Trend-Aware Rating System

Caroline Chang
Kenta Takatsu

Introduction

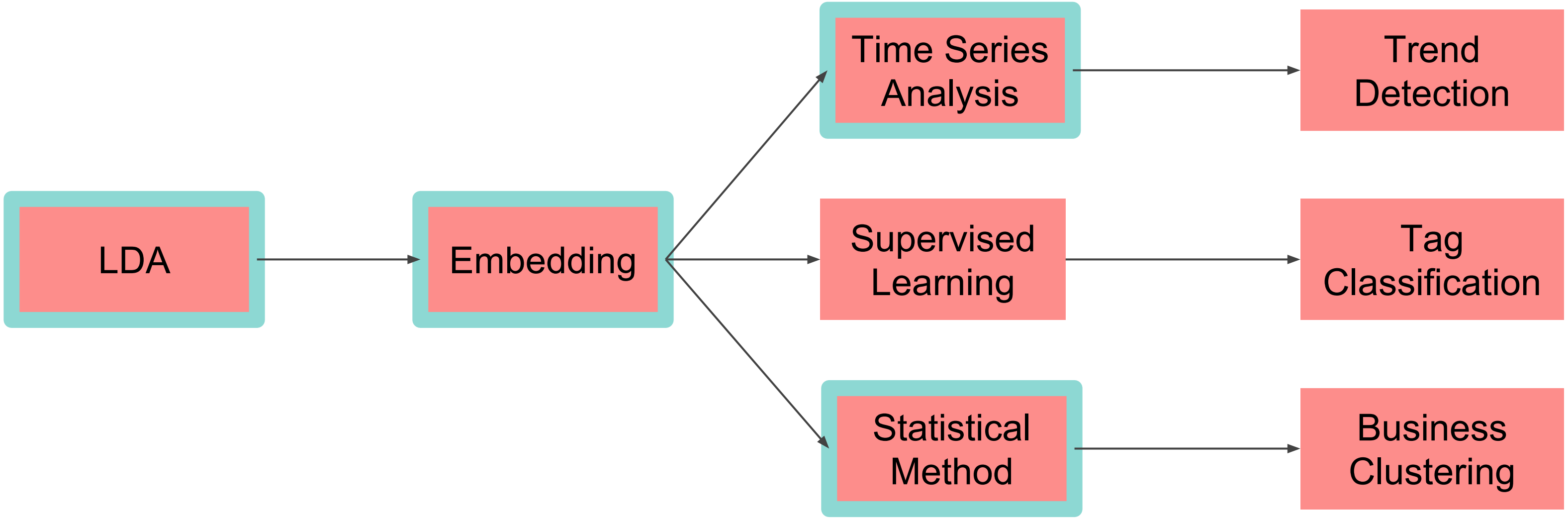
Motivation

- Trend-Aware Rating System
    - Generate time-sensitive rating metric for the better representation of current business
        - Training and testing classifiers that most accurately predict star ratings for a single business
    - Detecting trending topics while accounting for increase in Yelp users over time
        - Possibly could explain drastic fluctuations in business performances

- Topic-Based Tag Matching
    - Define the similarity of 2 businesses by the generalized embedding of topic distributions
    - Use this embedding to improve Yelp user experiences by
        - Removing ambiguous tags such as ‘Real Estate Agents’ and ‘Real Estate Services’
        - Defining the better representation of business categories in a continuous metric
        - Extracting implicit business traits such as ‘good portion’ and ‘friendly staff’

Methods

Roadmap


LDA
Embedding
Time Series Analysis
Supervised Learning
Statistical Method
Trend Detection
Tag Classification
Business Clustering

Progress

- Trained LDA and NMF topic model
- Generated distribution of topics for each review
- Observed cosine similarity and its confidence intervals
- Began time series analysis and star rating prediction with classifiers

Pre-Processing

1. Latent Dirichlet Allocation (LDA)
   - Extracting latent topics from review text
2. K-Index Embedding
   - This gives us an idea of the relative topic distribution per review text
   - We use this in our preliminary stages of coming up with a classifier model

Embedding

- Additive NMF topics
  - Since the NMF uses Tf-Idf vectorizer, the product of topics usually ends up 0
  - Cumulative sum of topics and scaled to be 1

- Multiplicative LDA topics
  - Since LDA uses Count vectorizer and it is a probabilistic model, we can apply bayes rule
  - Cumulative product of LDA topics

Cosine Similarity

- Calculated the similarity of 2 topic embeddings
- Randomly select 2 reviews from businesses and constructed confidence interval
- LDA, MNF are not generalized enough to capture the similarity of 2 businesses

$$\frac{\sum_{i=1}^{n} A_i B i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Topic Distributions Over Time

- Preliminary assessment of topic trends over time
- For a given topic, we plot the k-index values over time
  - This was done across all the reviews for a single business
- Want to find most relevant and consistent topics for later linear regression, SVM training
- Despite small k-index values, we do notice some trends with the values

