



دانشگاه شهید باهنر کرمان

دانشکده فنی و مهندسی

گروه مهندسی کامپیوتر

دستورالعمل پروژه داده کاوی

استاد راهنما

دکتر مصطفی قاضی زاده احسانی

محمد مجتبی روشنی

اردیبهشت ۱۴۰۱

راهنما

در این پروژه به دنبال استخراج و آنالیز دیتاست نظرات در مورد رستوران‌ها که توسط *Yelp* جمع‌آوری شده است هستیم تا به افراد در مورد تصمیم‌گیری در مورد غذا کمک کنیم. هر تسک یک گزارش به خصوص دارد و یک گزارش نهایی برای تمامی تسک‌ها مورد نیاز است.

- متون انگلیسی در جایی که منجر به درک بهتر مطلب می‌گردند، قید می‌شوند.
 - ترجمه فارسی ممکن است، برای درک بهتر، با ترجمه تحت الفظی متن انگلیسی متفاوت باشد.
 - تمامی تصاویر و فایل‌های مربوطه با کیفیت قابل قبول در دایرکتوری "assets" قابل مشاهده هستند.
 - در گزارشی که برای هر Task می‌نویسید، می‌بایست به ترتیب زیر:
 - I. تحت عنوان Objective توضیح دهید چگونه به هدف مطلوب رسیده‌اید.
 - II. تحت عنوان Tools توضیح دهید از چه ابزارهایی (کتابخانه، بسته، زبان برنامه‌نویسی و...) استفاده کرده‌اید.
 - III. تحت عنوان Other Possibilities توضیح دهید چه پیشنهادهای دیگری علاوه بر آنچه در این مقاله برای Task ذکر شده یا آنچه که آموخته‌اید، دارید.
 - IV. تحت عنوان Analysis توضیح دهید که چگونه مسئله را تحلیل و حل کرده‌اید.
- توجه کنید که برای تسک‌هایی که از چند قسمت تشکیل شده‌اند باید چهار مرحله بالا را در هر قسمت تکرار کنید. برای متوجه شدنِ طریقه نوشتن یک گزارش، برای Task سوم در مسیر
- "assets/templates/Task3_report.pdf" نمونه‌ای قرار داده شده است.
- گزارش نهایی را که شامل ارزیابی خلاصه‌ای از تمامی شش Task می‌باشد در فایل به فرمت Report-Final.pdf در مسیر اصلی پروژه قرار دهید.
 - مطلب زیر را در گزارش نهایی بررسی کنید:
 - I. ویژگی‌هایی از دیتاستی که در اختیارتان قرار داده شده است را ذکر کنید که مشتری‌ها و صاحبان رستوران به آن‌ها اهمیت می‌دهند و در Task‌ها ذکر نشده‌اند.
 - II. how does the review topic distribute for those
 1. frequent yelp users
 2. return customer
 - ترجیحاً از زبان برنامه‌نویسی پایتون استفاده کنید (استفاده از سایر زبان‌ها به شرط ارضای مسئله محدودیتی ندارد).

First Task: Data Exploration.....	5
۱-۱- هدف	۶
۱-۲- آماده سازی دیتا	۶
۱-۳- قسمت اول Task	۶
۱-۴- قسمت دوم Task	۱۱
۱-۵- خلاصه	۱۵
۱-۶- نتایج	۱۵
Second Task: Cuisine Clustering.....	16
۲-۱- توضیحات	۱۷
۲-۲- هدف	۱۷
۲-۳- قسمت اول Task (Visualization of the Cuisine Map)	۱۸
۲-۴- قسمت دوم Task (Improving the Cuisine Map)	۱۹
۲-۵- قسمت سوم Task (Incorporating Clustering in Cuisine Map)	۲۰
۲-۶- خلاصه	۲۱
۲-۸- نتایج	۲۱
Third Task: Dish Recognition	22
۳-۱- هدف	۲۳
۳-۲- آماده سازی	۲۳
۳-۳- قسمت اول Task (Manual Tagging)	۲۳
۳-۴- قسمت دوم Task (Mining Additional Dish Names)	۲۴
۳-۵- خلاصه	۲۵
۳-۶- نتایج	۲۵
Fourth Task: Mining Popular Dishes Report.....	26
۴-۱- هدف	۲۷
۴-۲- خلاصه	۲۸
۴-۳- نتایج	۲۸
Fifth Task: Restaurant Recommendation Report.....	29
5-1- هدف	۳۰
۵-۲- خلاصه	۳۲
۵-۳- نتایج	۳۲
Sixth Task: Hygiene Prediction	33

۳۴	۱-۶- هدف
۳۴	۲-۶- آماده سازی
۳۴	۳-۶- خلاصه
۳۴	۴-۶- نتایج
۳۶	منابع و مراجع

صفحه

فهرست اشکال

شکل ۱-۱	نمونه visualization اول	۸
شکل ۱-۲	نمونه visualization دوم	۹
شکل ۱-۳	نمونه visualization سوم	۱۰
شکل ۱-۴	نمونه visualization چهارم	۱۲
شکل ۱-۵	نمونه visualization پنجم	۱۳
شکل ۱-۶	نمونه visualization ششم	۱۴
شکل ۱-۷	نمونه visualization هفتم	۱۴

First Task:

Data Exploration

۱-۱- هدف

هدف از این تسک بررسی داده‌ها، برای دریافت حسی نسبت به شکل ظاهری ویا ویژگی داده‌ها است. به طور کلی با جواب دادن به سوالات زیر می‌توانید به این هدف فکر کنید:

--۱. موضوعات اصلی در این بررسی‌ها چیست؟

--1. What are the major topics in the reviews?

--۲. آیا آنها در بررسی‌های مثبت و منفی متفاوت هستند؟

--2. Are they different in the positive and negative reviews?

--۳. آیا آنها برای غذاهای مختلف متفاوت هستند؟

--3. Are they different for different cuisines?

