**Newsletter Personalization：**

**Clustering of Users by Reading Behavior**

Time: 10/27/2019

Author: Ray Liu, David Chen, De Lu, Lily Chou, Kelly Zhao

**Summary**

**Dataset：**Users’ reading behavioral data from a newsletter website in San Francisco.

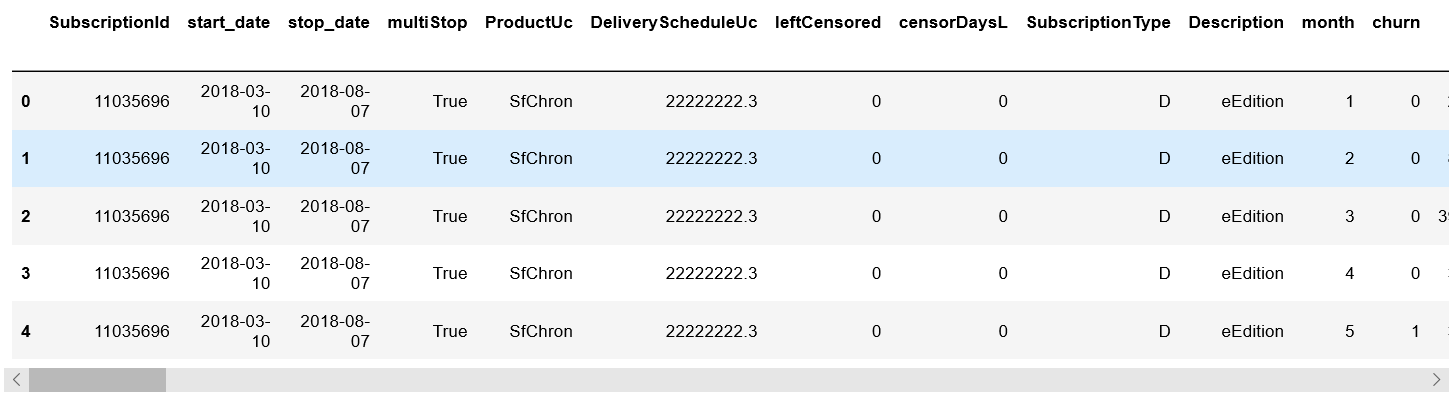
**Approach:** Explore data and cluster users based on reading interest.

**Goal :** Use the clusters to personalize a newsletter to readers each day and perhaps develop other specialized newsletters/products.

**Outcome and take-away:** Identified 11 types of users with different reading interests, and respectively 8, 8, 5 types of users for news, sports and food subcategories.Finally provided personalization suggestions accordingly to the types of users and ways of testing the efficiency of the personalization clustering strategy.

**1. Explore the data. Understand the variables and identify the most relevant variables. Choose the list of variables to work with.**

Data before transformation and aggregation:

****

**…**

Pick the variables used in model:

**Variables taken into clusters - Overall interests and their corresponding page views**

|  |  |  |
| --- | --- | --- |
| **Topics of Articles** | **Sum of Page Views** | **Percentages** |
| news | 118754 | 25.11% |
| business | 74752 | 15.81% |
| sports | 68555 | 14.50% |
| food | 40537 | 8.57% |
| entertainment | 29922 | 6.33% |
| local | 27528 | 5.82% |
| trivia | 24967 | 5.28% |
| chronicle vault | 23546 | 4.98% |
| travel | 22912 | 4.85% |
| ion | 14224 | 3.01% |
| opinion | 13312 | 2.82% |
| living | 8536 | 1.81% |
| biz & tech | 5313 | 1.12% |

**Variables take into clusters - Subclusters related to news**

|  |  |  |
| --- | --- | --- |
| **Topics of Articles** | **Sum of Page Views** | **Percentages** |
| news:bay area & state | 376349 | 59.42% |
| news:crime | 92605 | 14.62% |
| news:politics | 88045 | 13.90% |
| news:california wildfires | 27196 | 4.29% |
| news:us news | 16832 | 2.66% |
| news:science | 11835 | 1.87% |
| news:world news | 8786 | 1.39% |
| news:education | 7049 | 1.11% |
| news:houston & texas | 3198 | 0.50% |
| news:health | 1517 | 0.24% |

**Variables take into clusters - Subclusters related to sports**

|  |  |  |
| --- | --- | --- |
| **Topics of Articles** | **Sum of Page Views** | **Percentages** |
| sports:golden state warriors | 58533 | 30.74% |
| sports:giants | 49315 | 25.90% |
| sports:49ers | 25696 | 13.49% |
| sports:sports columnists | 23097 | 12.13% |
| sports:oakland athletics | 15743 | 8.27% |
| sports:college | 8361 | 4.39% |
| sports:oakland raiders | 7022 | 3.69% |
| sports:golf | 1628 | 0.85% |
| sports:san jose sharks | 744 | 0.39% |
| sports:Pro teams | 285 | 0.15% |