Predicting Star Ratings Over Time

- Training models
  - Constructed **linear regression classifier** with topic distributions as features and star rating as the response variable
- For “n” number of reviews in the training set, how accurate is the next star rating for the next “n” reviews
- Establish a baseline of how much we can improve in our star rating predictions

Results

Topic models

Chinese

- dim sum carts cart mai places har items vegas
- soup noodle wonton bowl broth wontons base drop ton
- sushi roll fresh bar rolls sashimi tuna salmon fish
- tea milk bubble ice boba drink green hk drinks

Pizza

- staff friendly super helpful clean fast attentive wait family
- wings hot buffalo crispy ranch honey bbq fries medium
- beer selection beers tap craft wine local list burger
- free gluten options vegan menu offer option regular eat

Similarity and Confidence Interval

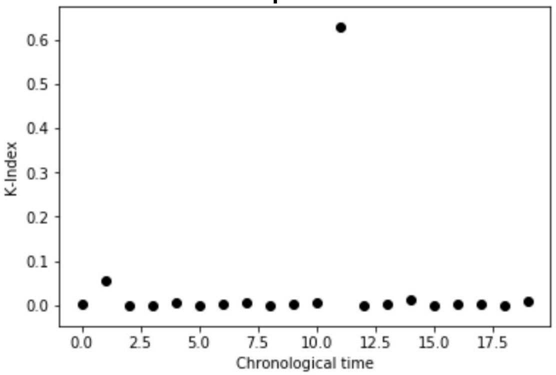
- Took the average of all reviews for each business
- 95% - 91% matching was observed in 10 Canadian Chinese Restaurants
- Constructed 95% confidence intervals with 90%+ matched restaurants.

(0.38177451520125377, 0.54461843494557982)
(0.4010228211599064, 0.55103626238774905)
(0.30415087880980851, 0.44544987447367779)
(0.30869657664824723, 0.47429023439115037)
(0.36271312573149744, 0.53339694752447109)

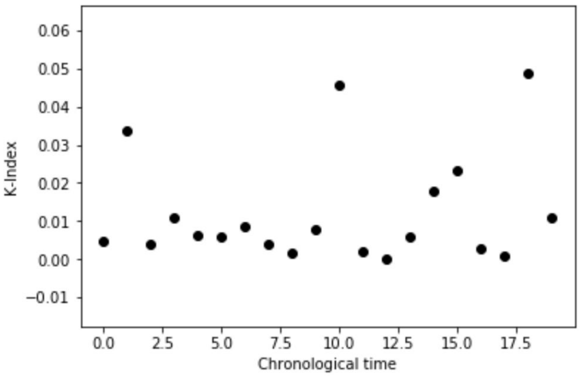
- Each review contains huge variance in topic distribution, and we need more generalized model