--۴. توزیع تعداد نظرات بر سایر متغیرها (به عنوان مثال، آشپزی، مکان) چگونه است؟

--4. What does the distribution of the number of reviews over other variables (e.g., cuisine, location) look like?

--۵. توزیع رتبه بندی‌ها چگونه است؟

--5. What does the distribution of ratings look like?

این سوالات را بر اساس نظرات و داده‌های موجود در دیتاست می‌توانید پاسخ دهید.

۱-۲- آماده سازی دیتا

ابتدا دیتاست را از این [لینک](#) دانلود کرده و extract کنید.

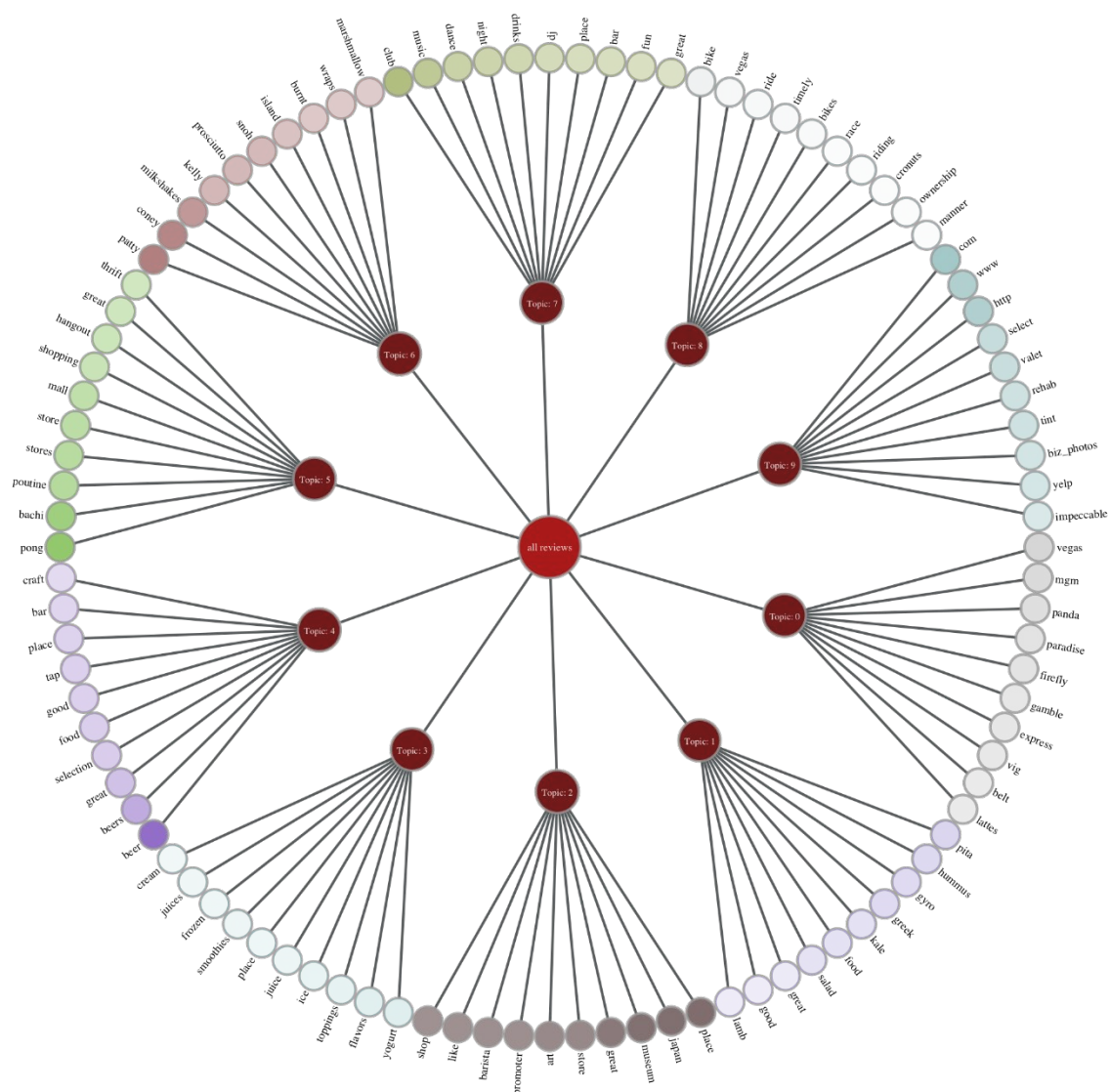
۱-۳- قسمت اول Task

از یک topic model برای استخراج موضوعات از تمامی متن نظرات و visualize کردن آنها استفاده کنید تا متوجه شوید مردم در چه موردی در این موضوعات صحبت کرده‌اند. برای مثال بعد از apply

کردن Latent Dirichlet Allocation بر روی نمونه‌ای از نظرات visualization زیر را بدست آورده‌ایم در اینجا میزان هر گره با وزن آن در هر موضوع مطابقت دارد.