**Variables take into clusters - Subclusters related to food**

|  |  |  |
| --- | --- | --- |
| **Topic of Articles** | **Sum of Pages Views** | **Percentages** |
| food:restaurants | 48805 | 56.65% |
| food:wine & beer | 18645 | 21.64% |
| food:inside scoop | 13874 | 16.10% |
| food:recipes | 4828 | 5.60% |

**2. Create variables by aggregating and transforming the data.**

Aggregation: Based on the raw data, aggregate the variables, grouping by subscription id regarding each customer as an observation:

|  |  |
| --- | --- |
| Sum | "news", "business", "sports", "food”, "entertainment", "local", "trivia", "chronicle vault", "travel", "ion", "opinion", "living", "biz & tech" |

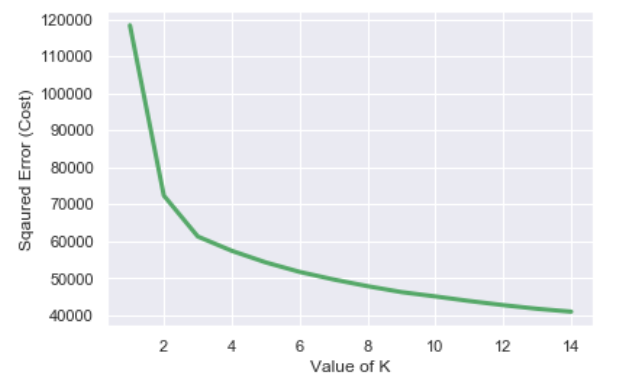
Transformation: As shown in the Pair plot, number of articles read in each category is right skewed so perform log transformation to all variables of this kind.

**3. Identify clusters based on reading behaviors**

**(1) Overall interests**

**Finding the best number of clusters:**

Sum variance within clusters VS number of clusters

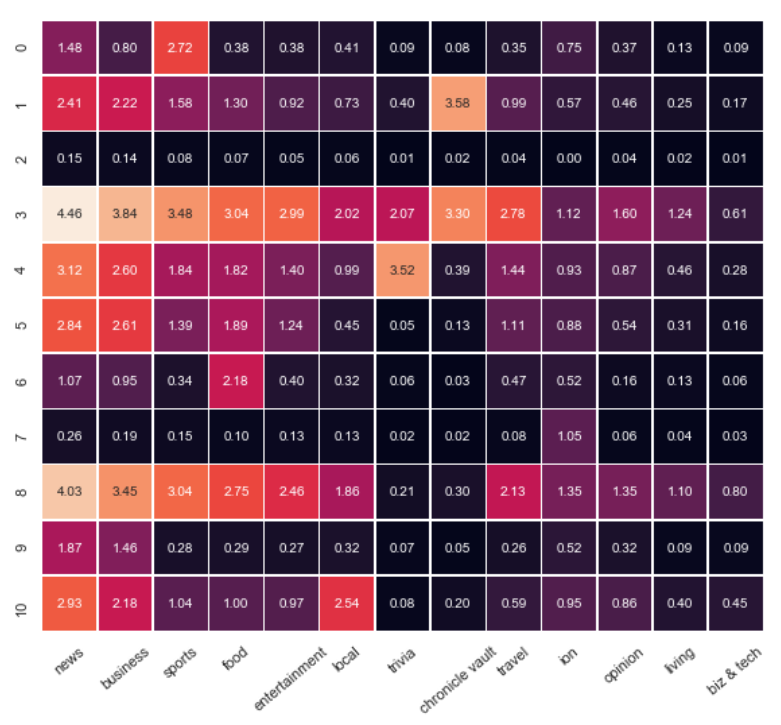


The more clusters there are, the smaller the with-in cluster variance is, with diminishing returns.

We tried cluster numbers 6-12 and found that when cluster number = 11, the characteristics of each cluster is pretty clear to describe. Especially, the original big non-active user cluster is split into two different clusters. This helps us identify some potential marketing strategies towards the non-active user cluster.

**Display the clustering means with visualization**

As shown in the following image, on the Y axis are our 11 user clusters. On the X axis are our 13 topics. The color of the cells indicate how high or how low a value is. The darker colors, the lower the values.

