Using weather data for wine recommendations

Paul Shved, Stanford University, CS221, Fall 2018

Objectives

- **unsupervised** wine recommendation
- based on the **review text analysis**
- Use **weather data** to improve

Dataset

- **Wine**: Wine Spectator expert reviews (270,000 reviews, 20,000 wineries), $60/year
- **Weather**: GHCN-M world data from NCEI (471,000 station-years)
- **Google Maps API**: 580 calls for $2.90

Feature extraction

- $T \in \{0...10\}$: monthly avg → vector of $[0,1]$ min-max-normalized → North-South unification → Hierarchical clustering
- $W \in [0,1]^{20}$: description → 1- and 2-grams → Term Frequency → LDA with 20 categories (>20 tends to overfit)

Problem statement

If the user likes the wine (W_0, T_0), then they must also like the wine (W, T):

$$T = T_0 \wedge ||W - W_0||_1 < \tau \qquad (1)$$

We also introduce **winery proximity**,

$$||W - W_0||_1 \cdot \log \frac{\text{total count from winery X}}{\text{count from X matching (1)}}$$

LDA extracted features

Topic #13: white, lemon, peach, apple, acidity, grapefruit, fresh, lime → **Sauv Blanc**
Topic #19: dark, blackberry, plum, chocolate, ripe, licorice, syrah → **Southern France**

TF-IDF does not work

- Wine reviews use small vocabulary.
- TF-IDF filtered out such signifiers as *"crisp"*, *"plum"*, *"cherry"*, and *"apple"*.

Results

- Correctly predicts similarity within established regions
- Weather data **reduces** amount of prediction outside the region (due to fewer matching weather). Matches in CA for FR wines:

	wine 1	wine 2
use weather	92	43
ignore weather	7	14

Surprising results

- LDA says many white wines taste similar
- Ignoring weather **reduces matching**
- Significant cross-varietal matching

- Using weather feature T of cardinality 10 requires more data (10x more?)
- When data are available, weather improves recommendations
- North-South unification requires more work