Clustering beer text reviews using the K-means algorithm

Stefan Seastian 242

Motivation

- Beer is the most popular alcoholic drink
- Global beer market is huge (~600$ billion)
- Data about consumer needs => useful for beer marketing

Proposed experiment

- Extract information from beer text reviews using unsupervised learning (clustering)
- Analyze resulting clusters and correlation with initial beer styles

```
,beer_style,review_text
0,Amber,"A - Pours up a gorgeous, crystal clear honey color with a creamy head that
1,Amber,"This bottle popped up in a mixed case from Shangys that a bunch of guys go
2,Amber,"Reviewed 8/19/2008 (bottle BB10Oct08): Pours an orangish amber body. Big f
3,Amber,"Thanks to Eyncognito for this 12 oz bottle. A: The oktoberfestbier is a br
4,Amber,"Found a 22 oz bottle at Sams wines and liquors in Chicago. Dark red amber
```

Sample experiment data

Related work

- Text clustering
  - Anna Huang 2008, comparison of distance measures over tf-idf features
  - Hu et al 2009, enrich document information by using Wikipedia category data
- Review analysis
  - Iacob and Harrison 2013, extracts feature requests from app reviews, using linguistic rules

Related work

- Experiments on beer datasets
  - Braun and Timpe 2015, predict score based on text review, using SVM on bag of words model
  - McAuley and Leskovec 2013, recommender system that adapts to user experience; experience label is learned by model using rating and timestamp data

Dataset

- Started from 500k reviews from the BeerAdvocate website
- Reduced number of different styles by cutting those with few reviews, mapping others to parent style
- Select an equal distribution for each style

Style mapping
(using taxonomy published by BussinessInsider)

American Strong Ale
Saison / Farmhouse Ale
English Pale Ale
American Pale Ale (APA)
American Double / Imperial IPA
American IPA
→ Pale Ale

American Double / Imperial Stout
Russian Imperial Stout
American Stout
→ Stout

Witbier
Hefeweizen
American Pale Wheat Ale
→ Wheat Beer

American Amber / Red Ale
Märzen / Oktoberfest
→ Amber

American Porter → Porter

Belgian Strong Dark Ale
Tripel
Belgian Strong Pale Ale
Dubbel
→ Belgian Ale

Fruit / Vegetable Beer → Fruit / Vegetable Beer

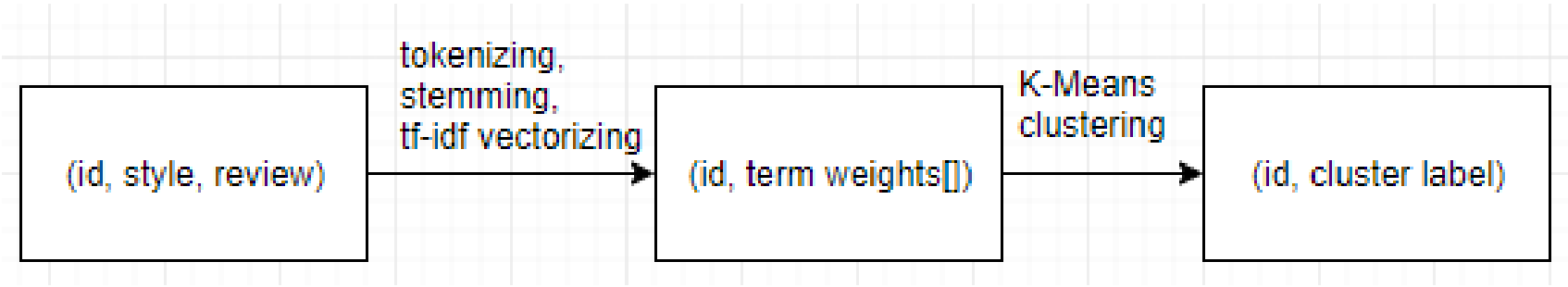
American Barleywine → Barleywine

American Adjunct Lager → Pale Lager

Model overview

- Use text review data from BeerAdvocate
- Extract features with tf-idf vectorizer
- Apply K-means with random seeds and k-d tree seeds
- Analyze cluster results

Model overview



Operations over the initial dataset

Technologies used

- Python programming language
- Clustering algorithms implemented from scratch
- Pandas for data analysis
- Nltk for tokenizing and stemming
- Scikit learn for tf-idf vectorizer
- Matplotlib and Seaborn for plotting

Model parameters

- K = nr of clusters = 9 (number of distinct beer styles)
- Number of features = 1000 (determined experimentally)

Nr features	Precision	Recall	F1
200	0.420	0.481	0.449
700	0.501	0.554	0.526
1000	0.537	0.532	0.534
1500	0.515	0.528	0.521

- Initialization : random seeds 0.53 F1 score, k-d trees 0.47 F1 score

Evaluation

- Correlation between initial beer styles and the clusters found within the dataset
- Confusion matrix : (x, y) = how many times a point of x has been misclassified as y
- Precision and recall are calculated for each row, then an average is taken