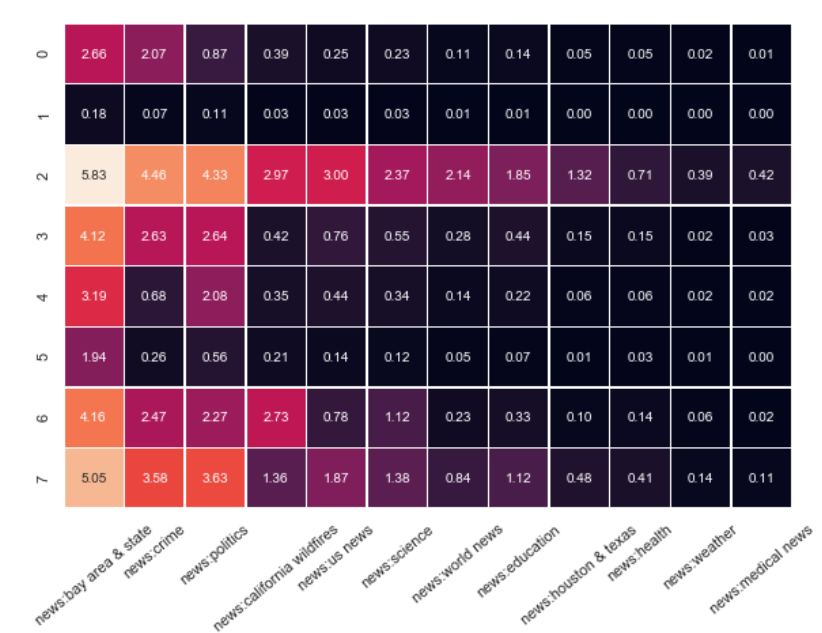


**Profile the clusters**

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster Numbers** | **Sum of Page Views** | **Percentages** | **Descriptions** |
| 0 | 579 | 5.55% | Light users interested in sports only |
| 1 | 255 | 2.45% | Users interested in business and news with considerably older age (Chronical Vault > News > Business) |
| 2 | 3786 | 36.31% | Inactive users |
| 3 | 237 | 2.27% | Highly active users with broad interests (News > Business > Sports) |
| 4 | 359 | 3.44% | Medium active users interested in trivia (Trivia > News > Business) |
| 5 | 733 | 7.03% | Medium active users (News > Business > Food) |
| 6 | 595 | 5.71% | Light users interested in food only |
| 7 | 1681 | 16.12% | Inactive users with interests in ion |
| 8 | 373 | 3.58% | Active users who are not interested in trivia and chronicle vault (News > Business > Sports) |
| 9 | 1425 | 13.67% | News and business oriented (News > Business) |
| 10 | 405 | 3.88% | Pay more attention to local news (News > Local > Business) |

