Tripadvisor
The Good, The Bad, and The Neutral
AC297r Final Presentation
Louie Hoang, Jiachen Song, Xingchi Dai, Sarah Mohamed
Background

World’s largest travel website
Provide the most updated travel-related information
Know Better, Book Better, Go Better
1. Problem Statement

Improve the quality of returned search results by interpreting past users’ feelings toward the specific features of each business.
The Data

- Total number of reviews: 582,636
- The most reviewed business: Boston Omni Parker House (9330)
  --The original Parker House Hotel opened on the site on October 8, 1855, making it the longest continuously operating hotel in the United States.
- The average reviews per business: 174.3

<table>
<thead>
<tr>
<th>Business</th>
<th>TimeStamp</th>
<th>ReviewTitle</th>
<th>ReviewText</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Krusty Krab</td>
<td>02:24:12 06-13-2015</td>
<td>Plankton Infestation</td>
<td>Great recipe, but overrun...</td>
</tr>
</tbody>
</table>
The Search Engine

- Parse the words of the query
- Look for matching “entities” - noun phrases that are discussed in the business reviews.
  - e.g. “good lobster”, “chicken tacos”
- Grab all of the businesses where these entities are discussed in its reviews
- Rank the businesses according to how their users “felt” about those entities - are they discussed in a positive way, or a negative way?
Our solution
Short Web Demo
Methodology And Modeling

2.

Named Entity Recognition (NER): What ‘things’ (noun phrases) are being discussed in a review?

Sentiment Analysis: What feelings (positive/negative) are being expressed about an entity?

Topic Modeling: When there are no specific entities matching a user query, what might be a related query?
We enjoyed excellent cocktails in the lounge as well as an excellent breakfast in our room.

- **Part of Speech Tagging**: What is the part-of-speech for each word?

- **Entity Recognition**: Detect noun-phrases
Part-of-speech Tagging

What is PoS Tagging?
- Assign a part-of-speech/lexical class marker to each word in the corpus.
- Example: Natural language processing (NLP) is a field of computer science

Why we need PoS Tagging?
- Word Sense Disambiguation (e.g. Time flies like an arrow / Your efforts will bear fruit)
- Information Extraction
Named Entity Recognition (NER)

Detect entities by chunking together words based on their POS tags:

- LDA method to treat entities as topics (previously mentioned)
- Grammar matching with regular expression: `<DT>?<JJ>*<NN>`
- Our approach (simpler): Greedy matching of continuous nouns
We enjoyed excellent cocktails in the lounge as well as an excellent breakfast in our room.

Positive/Negative/Neutral
Sentiment Analysis (Modeling)

- **Formulation:** Binary *(positive/negative)* or Multi-class problem (1-5 stars)
  a. Positive: The doorman and bellman are great.
  b. Negative: Was terribly disappointed with the room setup.

- **Approaches:** Use *(Positive, Negative, Neutral)* classification
  a. Rule-Based System
  b. Standard ML Techniques (LR, OVA, Neural Networks)
Sentiment Analysis (Modeling)

Linear model,

\[ \hat{y} = f(xW + b) \]

- \( W \in \mathbb{R}^{d_{in} \times d_{out}}, b \in \mathbb{R}^{1 \times d_{out}} \); model parameters
- \( f : \mathbb{R}^{d_{out}} \mapsto \mathbb{R}^{d_{out}} \); activation function
- Sometimes \( z = xW + b \) informally “score” vector.
- Note \( z \) and \( \hat{y} \) are not one-hot.
Sentiment Analysis Models

- NLTK Vader
- Convolutional Neural Networks (CNN)
- Deep Convolutional Generative Adversarial Networks
- FastText
- Semi-supervised LSTM & Logistic Regression
Model 1 - NLTK Vader (Baseline Approach)

ICWSM '14: VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text

- Quality Crowd-Sourcing to build a sentiment lexicon (-4 to +4 intensity)
  - Okay (0.9), Good (1.9), Horrible (-2.5), Sucks (-1.5)