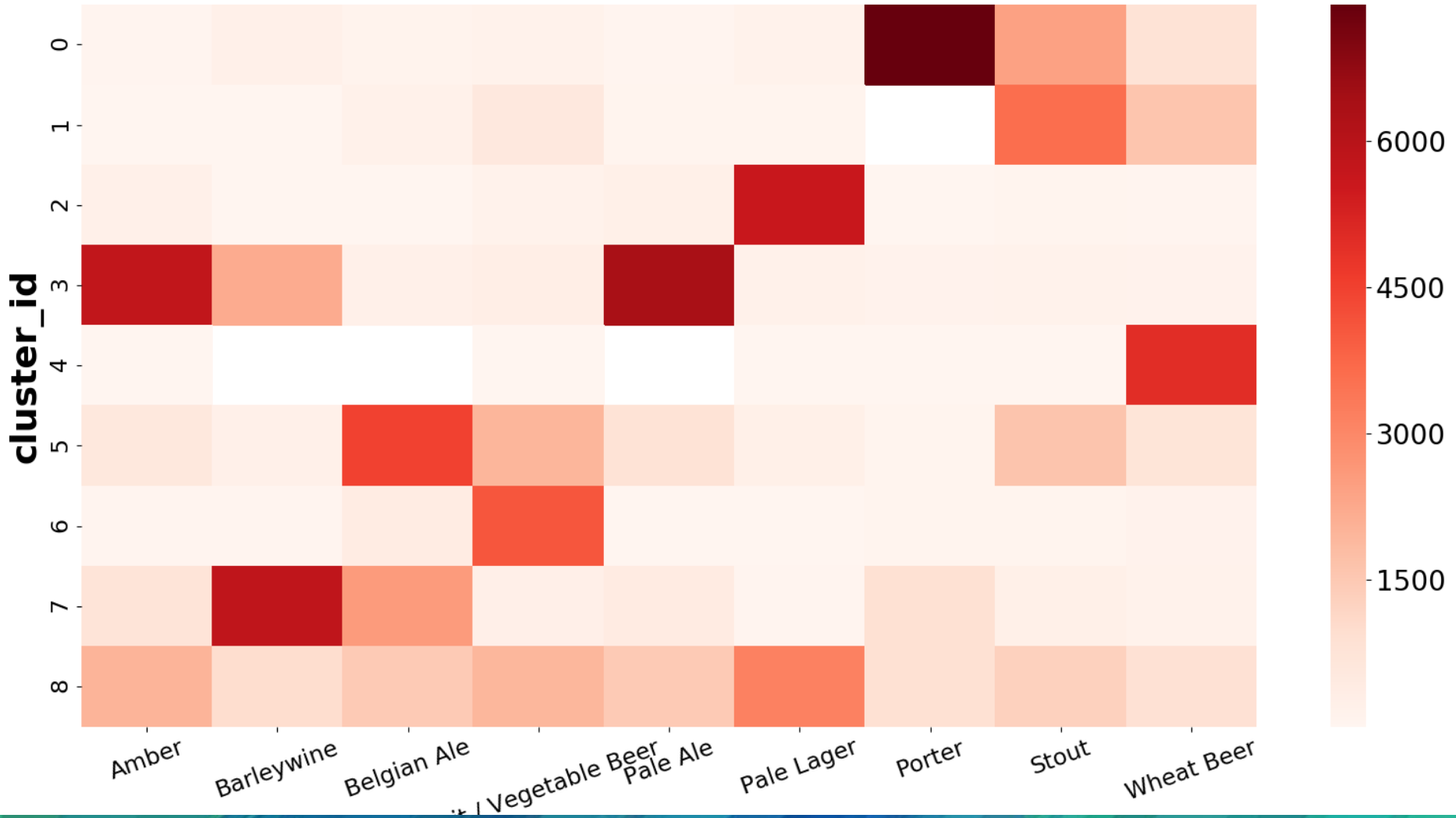
Results

- Multilabel classification score

Mean precision	0.537
Mean recall	0.532
F1 score	0.534

Results

Heatmap



Results

best rated words per cluster

```
0  :  ['chocol', 'coffe', 'roast', 'dark', 'black', 'stout', 'veri', 'beer', 'malt', 'flavor']
1  :  ['wheat', 'banana', 'lemon', 'beer', 'clove', 'light', 'veri', 'tast', 'orang', 'citrus']
2  :  ['lager', 'corn', 'macro', 'light', 'beer', 'adjunct', 'tast', 'yellow', 'grain', 'veri']
3  :  ['hop', 'malt', 'nice', 'bitter', 'veri', 'amber', 'citrus', 'caramel', 'sweet', 'beer']
4  :  ['porter', 'chocol', 'roast', 'coffe', 'dark', 'malt', 'brown', 'nice', 'veri', 'flavor']
5  :  ['appl', 'spice', 'veri', 'yeast', 'light', 'sweet', 'beer', 'white', 'orang', 'flavor']
6  :  ['cherri', 'raspberri', 'tart', 'beer', 'sweet', 'like', 'fruit', 'sour', 'veri', 'red']
7  :  ['alcohol', 'dark', 'sweet', 'caramel', 'brown', 'veri', 'fruit', 'malt', 'hop', 'barleywin']
8  :  ['beer', 'tast', 'like', 'veri', 'smell', 'good', 'drink', 'just', 'realli', 'flavor']
```

Obs. Cluster 8 contains common words for all reviews : beer, taste, smell, drink, flavour

Conclusions

- As expected, clusters were formed around similar beer styles
- The most important words for each cluster contain relevant characteristics
- Further improvements: more data, computation time optimizations