**(2) Subclusters related to news**

**Display the clustering means with visualization**

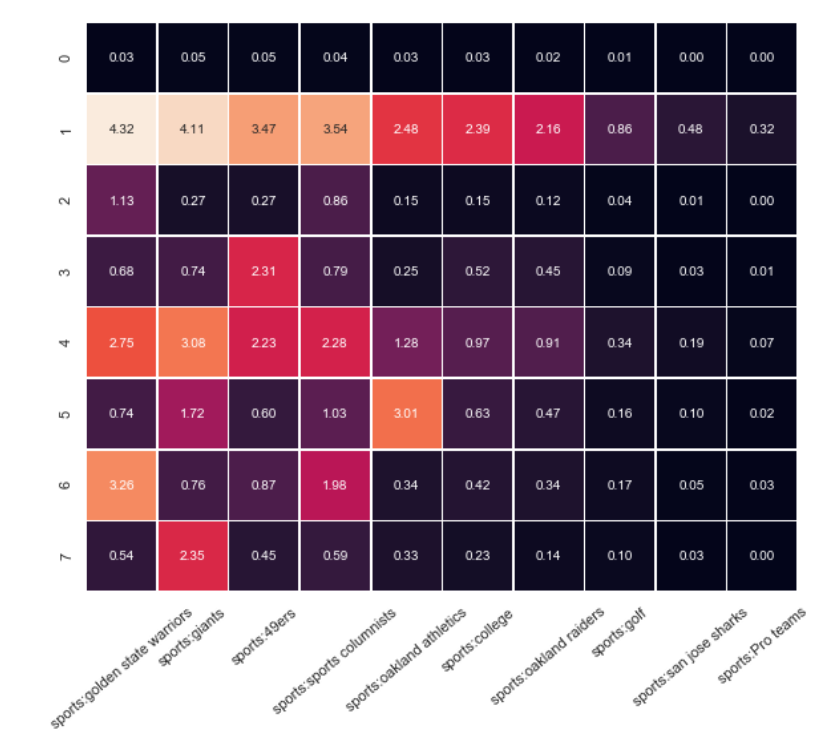
****

**Profile the clusters**

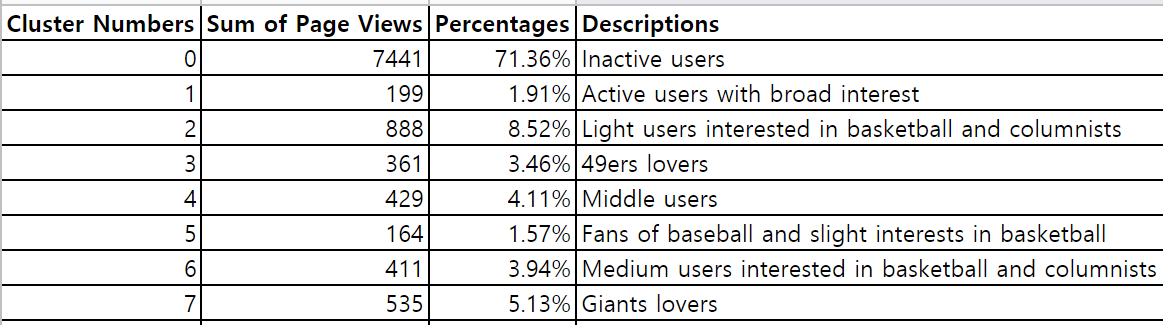
|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster Numbers** | **Sum of Pages Views** | **Percentages** | **Descriptions** |
| 0 | 799 | 7.66% | not active, bay area & crime |
| 1 | 4453 | 42.70% | Inactive in all kinds of news |
| 2 | 225 | 2.16% | Active readers with broad interests |
| 3 | 883 | 8.47% | Active readers; Local+crime+politics |
| 4 | 1073 | 10.29% | Active; Local news and politics |
| 5 | 1879 | 18.02% | Light users, local news only |
| 6 | 523 | 5.02% | Active readers; Local+crime+politics+wildfires |
| 7 | 593 | 5.69% | Active readers focus on local, political and crime news |

