objective

GENERAL GOAL

- Predict similar restaurants with Yelp reviews

TECH GOALS

- NLP preprocessing techniques
- Topic modeling with sentiment analysis
- Google Cloud Exposure
- Flask

DELIVERABLES

Build a usable flask web application that is:

- Given a restaurant as input
- Outputs similar restaurants in a another city

YELP! (I NEED SOMEBODY...)

USING MACHINE LEARNING TO RECOMMEND RESTAURANTS

AUGUST 2, 2019

THE TEAM

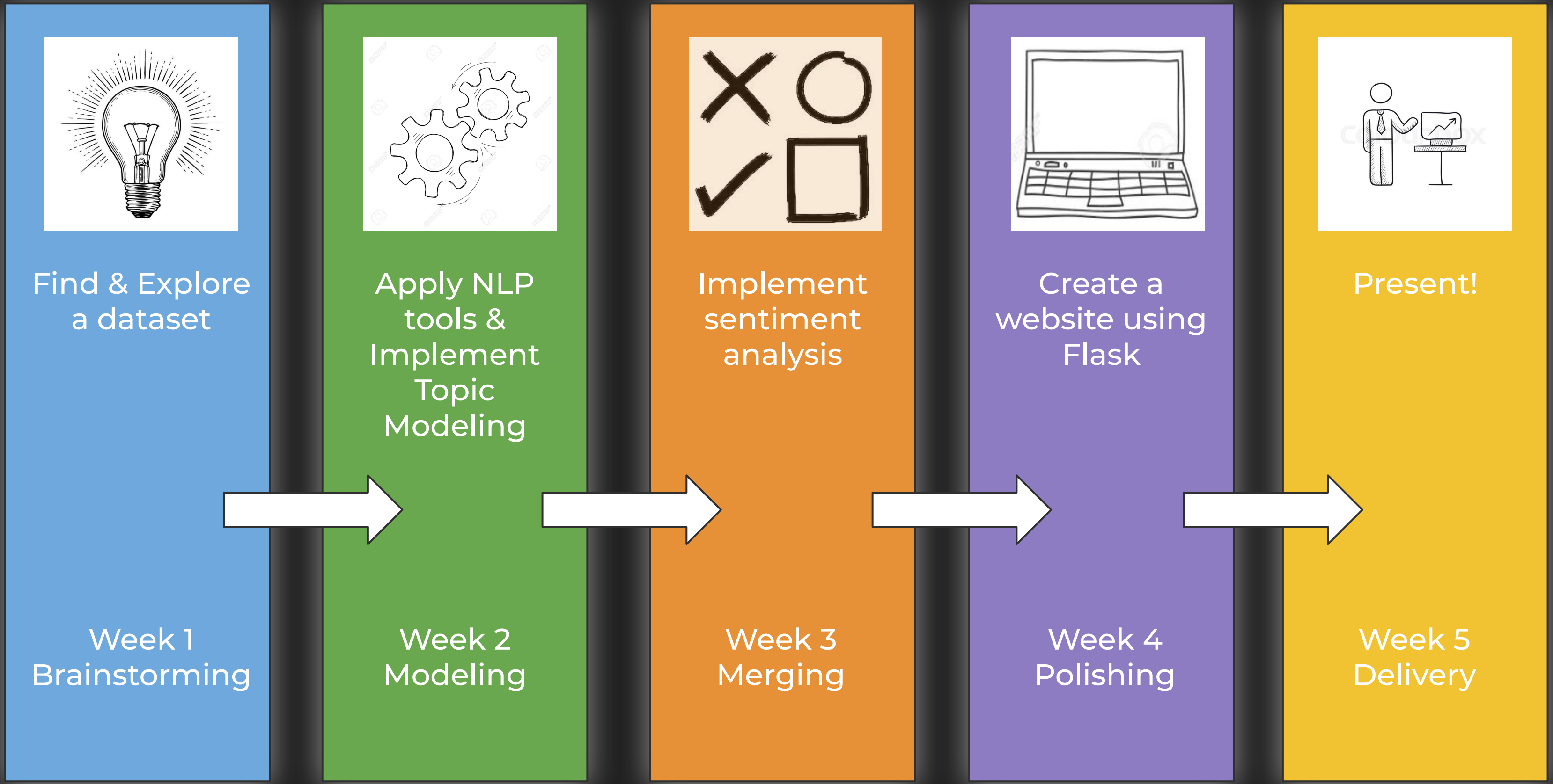
Xijie Guo
Mt. Holyoke College
South Hadley, MA

Marissa Kelley
CU Boulder
Boulder, CO

Marc Mascarenhas
South Dakota School
of Mines
Rapid City, SD

Jesse Stewart
Hofstra University
Long Island, NY

TIMELINE



Find & Explore a dataset

Week 1 Brainstorming

Apply NLP tools & Implement Topic Modeling

Week 2 Modeling

Implement sentiment analysis

Week 3 Merging

Create a website using Flask

Week 4 Polishing

Present!