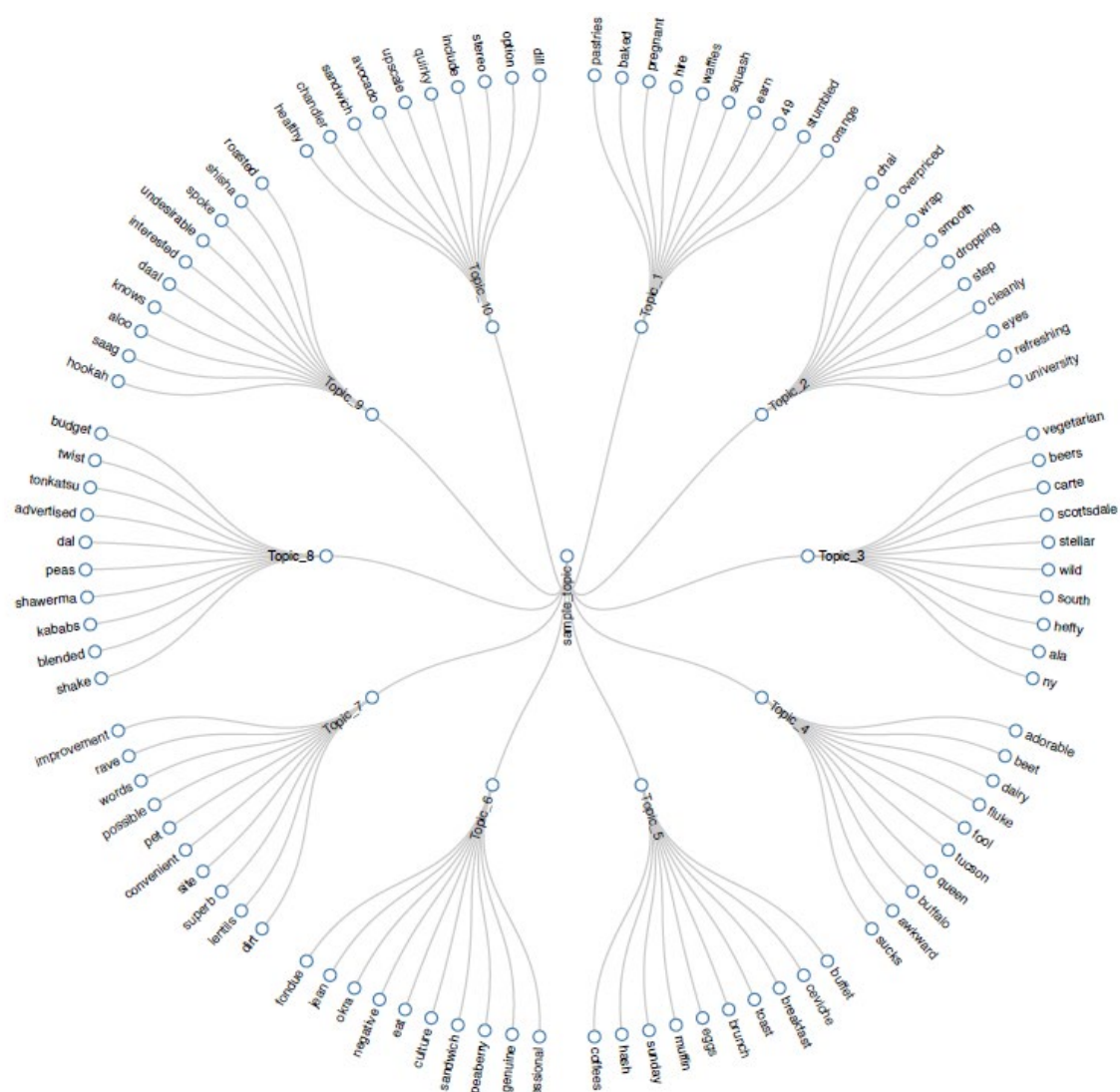
Use a topic model (e.g., PLSA or LDA) to extract topics from all the review text (or a large sample of them) and visualize the topics to understand what people have talked about in these reviews. For example, after applying LDA to a sample of the reviews, we obtained the following visualization. Here the opacity of each node corresponds to its weight in each topic.

First Task: Data Exploration

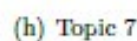
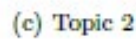
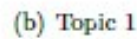


شکل ۱-۱ نمونه visualization اول.

(The intensity of color depends on weight of each word in the topic)



شكل ٢-٢ نمونه visualization دوم.

