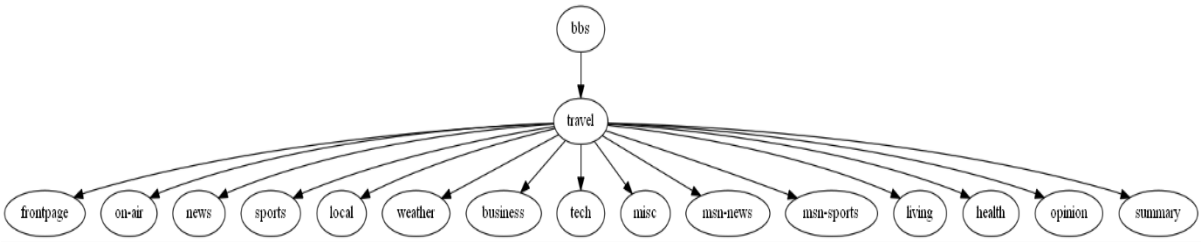
۲- علاوه بر این خو شه‌های بزرگ به خو شه‌های کوچک‌تر تقسیم می شوند به طوری که در شکل دوم دو خوشـه با ۱۰٪ داده وجود دارند که دارای ماتریس کوارایانس مشـابه هسـتند. که این ۲ خوشـه باید با یکدیگر ترکیب شوند.

۳- همانطوری که در شکل مشاهده می‌شوند این داده دارای ۴ خوشه اصلی می‌باشد که اکثریت داده را در خود جا داده‌اند.) که در کتاب نیز به این موضوع برای این مجموعه داده اشاره شده است.

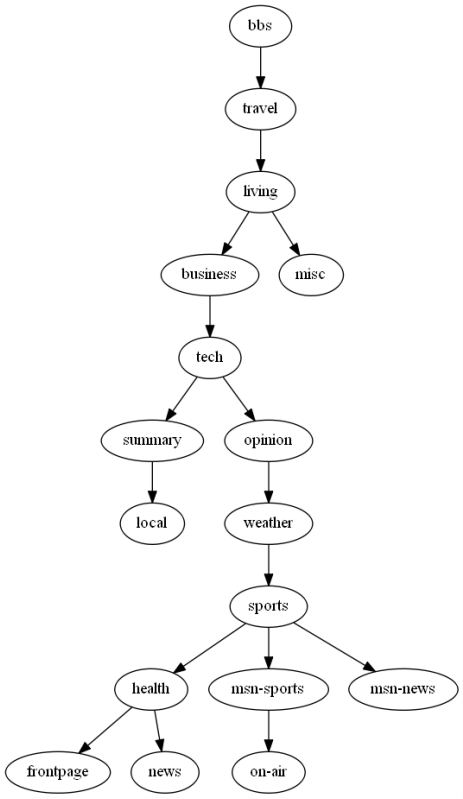
۴- با ا ستفاده از ۴ خو شه ا صلی می‌توان ساختار هر خو شه را (به معنی اینکه داده‌های موجود در هر خوشه شامل چه ویژگی‌هایی می‌باشند یا چه ویژگی‌هایی در ساختار شبکه بیزی به یکدیگر مرتبط هستند، را به دست آورد.) یافت. (به طور مثال در شکل دوم خوشه ای که ۴۰ درصد داده را در خود جا داده اســت همانطوری که مشـاهد می‌شـود، این داده شـامل ویژگی‌های شـماره ۱، ۲، ۳ و ۱۲ می‌باشـد که هر کدام به frontpage، news، tech و business اشـاره دارد که می‌توان آن را بردار نمایانگر یک خوشه دانست. که در کتاب نیز به خوشه‌ای مانند این مثال اشاره شده است.)

در ادامه با به دست آوردن بردارهای نمایانگر خوشه‌های اصلی از آن‌ها برای خوشه‌بندی استفاده خواهد شد.

ســـپس در ادامه با فرض کامل بودن داده با اســـتفاده از الگوریتم hillClimbing در کتابخانه pgmpy سـاختار شـبکه بیزی را به‌دسـت می‌آوریم. برای این کار ابتدا header های هر سـتون (attribute های شبکه) را به سطر ابتدایی داده csv ا ضافه می‌کنیم تا نتایج گراف به‌د ست آمده قابل فهم با شد. بع ضی از سطرهای داده شامل تعداد بسیار زیادی صفحه در یک سطر می‌باشند که این نقاط مانند نقاط outlier در این داده عمل می‌کنند که برای حذف آن‌ها در ابتدا سطرهایی که جمع مقادیر آن‌ها از ۲۰۰ بیشتر با شد را حذف می‌کنیم. ابتدا برای به دســـت آوردن ســـاختار، فرکانس attribute ها را نیز در نظر می‌گیریم که ساختار به دست آمده به صورت زیر خواهد بود که ساختار درستی نمی‌باشد.