Finding Reliable Topics

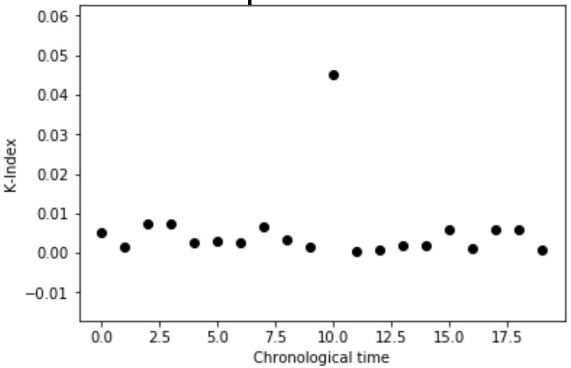
Topic 9

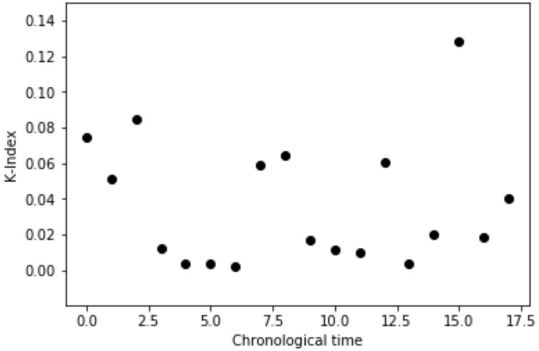


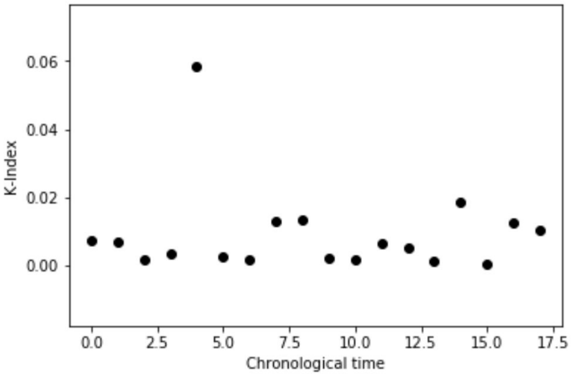
Topic 0

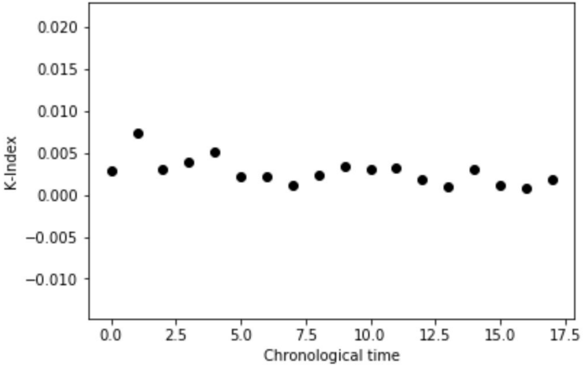


Topic 12

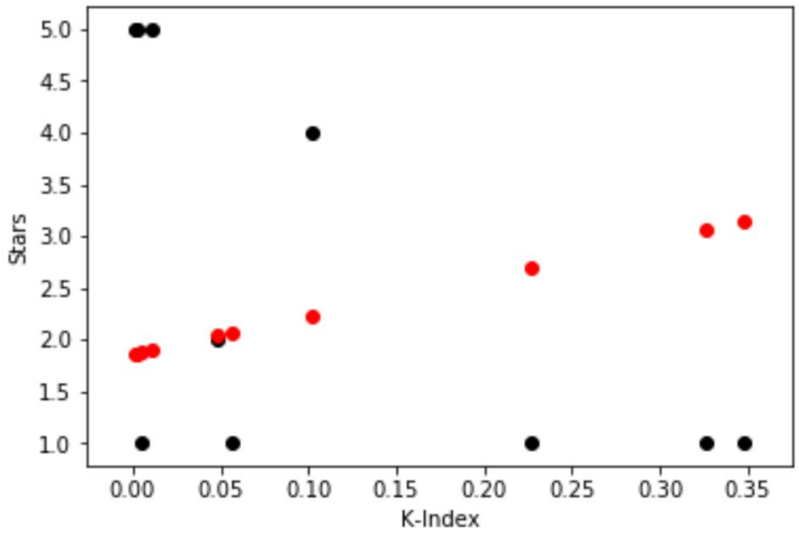




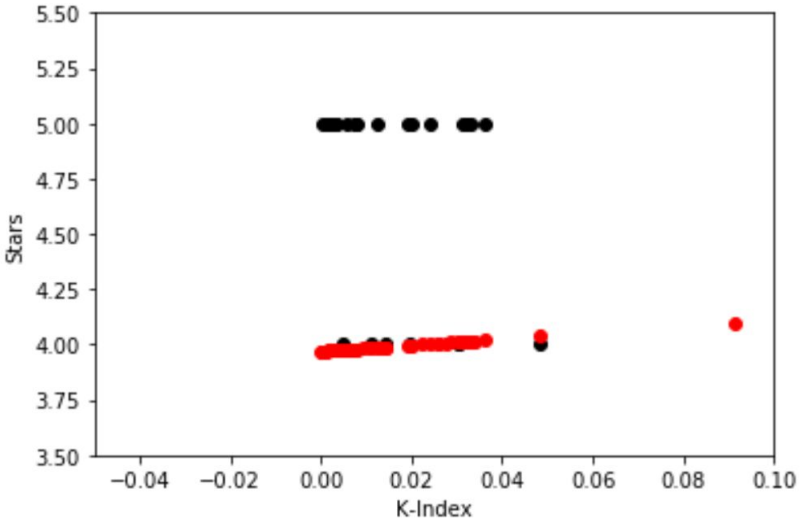




Linear Regression Preliminary Results



Testing and training using a single pizza business (why having “data-rich” businesses are so important)



Testing and training with multiple businesses
Classifier that solidly predicts “4-star” rating

Further Studies

- Hot topic trends
    - Determining such trends → substantiating drastic fluctuations in business performance
- Merging categories
    - Larger groups → More data-richness and less variability (i.e. sparse number of reviews)
- Finding the right subset (Chinese? Piazza? Pizza?)
- Better embedding of topics (will try Word2Vec, Stacked Denoising Autoencoder)

References

References

Trend-Awareness :

- Improving Restaurants by Extracting Subtopics from Yelp Reviews (J. Huang, S. Rogers, E. Joo)
- Personalizing Yelp Star Ratings: a Semantic Modeling Approach (J. Linshi)
- Sentence Level Recurrent Topic Model: Letting Topics Speak for Themselves (J. Huang, S. Rogers, E. Joo)

Topic-Based Tag Matching:

- Collaborative Topic Modeling for Recommending Scientific Articles (C. Wang, D. Blei)
- Sentence Level Recurrent Topic Model: Letting Topics Speak for Themselves (J. Huang, S. Rogers, E. Joo)
- Stacked denoising autoencoders for sentiment analysis (H. Sagha, N. Cummins, B.Schuller)