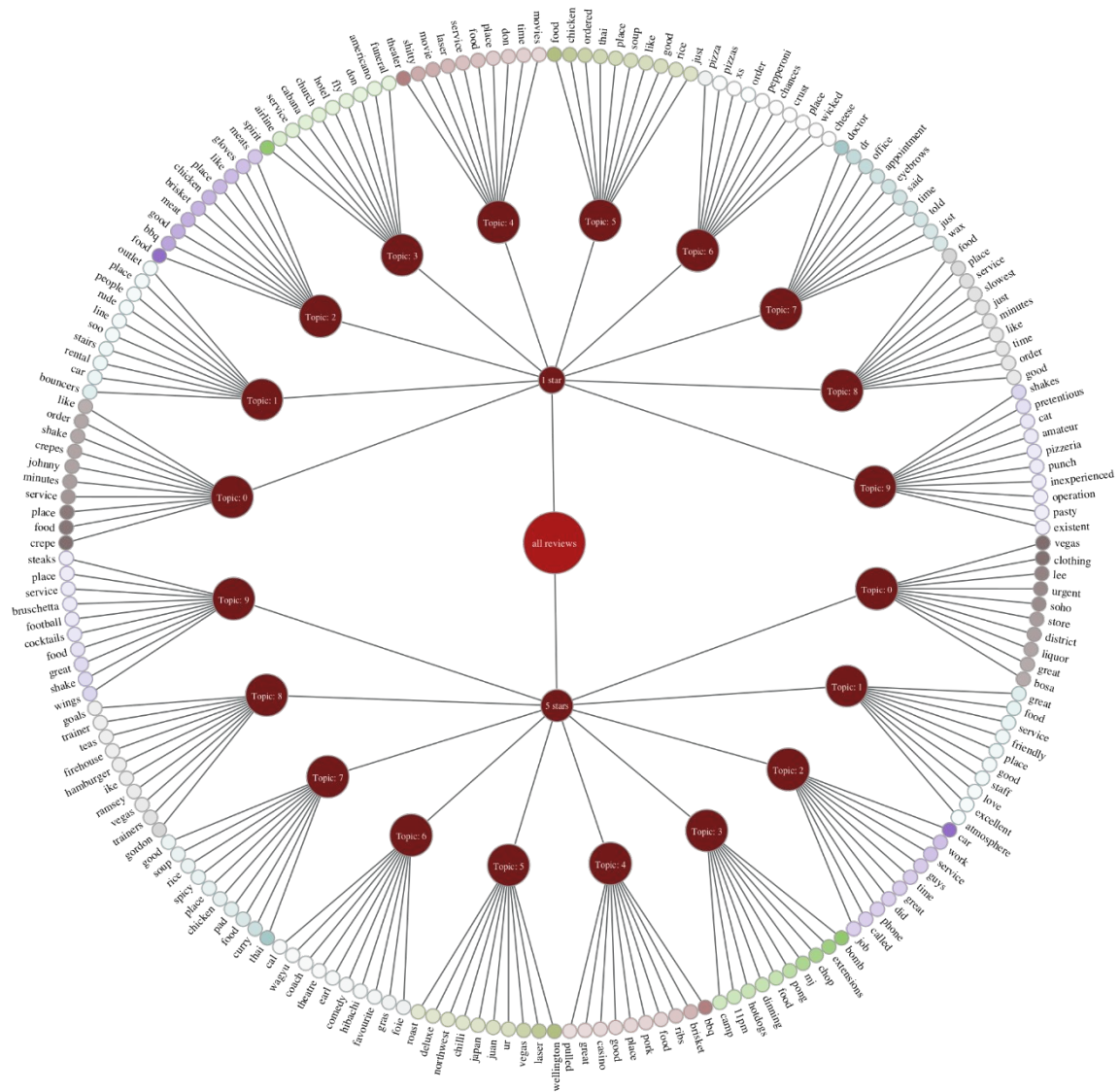


شکل ۳-۳ نمونه visualization سوم.

۴-۱ - قسمت دوم Task

همین کار را برای دو زیر مجموعه از نظرات که قابل مقایسه هستند (مثلا نظرات مثبت و منفی برای یک غذا یا رستوران خاص) انجام دهید و به صورت بصری موضوعات استخراج شده از این دو زیر مجموعه را برای کمک به درک شباهت و تفاوت بین موضوعات استخراج شده، مقایسه کنید. این ۲ زیر مجموعه را به دلخواه تشکیل دهید. در اینجا یک visualization برای نمونه‌ای از نظرات با رتبه‌های بالا و پایین را نشان می‌دهیم.

Do the same for two subsets of reviews that are interesting to compare (e.g., positive vs. negative reviews for a particular cuisine or restaurant), and visually compare the topics extracted from the two subsets to help understand the similarity and differences between these topics extracted from the two subsets. You can form these two subsets in any way that you think is interesting. Here we show a sample visualization for a sample of reviews with high and low ratings .





شکل ۴-۴ نمونه visualization چهارم.

First Task: Data Exploration



شکل ۵-۵ نمونه visualization پنجم.

First Task: Data Exploration

 <p>(a) Topic 0</p> <p>(b) Topic 1</p> <p>(c) Topic 2</p> <p>(d) Topic 3</p> <p>(e) Topic 4</p> <p>(f) Topic 5</p> <p>(g) Topic 6</p> <p>(h) Topic 7</p> <p>(i) Topic 8</p> <p>(j) Topic 9</p>	 <p>(a) Topic 0</p> <p>(b) Topic 1</p> <p>(c) Topic 2</p> <p>(d) Topic 3</p> <p>(e) Topic 4</p> <p>(f) Topic 5</p> <p>(g) Topic 6</p> <p>(h) Topic 7</p> <p>(i) Topic 8</p> <p>(j) Topic 9</p>
<p>شکل ۷-۷ نمونه visualization هفتم.</p> <p>برای نظرات منفی</p>	<p>شکل ۶-۶ نمونه visualization ششم.</p> <p>برای نظرات مثبت</p>

۱-۵- خلاصه

به طور کلی این تسک به ۲ قسمت تقسیم شده است.

در قسمت اول می‌بایست یک نمایش بصری از کل دیتاست ارائه شود و در قسمت دوم باید یک نمایش بصری برای دو مجموعه در تضاد و قابل مقایسه ارائه شود.

۱-۶- نتایج

- تمامی کد ها باید تحویل گردند.
- نتایج مورد انتظار باید در یک فایل با فرمت Task-1_report.pdf پیوست گردند.
- اگر برنامه نوشته شده خروجی‌ای علاوه بر نتایج مورد انتظار داشته باشد، باید در دایرکتوری‌ای به نام "output" با فرمت دلخواه قرار گیرند.

Second Task: Cuisine Clustering

۲-۱- توضیحات

در این Task ما به دنبال mine کردن دیتاست برای پیدا کردن knowledge در مورد غذا هستیم. در دیتاست کسب و کارها با "categories" تگ گذاری شده‌اند. برای مثال دسته‌بندی "restaurant" نشان دهنده تمامی کسب و کارهای مرتبط با رستوران است. رستوران‌های مخصوص با نوع غذاهایی که ارائه می‌کنند تگ گذاری می‌شوند (برای مثال "Indian" یا "Italian"). مزیتی که این قابلیت به ما می‌دهد باعث می‌شود که بتوانیم که یک نمایش مفید در مورد یک غذای خاص را بدست آوریم. برای مثال متن نظرات تمامی رستوران‌هایی که یک غذای خاص را ارائه می‌دهند. سپس می‌توان از چنین نمایشی برای ارزیابی شباهت‌های بین دو غذا استفاده کرد که امکان خوشه بندی بهتر غذاها را فراهم می‌کند.

۲-۲- هدف

هدف ما در این تسک استخراج دیتاست است که منجر به ساختن یک cuisine map^۱ برای بدست آوردن یک درک بصری از انواع مختلف غذاها و شباهت‌های آن‌هاست. cuisine map به کاربر کمک می‌کند تا متوجه شود چه غذاهایی موجود هستند و روابط بین آن‌ها چیست. روابط بین غذاها کمک می‌کند تا کاربر غذاهای جدید را کشف کند یعنی کاوش کردن غذاهای ناآشنا برای کاربر راحت‌تر می‌گردد.