اما در ادامه با توجه به گفته کتاب، حضور یا عدم حضور یک دسته از اهمیت برخوردار است نه تعداد رخداد آن؛ فرکانس داده‌ها را حذف کرده و هر مقداری که یک یا بیشتر از یک بار رخ داده باشد را یک در نظر می‌گیریم که مجموعه داده را به یک مجموعه داده باینری تبدیل می‌کند بعد از این تغییر ساختار شبکه بیزی به دست آمده به صورت زیر خواهد بود. که برای به دست آوردن خوشه‌ها از اهمیت بالایی برخوردار است.



همانطوری که مشاهده می شود نتایج به دست آمده در ماتریس‌های کواریانس در این ساختار نیز مشاهده می‌شود (business و tech به هم مرتبط هستند.) در ادامه از این ساختار و نتایج به دست آمده در مرحله قبل برای به‌دست آوردن خوشه های اصلی استفاده می‌شود. (کد این بخش در فایل BayesianGraph.py موجود است.)

خوشه‌بندی

با مطالعه روش خو شه‌بندی با ا ستفاده از Naïve bayes، ارائه شده تو سط دکتر جیمز مکافری، (که در ادامه به توضیح آن پرداخته می‌شود) به خوشه‌بندی داده‌های این شبکه پرداختیم.

پس از حذف نقاط outlier و داده‌های باینری در کد، در این روش برای هر خوشـــه یک نماینده در نظر گرفته می شود که انتخاب این نماینده از اهمیت زیادی برخوردار ا ست زیرا اع ضای خو شه برا ساس میزان شباهت با این نماینده انتخاب می شوند که ما برای به د ست آوردن این نماینده‌ها از ساختار شبکه بیزین و ماتریس هم‌بستگی attribute ها استفاده می‌کنیم.

بعد از به د ست آوردن این نماینده‌ها، آن‌ها را به صورت یک بردار ۱۷ تایی تشکیل شده از صفر و یک (که یک‌ها نمایانگر ویژگی‌های موجود در آن خوشه هستند) به ابتدای داده csv اضافه می‌کنیم و تعداد خوشه‌ها را نیز در کد مشخص می‌کنیم. (در این جا به ۴ خو شه ا صلی اکتفا می‌کنیم و این تعداد را در کد، ۴ در نظر می‌گیریم.)

در این روش با توجه به تعداد مشـــخص شـــده برای خوشـــه‌ها، داده‌های ابتدای مجموعه داده به عنوان پرچمداران هر خوشه، انتخاب می‌شوند و سپس برای هر داده جدید احتمال شرطی هر خوشه به شرط داده جدید به صـورت جداگانه محاسـبه می‌شـود و داده به خوشـه‌ای که احتمال بیشـتری دارد اختصـاص داده می‌شود. روش محاسبه این احتمال به شرح زیر است:

بعد از اختصاص دادن پرچمداران به خوشه‌های مختلف، احتمال هر خوشه به شرط داده جدید برای هر کدام از خوشـه‌ها به صـورت جداگانه و براسـاس attribute هایی از آن که دارای مقدار یک هسـتند، محاسـبه می‌شـود. به طور مثال برای محاســبه احتمال در خوشـــه اول ($P(C0 \mid X)$)، متناســب با مقدار attributeهایی داده X، به ازای هر ویژگی تعداد نمونه‌هایی که با شـــرایط مشـــابه در این خوشـــه یافت می‌شود را بر تعداد کل نمونه‌های موجود در خو شه صفر تقسیم می‌شود. و این احتمالات به ازای هر ویژگی محا سبه شده و در هم ضرب می شوند. در این مرحله احتمال توام داده X و خو شه صفرم به د ست می‌آید سپس برای به د ست آوردن احتمال شرطی باید این مقدار بر مجموع همین مقدار در تمام خو شه‌ها تقسیم