**(3) Subclusters related to sports**

**Display the clustering means with visualization**

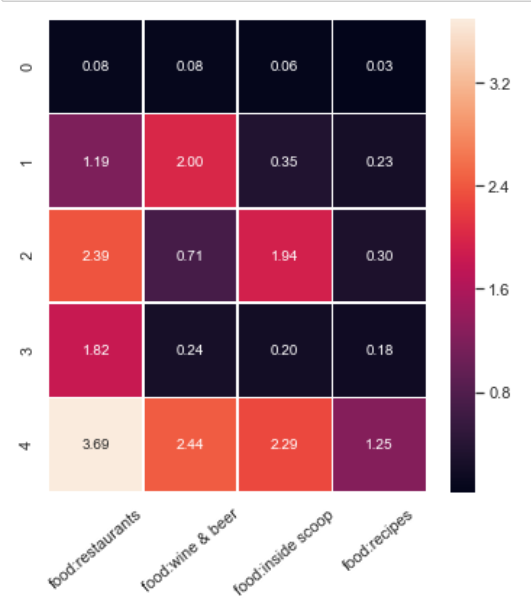
****

**Profile the clusters**

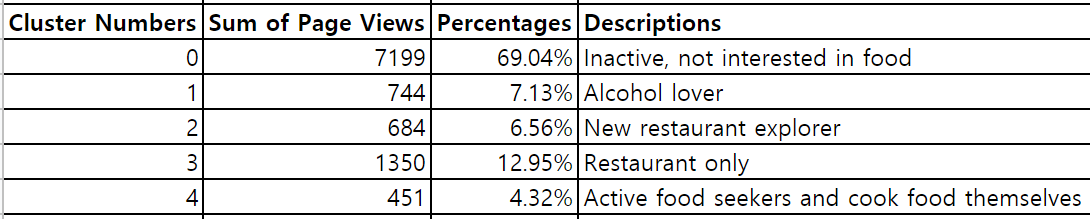
****

**(4) Subclusters related to food**

**Display the clustering means with visualization**

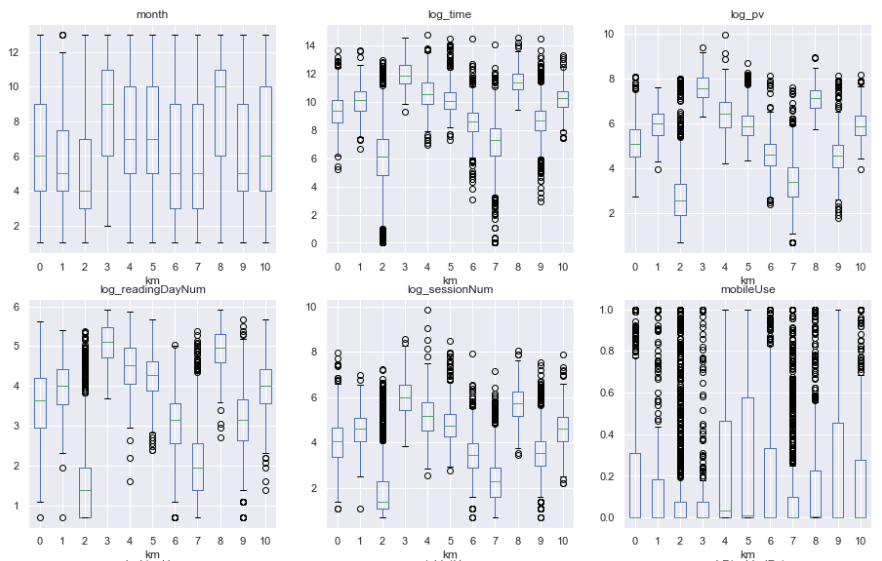
****

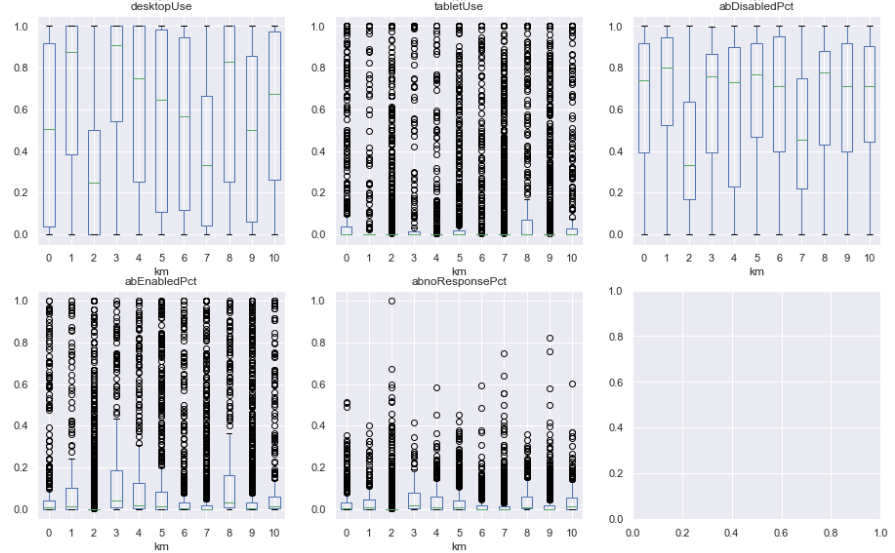
**Profile the clusters**



**4. Profiling and recommendations**

Merge Profilingvalues by cluster



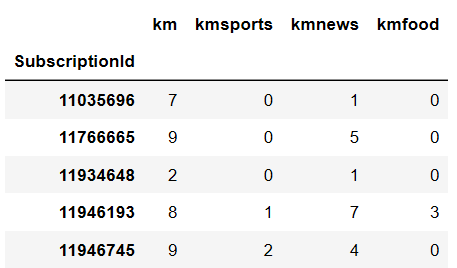


Summary to be added…

**5. Personalization strategy and testing**

**Personalization strategy**

**1) Add cluster labels to each subscription ID**



**2) According to each label to adjust the percentage of each kind of news recomended:**

1. Overall interest: adjust the percentage of news recommend on homepage the larger the cluster mean is, the more of this kind of news recommended
2. News/Sports/Food: adjust the percentage of news shown in each category tab.
3. Adjust the sequence of the tab: put the most interested tab to the left and the least interested one to the right

**A/B testing:**

Hypothesis: The mean reading time and number of sessions will increase after applying the new personalization strategy to San Francisco Chronicle.

The designed experiment

1. Randomly select 10,000 users for Group A and 10,000 Users for Group B
2. Day 1 to Day 7: Apply new personalization to Group A only.
3. Compare the mean reading time and number of sessions, between Group A (Day 1 to Day 7) and Group A (Day -6 to Day 0). Related data for the former group is expected to be larger.
4. Compare the mean reading time and number of sessions, between Group A (Day 1 to Day 7) and Group B (Day 1 to Day 7). Related data for the former group is expected to be larger.