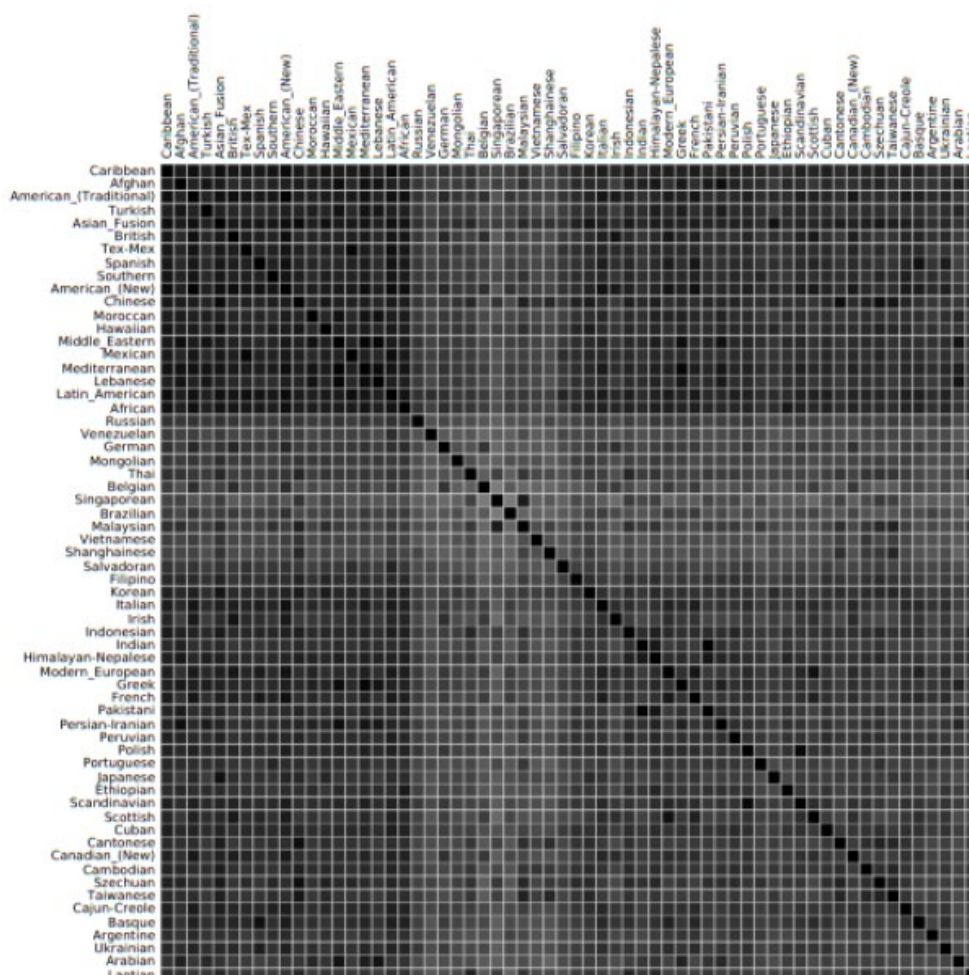
The goal of this task is to mine the dataset to construct a cuisine map to visually understand the landscape of different types of cuisines and their similarities. The cuisine map can help users understand what cuisines are available and their relations, which allows for the discovery of new cuisines, thus facilitating exploration of unfamiliar cuisines.

^۱ ترجمه تحت الفظی فارسی: نقشه غذایی

۲-۳ - قسمت اول Task (Visualization of the Cuisine Map)

در این بخش باید از تمام نظرهایی که در مورد هر غذا در رستوران‌ها داده شده است، برای نمایش همان غذا استفاده کنید و شباهت غذاها را از نظر نمایش‌های متنی مربوط به آن‌ها محاسبه کنید. انتظار می‌رود که شباهت‌هایی را که بدست آورده‌اید را Visualize کرده و آن‌ها را توصیف کنید. در زیر نمونه ای از یک ماتریس شباهت ارائه شده است.

Use all the reviews of restaurants of each cuisine to represent that cuisine, and compute the similarity of cuisines based on the similarity of their corresponding text representations. Visualize the similarities of the cuisines and describe your visualization.



شکل ۲-۱ نمونه visualization اول.

۴-۲ - قسمت دوم Task (Improving the Cuisine Map)

در این بخش باید cuisine map را از طرق زیر بهبود دهید:

- i. تغییر در نمایش متن (برای مثال، بهبود وزن term ها و یا apply کردن topic models)
- ii. تغییر similarity function (مثلا concatenate همه نظرات و سپس محاسبه شباهت و یا ابتدا محاسبه کردن تک، تک نظرات و در انتها جمع کردن مقادیر شباهت‌ها)

Try to improve the cuisine map by:

- 1) varying the text representation
(e.g., improving the weighting of terms or applying topic models)
and
- 2) varying the similarity function
(e.g., concatenate all reviews then compute the similarity, or, first compute the similarity of an individual review, then aggregate the similarity values.)

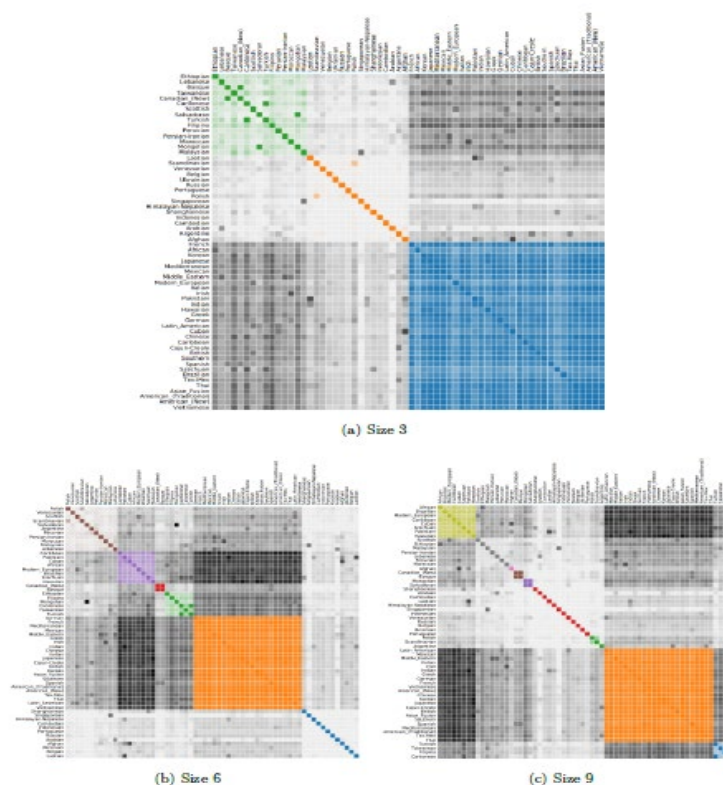
نتایج مورد انتظار همانند شکل ۱-۲ است.