شود. که فرمول‌های محاسبه آن به صورت زیر است: (به طور مثال اگر داده X دارای ۲ مقدار frontpage و on-air باشد)

$$P(C0 \mid X) = \frac{PP(C0 \mid X)}{PP(C0 \mid X) + PP(C1 \mid X) + PP(C2 \mid X) + PP(C3 \mid X)}$$

$$\begin{aligned} PP(C0 \mid X) = {} & P(frontpage(1) \mid C0) * P(on-air(1) \mid C0) \\ & * P(news(0) \mid C0) * P(tech(0) \mid C0) * \dots \\ = {} & \frac{count(frontpage(1) in\ C0)}{count\ (C0)} * \frac{count(on-air(1) in\ C0)}{count\ (C0)} \\ & * \frac{count(news(0) in\ C0)}{count\ (C0)} * \frac{count(tech(0) in\ C0)}{count\ (C0)} * \dots \end{aligned}$$

سپس پس از محا سبه این مقدار برای تمام خو شه‌ها، خو شه‌ای که دارای مقدار احتمال بی‌شتری با شد این داده به آن اخت صاص داده می شود و سپس count آن خو شه متنا سب با ویژگی‌های موجود در داده باید بروزرسانی شود.

دو نکته در این روش قابل ذکر است :

۱- برای ذخیره مقادیر count ها در این روش از یک ساختار داده به صورت آرایه‌ای ۳ بعدی (در این مثال ۲*۱۷*۴) اســتفاده می‌شــود به صــورتی که مقدار حضــور یا عدم حضــور هر ویژگی برای هر خوشه در آن ذخیره می‌شود و برای محاسبه احتمالات از آن بهره گرفته می‌شود.

۲- برای جلوگیری از تقسیم بر صفر در محاسبه احتمالات در این روش، از روش هموارسازی لاپلاس[8] ا ستفاده شده ا ست. به صورتی که در ابتدا ساختار داده ذکر شده در نکته اول با مقادیر اولیه یک، مقداردهی می‌شوند.

موارد ذکر شده در فایل DataClustering.py پیاده‌سازی شده است. برای آشنایی بیشتر با این روش می‌توان به لینک زیر مراجعه کرد:

https://msdn.microsoft.com/en-us/magazine/jj991980.aspx

[8] laplace smoothing

نتایج

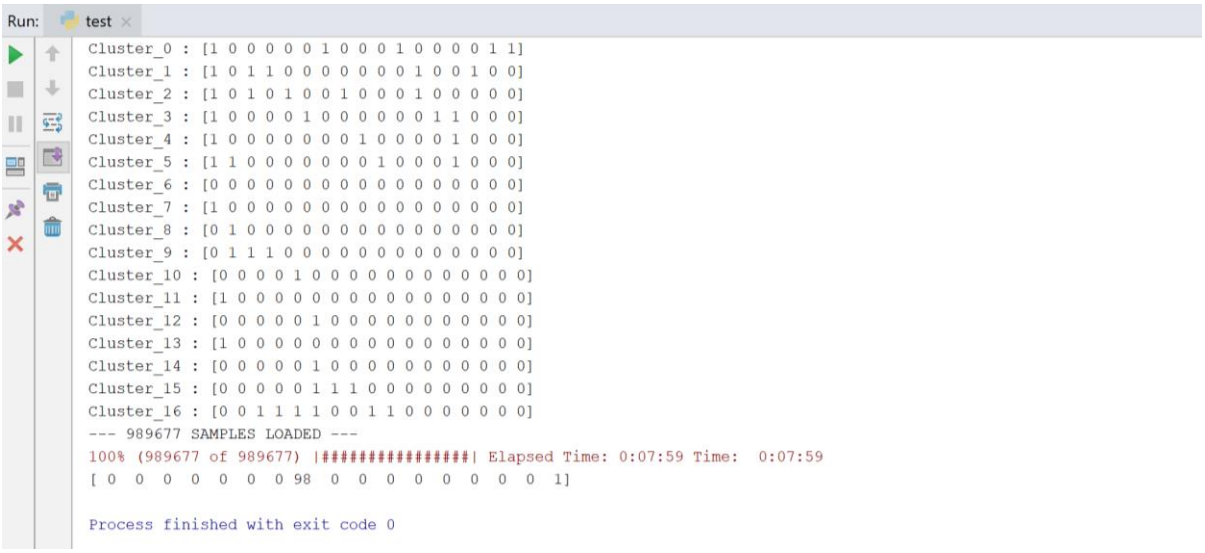
در انتها در یک فایل csv نتایج خروجی خوشه‌ها به ازای هر داده ورودی گزارش می‌شود. و مشخص می‌کند هر داده با استفاده از این خوشه‌بندی به کدام خوشه اختصاص داده شده است. و همچنین در console میزان درصدی که به هر خوشه اختصاص داده شده است مشخص می‌شود. نتیجه گزارش شده به نماینده خوشه‌ها وابسته است و با انتخاب نماینده‌های متفاوت می‌توان به خوشه‌بندی‌های متفاوتی دست یافت. به این صورت که در تقریبا ۹۹ درصد داده‌ها ویژگی frontpage به چشم می‌خورد. به همین دلیل درصورتی که نماینده‌ای فقط با یک رقم یک به ازای frontpage انتخاب شود اکثریت داده به این خوشه اختصاص داده می‌شود و دیگر خوشه‌ها خالی باقی می‌مانند در حالی که انتخاب این نماینده به عنوان نماینده یک خوشه، برای این داده مناسب نیست. (نمونه این نتیجه در تصاویر زیر گزارش شده است.). به چند نمونه از نتایج به دست آمده برای خوشه‌بندی این داده‌ها، در زیر اشاره شده است. ادامه نتایج با نماینده‌های دیگر در مخزن گیت‌هاب این پروژه گزارش شده است. (در هر نمونه نماینده‌های استفاده شده برای خوشه‌ها نیز ذکر شده است.)

```
test
Cluster_0 : [0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0]
Cluster_1 : [0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0]
Cluster_2 : [0 1 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0]
Cluster_3 : [0 0 0 1 1 0 1 1 0 1 1 0 0 0 1 1 1]
--- 989677 SAMPLES LOADED ---
 99% (989503 of 989677) |################ | Elapsed Time: 0:03:10 ETA:   0:00:00[12  7 52 27]
100% (989677 of 989677) |#################| Elapsed Time: 0:03:10 Time:  0:03:10

Process finished with exit code 0
```

نمونه خروجی ۱- در این نمونه ۹۸ درصد داده در ۴ خوشه اصلی قرار گرفته است.

```
test
Cluster_0 : [1 1 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0]
Cluster_1 : [1 1 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0]
Cluster_2 : [1 1 0 0 0 1 0 0 1 0 0 0 1 1 0 0 0]
Cluster_3 : [1 0 0 1 1 0 1 1 0 1 1 0 0 0 1 1 1]
--- 989677 SAMPLES LOADED ---
100% (989677 of 989677) |#################| Elapsed Time: 0:02:09 Time:  0:02:09
[ 7 50 16 25]

Process finished with exit code 0
```

نمونه خروجی ۲ – در این نمونه ۹۹ درصد داده در ۴ خوشه اصلی قرار گرفته است.



نمونه خروجی ۳ – همانطوری که در این خروجی مشخص است نماینده خروجی هشتم فقط شامل ویژگی frontpage و ۹۸ درصد داده فقط به این خوشه اختصاص داده شده است.

در این نمونه خروجی‌ها میزان در صد داده صفر به معنای این نیست که داده‌ای در آن‌ها وجود ندارد میزان درصد در این خروجی‌ها به صورت int گزارش شده است عدد صفر به این معنا است که تعداد نمونه موجود در آن‌ها کم تر از ۱ درصد است. زیرا با وجود انتخاب یک نماینده برای هر خوشه امکان خالی بودن خوشه‌ها وجود ندارد.

کتابخانه و ابزارهای دیگر

به غیر از کتابخانه‌های استفاده شده در این پروژه، کتابخانه‌های مفید دیگری که مرتبط با این پروژه و درس مدل‌های گرافی احتمالاتی که در حین این پروژه به مطالعه آن‌ها پرداخته شد به شرح زیر است : (لینک آن‌ها در زیر قرار داده شده است.)

کتابخانه pomegranate در پایتون برای پیاده‌سازی مدل‌های گرافی احتمالاتی:

https://pomegranate.readthedocs.io/en/latest

ابزار bayespy در پایتون برای شبکه‌های بیزین:

https://pypi.org/project/bayespy

لینک گیت‌هاب پروژه به شرح زیر است:

https://github.com/farzanefakhrian/MSNBC-Clustering