1. Introduction

- Motivation for choosing this project and a brief description of the goals
- Beer is one of the oldest and most popular drinks, with a huge global market and an emerging craft beer scene
- The main goals: extracting information from beer reviews using clustering algorithms

2. Related work

- Text clustering: Anna Huang 2008, comparison of distance measures for tf-idf features; Hu et al 2009, document clustering enriched with Wikipedia category data
- Review analysis: Iacob and Harrison 2013, extract feature requests from app reviews using linguistic rules
- Experiments on beer datasets: Braun and Timpe 2015, predict score based on text reviews using SVM on bag of words

3. Dataset

- Started from 500k review from the BeerAdvocate website
- Removed beer styles with a low number of reviews, made a mapping to unify similar beer types, selected an equal distribution of reviews
- Final count: 9500 per style, 85500 in total

4. Model overview

- Apply stop word removal, tokenizing and stemming on initial review data
- Extract scores using the tf-idf vectorizer
- Cluster scores using K-means with random and k-d tree seed initialization

5. Evaluation

- Correlation between initial beer style and the discovered clusters can be measured as a multi-label classification problem
- Confusion matrix => mean precision and recall
- Best scores: 0.537 precision, 0.532 recall, 0.534 F1 score
- Analysis can also be done on the most relevant terms for each discovered cluster

6. Conclusions

- Clustering results indicate a correlation between beer style and review clustering results
- Most relevant terms per cluster are descriptive of the associated beer style