۵-۲- قسمت سوم Task (Incorporating Clustering in Cuisine Map)

از نتایج شباهتهایی که در قسمت اول و دوم بدست آورده‌اید برای خوشه‌بندی کردن استفاده کنید. نتایج خوشه‌بندی را به صورت بصری نمایش دهید تا دسته‌بندی‌های اصلی غذاها را نشان دهد. از حداقل دو خوشه شروع کنید و تعداد خوشه‌ها را تغییر دهید و بررسی کنید که این تغییرات چه تاثیری بر کیفیت / سودمندی cuisine map دارند. صرفاً به یک الگوریتم خوشه‌بندی بسنده نکنید و از چندین الگوریتم خوشه‌بندی استفاده کنید.

Use any similarity results from Task 2.1 or Task 2.2 to do clustering. Visualize the clustering results to show the major categories of cuisines. Vary the number of clusters to try at least two very different numbers of clusters, and discuss how this affects the quality or usefulness of the map. Use multiple clustering algorithms for this task.

نتایج مطلوب در زیر آورده شده است.



شکل ۲-۲ نتیجه خوشه‌بندی.

۲-۶- خلاصه

در این Task به دنبال:

- i. نمایش بصری Cuisine Map.
- ii. بهبود و نمایش بصری Cuisine Map.
- iii. خوشه‌بندی Cuisine Map و بررسی نتایج ناشی از تغییرات تعداد خوشه‌ها.

۲-۸- نتایج

- تمامی کدها باید تحویل گردند.
- نتایج مورد انتظار باید در یک فایل با فرمت Task-2_report.pdf پیوست گردند.
- اگر برنامه نوشته شده خروجی‌ای علاوه بر نتایج مورد انتظار داشته باشد، باید در دایرکتوری‌ای به نام "output" با فرمت دلخواه قرار گیرند.

Third Task: Dish Recognition

۳-۱- هدف

هدف این Task تشخیص غذاهای رایج یا معروف برای یک غذای خاص است. برای درک بهتر میتوان تصور کرد که سوشی یک غذای خاص است و انواعی همانند "ناره زوشی"، "سوشی ادو" و... را شامل می شود. معمولاً هنگامی که برای امتحان یک غذای جدید می روید، انواع مختلف آن غذا را نمیشناسیم. به همین جهت می خواهیم غذاهایی که برای آشپزی در دسترس هستند را با ساخت یک dish recognizer شناسایی کنیم.

۳-۲- آماده سازی

به مسیر "assets/Task-3" مراجعه کرده و فایل "manualAnnotationTask.zip" را extract کنید.

۳-۳- قسمت اول Task (Manual Tagging)

فهرستی از نام غذاهای کاندید به شما داده می شود که همگی متداول هستند. این فهرست توسط برچسب گذاری خودکار "SegPhrase" تولید شده است. طریقه دسترسی به این لیست ها در قسمت ۳-۲ توضیح داده شده است.

برخی از نام ها توسط یک پایگاه دانش خارجی تایید شده اند به طوری که همه آن ها عبارات خوبی هستند و برخی دیگر ممکن است نام غذاهای خوبی باشند با این حال، برخی از برچسب ها ممکن است اشتباه باشند. بنابراین، وظیفه شما اصلاح این لیست برچسب برای یک غذا است. می توانید عبارات را اضافه یا حذف کنید. برخی از اقداماتی که ممکن است متمایل به انجام آن ها باشید عبارت است از:

- I. عبارات false positive را حذف کنید. این عبارات شامل عباراتی می شوند که نام غذا نیستند برای مثال: "hong kong 1" در فایل Chinese.label قابل حذف شدن است. (توصیه شده)
- II. میتوانید عباراتی همانند "hong kong 1" که نام غذا نیستند را به "hong kong 0" تغییر دهید که این عبارت به فرم "negative label" در می آید.
- III. عباراتی که نام غذا هستند اما به اشتباه "negative label" دارند را می توانید حذف کنید. برای مثال "wonton strips 0" که نام غذا است را می توان حذف کرد. (توصیه شده)

IV. همانند حالت دوم می‌توان برچسب این عبارات را تصحیح کرد برای مثال: "wonton strips 0" را به "wonton strips 1" تغییر داد.

V. در صورت تمایل به اضافه کردن عبارت دلخواه، می‌توانید آن را به فرمت شرح داده شده اضافه کنید. نکته قابل توجه این است که کاراکتر بین نام غذا و برچسب آن space نیست و یک tab می‌باشد.

توجه داشته باشید که ابزارهایی که استفاده می‌کنیم به طور کلی برای general phrase mining طراحی شده‌اند برای همین بسیار ایمن‌تر است که برچسب‌های مبهم را حذف کنیم تا تغییر دادن برچسب چرا که تغییر برچسب‌ها ممکن است ما را به سمت رخ دادن ریسک‌های نامشخصی روانه کند اگر چه هنوز ارزش امتحان کردن را دارد.