- Rule-based system to determine whole-sentence intensity:
  - Exclamation point (food is good!!!) = increase magnitude
  - Capitalization (food is GREAT!) = increase magnitude
  - Modifiers (food is extremely good) = increase/decrease intensity
  - Contrastive Conjunction (food is good, but service is horrible) = polarity shift, latter dominant
  - Preceding Tri-gram (food isn't all that great) can catch negated cases
Model 2 - CNN (EMNLP '14)

wait for the video and do n't rent it

$n \times k$ representation of sentence with static and non-static channels

Convolutional layer with multiple filter widths and feature maps

Max-over-time pooling

Fully connected layer with dropout and softmax output
Figure 1: DCGAN generator used for LSUN scene modeling. A 100 dimensional uniform distribution $\mathbf{Z}$ is projected to a small spatial extent convolutional representation with many feature maps. A series of four fractionally-strided convolutions (in some recent papers, these are wrongly called deconvolutions) then convert this high level representation into a $64 \times 64$ pixel image. Notably, no fully connected or pooling layers are used.
Figure 1: Model architecture of **fastText** for a sentence with $N$ ngram features $x_1, \ldots, x_N$. The features are embedded and averaged to form the hidden variable.
Model 5 - Semisupervised LSTM + LR (novel*)

- **Unsupervised**: Train a LSTM Language Model on reviews.

- **Supervised**:
  - Use word embeddings & LSTM parameters from *Unsupervised* model as features.
  - Train logistic regression on review ratings.
Topic Modeling - Latent Dirichlet Allocation

- **Topic**: Defined by a distribution over a set of vocabulary words

- **Document**: Sequence of words

- Assume the data come from a **generative** model:
  - Each document is generated from a random mixture of topics
  - Each topic is described by a distribution over words
Generate candidate labels (n-grams or entities/noun phrases)

Score each candidate label according to negative KL divergence between the topic and the label:

\[
Score(l, \theta) = -D(\theta || l) = - \sum_w p(w | \theta) \log \frac{p(w | \theta)}{p(w | l)}
\]

KL divergence captures the distance between two distributions and When the two distributions are close together, the label is “good”
3. Modeling Result
# Sentiment Analysis - Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nltk Vader</td>
<td>61%</td>
</tr>
<tr>
<td>DCGAN</td>
<td>Failed</td>
</tr>
<tr>
<td>CNN</td>
<td>89%</td>
</tr>
<tr>
<td>FastText</td>
<td>90.5%</td>
</tr>
<tr>
<td>Semi-supervised</td>
<td>92%</td>
</tr>
</tbody>
</table>
Topic Modeling - Results

**Elbow Plot for Optimal Number of Topics**

- Sum of Errors vs. K

**Topic Frequency Among Businesses**

- Number of Businesses vs. Topic
Topic Modeling - Results (continued)

Topic 0: room hotel stay location staff clean boston
Topic 1: hotel airport shuttle hilton boston stay service
Topic 2: museum tea kid aquarium exhibit fun great
Topic 3: trail history boston freedom tour walk church
Topic 4: time u experience like minute didnt got
Topic 5: hotel boston room great view marriott service
Topic 6: hotel room service boston staff stay location
**Topic 7: pizza sandwich place good food lunch cheese**
Topic 8: beer bar pub drink great boston fun
Topic 9: restaurant food service wine menu dinner table
Topic 10: view harbor boat boston water wharf ship
Topic 11: pastry line chocolate mike boston cannoli end
**Topic 12: food good service burger place great drink**
Topic 13: u hotel rock great verb kimpton wine
**Topic 14: food restaurant good dish service sushi chicken**
Topic 15: food restaurant italian pasta end north good
Topic 16: market boston hall food restaurant shop quincy
Topic 17: harvard cambridge square charles river mit campus
**Topic 18: e la que boston en le da**
Topic 19: lobster seafood food boston clam oyster chowder
Topic 20: copley bay shopping prudential hotel newbury mall
Topic 21: room boston breakfast house stay u fairmont
Topic 22: breakfast coffee inn staff egg morning good
Topic 23: museum art visit exhibit collection library building
Topic 24: boston park beautiful garden place walk common
Topic 25: great place good time really like small
Topic 26: hotel room dear view service boston hyatt
Topic 27: thank u boston stay hope review time
Topic 28: tour boston guide great city time fun
Topic 29: fenway game park sox great baseball red
Topic Labeling - Results

**Topic 7:** pizza sandwich place good food lunch cheese
**Topic 8:** beer bar pub drink great boston fun
**Topic 9:** restaurant food service wine menu dinner table
**Topic 10:** view harbor boat boston water wharf ship
**Topic 11:** pastry line chocolate mike boston cannoli end
**Topic 12:** food good service burger place great drink
**Topic 13:** u hotel rock great verb kimpton wine
**Topic 14:** food restaurant good dish service sushi chicken

**Topic 7:** cheese delicious sandwich
**Topic 8:** bar brewery drink
**Topic 9:** menu wine meal
**Topic 10:** wharf whale boston harbor
**Topic 11:** mike pastry cream
**Topic 12:** burger table drink
**Topic 13:** kimpton stay verb
**Topic 14:** sushi rice dish
Thanks!

Any questions?