Fun with Yelp

for CS 199, sp14

Sean Massung

1 Yelp Dataset

The Yelp Academic Dataset [3] consists of three main objects encoded as JSON files:

- **Business**
  - type, id, name, location, stars, review count, category, open, URL
- **Review**
  - type, business id, user id, stars, **text**, date, votes
- **User**
  - type, id, name, review count, average stars, votes

They're all linked together in a huge network. How can we use this network (or parts of it) to gain knowledge or solve useful problems? Below are two recent papers that make use of the Yelp data.

2 Recommendation Dialogs

- "Generating Recommendation Dialogs by Extracting Information from User Reviews" [2].
- **Purpose**: generate and dynamically rank relevant questions from user reviews.
  - IR: browsing *vs* searching. Include speech recognition?
  - Why is "user-generated" important?
  - Typically, systems draw questions from small, fixed set of topics (*e.g.* price or cuisine)
- **Main Idea I**: Use topic modeling and sentiment-based aspect extraction to identify fine-grained attributes for each business
  - Run LDA (topic model) on already partitioned cuisine types to get subtypes for Japanese, Coffee and Tea, Vegetarian, *etc*

  - Topics are then manually labeled, and businesses assigned a highest subcategory
  - Extract noun phrases (NPs) with sentiment tags
  - Apply learned syntactic patterns to identify other NPs:
    * Business + have + ADJ + NP → *This place has some really great yogurt and toppings*
    * NP + be + ADJ → *Our pizza was much too jalapeno-y*
  - Syntactic patterns are used as templates to create questions: *Do you want a place with good burritos?*

- **Main Idea II**: use information gain at each dialog step to select the most informative question (see Decision Trees)
  - Select questions that maximize the entropy of the resulting document set (list of businesses)
  - After the yes/no question is answered, remove businesses that don't satisfy answer from set
  - Do this in a loop until there is only one business left or until a fixed number of iterations has run
  - Result: recall was improved against terrible baselines

- **Tools overview**:
  - Topic modeling
  - Sentiment analysis
  - Information gain

- **Data overview**:
  - Yelp Academic Dataset (Yelp Reviews from Phoenix)
  - Created sentiment lexicon (1329 adjectives)

- This paper is glue connecting a bunch of existing tools—you can do this sort of thing, too!

3 Where *Not* to Eat

- "Where *Not* to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews" [1].

- **Purpose**: Find textual signals in restaurant reviews to predict hygiene inspection records and identify cues in reviews that are indicative of sanitary conditions
  - Department of Health has limited resources

  - Let citizen reviews help!
  - But are average citizens familiar with hygiene laws and inspection codes?

- **Main Idea**:
  - First, only keep businesses that have public health records
  - Then, filter out outlier reviews per business with a deceptiveness classifier
  - Initial findings:
    - average review is negatively correlated with inspection penalty scores
    - If deceptive reviews are added back in, the correlation is weaker. . .
  - Customer opinion features: average rating, unigram and bigram words from reviews
  - Restaurant metadata features: cuisine, location, **inspection history**, review count, non-positive review count

- **Results**:
  - Accuracy is measured in discriminating severe *vs* no violations.
  - 57% accuracy using average rating
  - 72% accuracy using inspection history
  - 82% accuracy with unigram and bigram word features
  - 81% accuracy using all features
  - Are these good?

- **Tools overview**:
  - Web crawler
  - liblinear SVM classifier
  - Deception detection classifier
  - **No** sentiment analysis tools used, just looked at rating metadata

- **Data overview**:
  - Crawled Yelp reviews from Seattle
  - Hygiene inspection records from Department of Public Health

References

[1] Jun Seok Kang, Polina Kuznetsova, Michael Luca, and Yejin Choi. Where not to eat? improving public policy by predicting hygiene inspections using online reviews. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1443–1448, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[2] Kevin Reschke, Adam Vogel, and Dan Jurafsky. Generating recommendation dialogs by extracting information from user reviews. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 499–504, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[3] Yelp Inc. Yelp Academic Dataset, April 2014. `https://www.yelp.com/academic_dataset`.