۴-۳- قسمت دوم Task (Mining Additional Dish Names)

بعد از مرحله ۳-۳ ممکن است هنوز نام برخی از غذاها وجود نداشته باشند در این مرحله باید این لیست را با استفاده از تکنیک‌های pattern mining و یا متدهای word association, expand کنید. برای مثال یک روش unsupervised frequent pattern-based phrase mining, الگوریتم ToPMine می‌باشد. این الگوریتم کلمات متوالی را بر اساس اهمیت آماری ادغام می‌کند (stopword ها ابتدا حذف و بعدا بازگردانده می‌شوند). فریم‌ورک state of the art, روش SegPhrase می‌باشد. SegPhrase به برچسب‌های اصلاح شده نیاز دارد و دارای classifier-ای است که به هر phrase candidate بر اساس ویژگی‌های آماری آن‌ها quality score, assign می‌کند.

The classification procedure will be enhanced by phrasal segmentation results.
These two parts could mutually enhance each other.^۱

رویکرد دیگر برای گسترش نام غذاها، روش word association و یا روش‌های پیشرفته تری همانند word2vec وجود دارند که می‌توانید آن را آزمایش کنید.

^۱ ترجمه نشدن این بخش به درک مطلب خواننده کمک می‌کند.

۳-۵- خلاصه

در این Task به اصلاح و گسترش لیستِ نامِ غذاها می‌پردازیم.

۳-۶- نتایج

- کد مربوطه باید تحویل گردد.
- نتایج مورد انتظار باید در یک فایل با فرمت Task-3_report.pdf پیوست گردند، در انتهای فایل نظر خود را در مورد روش‌ها/رویکردهای شرح داده شده مطرح و مقایسه کنید.
- در فایل Task-2_report.pdf شرح دهید که از کدام یک از اقدامات قسمت ۳-۳ استفاده کرده‌اید و چه ارزیابی‌ای از نتیجه حاصل شده دارید.
- اگر برنامه نوشته شده خروجی‌ای علاوه بر نتایج مورد انتظار داشته باشد، باید در دایرکتوری‌ای به نام "output" با فرمت دلخواه قرار گیرند.

Fourth Task: Mining Popular Dishes Report

۴-۱- هدف

هدف از این تسک ایجاد یک نمایش بصری که نشان دهنده رتبه بندی غذاها برای یک غذای خاص از دیتاست معهود است. شما می‌توانید از فهرست غذاهایی که در Task سوم ایجاد کرده‌اید یا هر فهرست دیگری استفاده کنید.

رویکردهای زیادی برای انجام این Task وجود دارد. چالش اصلی، نحوه ایجاد رتبه بندی خواهد بود. شما می‌توانید روش خود را ابداع و استفاده کنید و یا از روش‌های [Text Retrieval MOOC](#) استفاده کنید.

یک پیشنهاد ساده این است که تعداد غذاهای یک غذای خاص را در تمامی نظرات رستوران‌ها را شمارش کنید اما مهم است که شما جست‌وجو کنید تا این رویکرد ساده را بهبود دهید مثلاً احساس را در یک جمله که به غذایی اشاره می‌کند، فاکتور موثر در این رتبه بندی اعمال کنید. حتی اگر رویکرد ساده ذکر شده را در نظر بگیرید باز هم ممکن است لازم باشد که حالتی که α^1 را در مقابل β^2 در نظر بگیرید، بنا بر این همواره به این سوال توجه کنید:

بهترین راه برای رتبه بندی کردن غذاها برای یک غذای خاص چیست؟

In this task, you will create a visualization showing a ranking of the dishes for a Yelp cuisine of your choice. You may use the dish list we have provided, the list based on your annotations from Task 3 (or a subset of that list), or any other list for other cuisines. You might find it more interesting to work on cuisine for which you can recognize many dishes than one with only a few dish names that you recognize.

There are many ways to approach this task; the main challenge will be how to create the ranking. You can devise your own method or use other methods you have learned in the Text Retrieval MOOC. The simplest approach can be to simply count how many times a dish is mentioned in all the reviews of restaurants of a particular cuisine, but you are encouraged to explore how to improve over this

^۱ شمارش تعداد دفعات ذکر نام غذا بر اساس نظرات

^۲ تعداد رستوران‌ها

simple approach, e.g., by considering ratings of reviews or even sentiment of specific sentences that mention a dish. Even if you just try this simple approach, you may still need to consider the options of counting dish mentions based on the number of reviews vs. the number of restaurants, so keep this question in mind: What do you think is the best way of ranking dishes for a cuisine? This is an **open research question**, but your exploration may help us better understand it.

انتظار می‌رود فارغ از روش پیاده سازی این Task، نموداری/نمودارهایی قابل تفسیر ارائه شود/شوند.

۴-۲- خلاصه

در این تسک باید نحوه‌ای برای رتبه بندی غذاها ارائه کنید.

۴-۳- نتایج

- تمامی کدها باید تحویل گردند.
- نتایج مورد انتظار باید در یک فایل با فرمت Task-4_report.pdf پیوست گردند.
- اگر برنامه نوشته شده خروجی‌ای علاوه بر نتایج مورد انتظار داشته باشد، باید در دایرکتوری‌ای به نام "output" با فرمت دلخواه قرار گیرند.

Fifth Task: Restaurant Recommendation Report

۱-۵- هدف

هدف شما در این Task این است که رستوران‌های خوب را به کسانی که می‌خواهند یک یا چند غذا را از نوعی غذای خاص امتحان کنند، توصیه کنید. در مورد یک غذا ایده کلی حل این مسئله این است که ارزیابی شود آیا رستوران \mathcal{V} برای این غذا خوب است یا خیر. معیار خوب بودن بر این اساس تعیین می‌شود که آیا نظرات کاندید برای رستوران \mathcal{V} شامل تعداد زیادی نظرات مثبت (تعداد خیلی کمی نظرات منفی) برای غذای داده شده می‌باشد یا خیر. شما می‌توانید غذا/غذاهای محبوب را که توسط رتبه بندی‌ای که در Task چهارم انجام داده‌اید انتخاب کنید و یا هر غذایی که نام آن به تعداد زیادی در نظرات ذکر شده است را انتخاب کنید (هر چه تعداد نظرات بیشتری داشته باشید که شامل نام غذا هم باشد، پایه محکم‌تری برای رتبه بندی رستوران‌ها خواهید داشت)

شما می‌بایست یک نمایش بصری برای برای رتبه بندی رستوران‌های توصیه شده داشته باشید.

