grapevine

a sommelier in your pocket

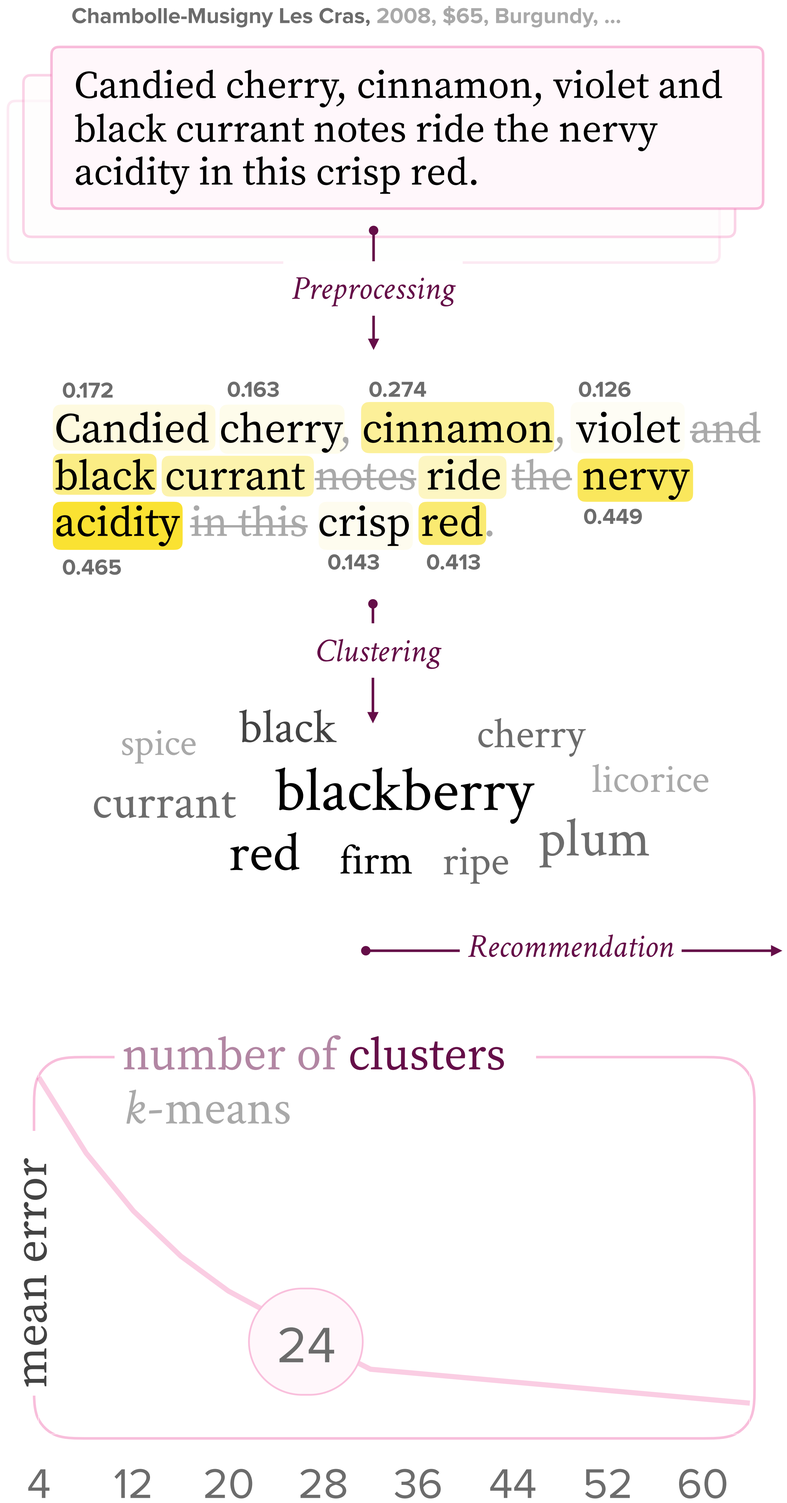
Motivation

Wine is an expensive hobby, and the majority of people have neither the time nor the money to discover their own flavor palate. Typically, the only claim one can make about a given glass of wine is whether or not it was enjoyable.

For this reason it is incredibly difficult for one to discover new wine. Our project looks to simplify the endeavor, through techniques in the realm of unsupervised learning and recommendation.

Formally, we define the problem as follows. Given design matrix X, we look to assign each example x to one of k clusters. Then we seek to output an optimal recommendation w.

written by
Geoffrey Angus, gangus@stanford.edu
Rooz Mahdavian, rooz @stanford.edu
Richard Martinez, rdm@stanford.edu



Pipeline

Our dataset consists of ~300k wine reviews, augmented by name, price, rating, region, vintage, and county (aggregated from WineSpectator.com).

We first filter out extremely common words, and then apply tf-idf; sommeliers are quite precise with their language, so a unique word across two distinct reviews could signal meaningful similarity.

We then cluster using mixture of k-Gaussians via EM, which reduces the search space drastically for recommendation. The optimal cluster count was found efficiently via k-means, using the SSE (Sum Squared Error) falloff as our success signal.

We finally recommend by sampling from a Multinomial over the k outcomes, weighted by previous preferences. Within the cluster, a previously preferred wine is sampled (with multivariate Gaussian noise) and becomes the benchmark for finding the two clusters with highest probability; within this space, we minimize the following cost:

$$J(\boldsymbol{w}, \boldsymbol{w}') = \frac{\text{price}(\boldsymbol{w}')}{\text{cost}(\boldsymbol{w}')} + \lambda \frac{\boldsymbol{w} \cdot \boldsymbol{w}'}{||\boldsymbol{w}|| \cdot ||\boldsymbol{w}'||}$$

Future Work

Tuning score, price and similarity hyper-parameters within cost; implementing GloVE vectors over tf-idf scores.