رتبه بندی رستوران‌ها فارغ از نوع غذا یک رتبه بندی کلی است و به سادگی بدست می‌آید اما این رتبه بندی برای کسی که به دنبال غذای خاصی است کارایی ندارد بنابراین باید رتبه بندی را طوری ارایه کنید که موثر از نام غذا باشد.

سوال اصلی این است که چگونه می‌توان یک الگوریتم رتبه بندی غذا^۱ را رستوران‌ها تعیین کرد.

یک روش ساده این است که بیایم تمام نظراتی که شامل غذای داده شده است را پیدا کنیم سپس رتبه هر نظر را پیدا کنیم و میانگین تمام نظرات را بگیریم بعد هر رستورانی که میانگین نظرات بالاتری برای غذای داده شده داشت را به کاربر پیشنهاد کنیم. شما می‌توانید هر فاکتور دلخواه را در این رتبه بندی تاثیر بدهید مثلاً رتبه کلی رستوران بین سایر رستوران‌ها.

شما باید در نظر داشته باشید که حاصل کار شما در نهایت برای کاربران مفید باشد، یعنی بتوان رستوران محبوب را براسا غذای داده شده به کاربر پیشنهاد دهد، همانند یک search engine.

^۱ منظور از غذا، نوعی خاص است. مثلاً بهترین رستورانی که سوپ مرغ را سرو می‌کند.

In this task, your goal is to recommend good restaurants to those who would like to try one or more dishes in a cuisine. Given a particular dish, the general idea of solving this problem is to assess whether a restaurant is good for this dish based on whether the reviews of a candidate restaurant have included many positive (and very few negative) comments about the dish. You may choose a target dish or a set of target dishes from the list of "popular dishes" you generated from Task 4, or otherwise, choose any dishes that have been mentioned many times in the review data (the more reviews you have for a dish, the more basis you will have for ranking restaurants).

You are required to create a visualization to show the ranking of the recommended restaurants. While a generic ranking of restaurants based on their overall ratings can be easily obtained, such a generic ranking is not as useful as one customized for a particular dish if one has decided to try this "particular dish." Thus, the ranking of restaurants you generated should be influenced somehow by the dish names you assumed to represent a diner's dining preference. The central question is thus how to design a dish-specific ranking algorithm for ranking restaurants. A simple approach easy to implement is to collect all the reviews mentioning a dish, and compute the average ratings of these reviews for each restaurant so that a restaurant whose reviews containing the dish have the highest average rating would be ranked on the top. But you are free to experiment with any parameters such as the rating of the restaurant, among other things.

Something to consider is to make your visualization general enough such that it could be used in a search engine or system and generate something useful for the users by recommending popular restaurants based on different dishes.

انتظار می‌رود فارغ از روش پیاده سازی این Task، نموداری/نمودارهایی قابل تفسیر ارائه شود/شوند.

۲-۵- خلاصه

در این تسک باید نحوه‌ای برای رتبه بندی و پیشنهاد دادن رستوران برتر برای غذایی که مورد علاقه کاربر طراحی کنید.

۳-۵- نتایج

- تمامی کد ها باید تحویل گردند.
- نتایج مورد انتظار باید در یک فایل با فرمت Task-5_report.pdf پیوست گردند.
- اگر برنامه نوشته شده خروجی‌ای علاوه بر نتایج مورد انتظار داشته باشد، باید در دایرکتوری‌ای به نام "output" با فرمت دلخواه قرار گیرند.

Sixth Task:

Hygiene Prediction

۶-۱- هدف

در این Task می‌خواهیم بررسی کنیم که آیا مجموعه‌ای از رستوران‌ها تست بازرسی سلامت عمومی را pass می‌کنند یا خیر. این تست‌ها با توجه به متن نظرات دیتاست به همراه برخی اطلاعات اضافی مانند: location و غذاهای ارائه شده برای مجموعه رستوران‌های داده شده تعیین می‌گردند. پیش‌بینی یک attribute مشاهده نشده توسط داده کاوی، نشان دهنده طیف وسیع و مهم داده کاوی است. از طریق کار بر روی این Task شما تجربه‌ای از کار در چنین شرایطی را بدست خواهید آورد. شما از روش‌هایی که تا کنون آموخته‌ای در برخورد با این مسئله جهان واقع استفاده خواهید کرد و بررسی می‌کنید که کدام یک از روش‌ها مفیدتر/سودمندتر هستند و تجربه کسب می‌کنید.

۶-۲- آماده سازی

به مسیر "assets/Task-5" مراجعه کرده و فایل "Hygiene.zip" را extract کنید و محتویات سه فایل داده شده را بررسی کنید.

۶-۳- خلاصه

در این تسک به دنبال حل یک مسئله جهان واقع هستیم تا از دانش کسب شده در طول ۵ دوره گذرانده شده بهره ببریم.

۶-۴- نتایج

- تمامی کدها باید تحویل گردند.
- نتایج مورد انتظار باید در یک فایل با فرمت Task-6_report.pdf پیوست گردند.
- اگر برنامه نوشته شده خروجی‌ای علاوه بر نتایج مورد انتظار داشته باشد، باید در دایرکتوری‌ای به نام "output" با فرمت دلخواه قرار گیرند.

منابع و مراجع

- | | |
|---|-----|
| https://www.coursera.org/learn/data-mining-project | [۱] |
| https://github.com/yfliu87/DataMining_Capstone/tree/master/task6 | [۲] |
| https://github.com/jiachengpan/dataminingcapstone | [۳] |
| https://github.com/igor-sokolov/dataminingcapstone | [۴] |
| https://github.com/ducthienbui97/DataMiningCapstone/tree/master/Reports | [۵] |
| https://downloadly.ir/elearning/video-tutorials/data-mining-specialization/ | [۶] |