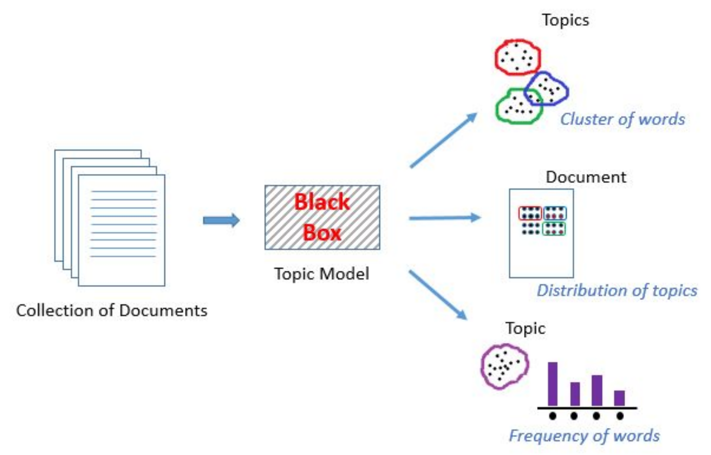
Week 5 Delivery

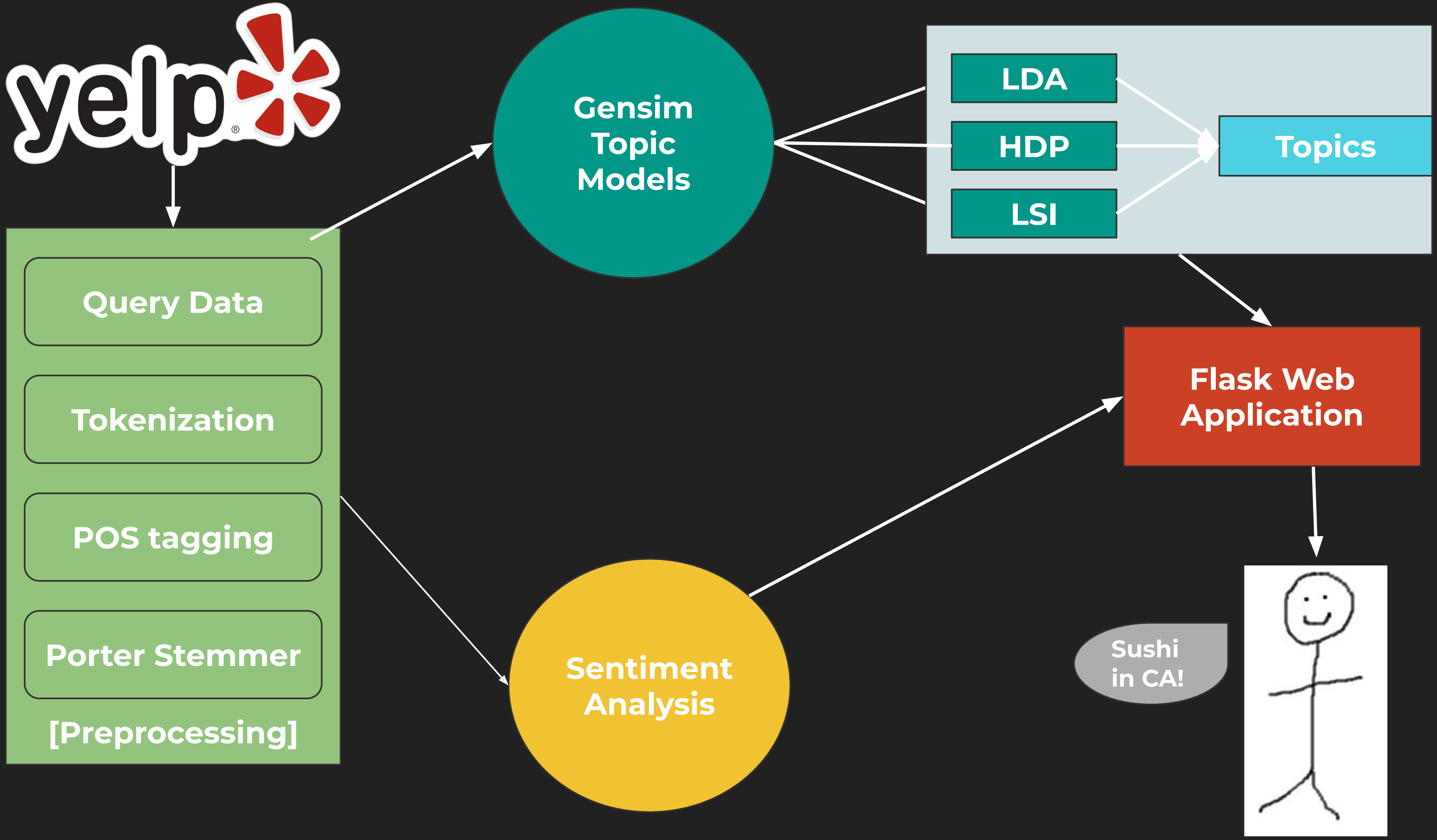
WHAT IS TOPIC MODELING?

- Traditional recommendation engines rely on defined categories

- Defined categories can be misleading, wrong, or uninformative

- Potential to uncover hidden similarities between reviews





Query Data

Tokenization

POS tagging

Porter Stemmer

[Preprocessing]

Gensim Topic Models

LDA

HDP

LSI

Topics

Sentiment Analysis

Flask Web Application

TOPIC MODELS

LDA (Latent Dirichlet Allocation)

HDP (Hierarchical Dirichlet Process)

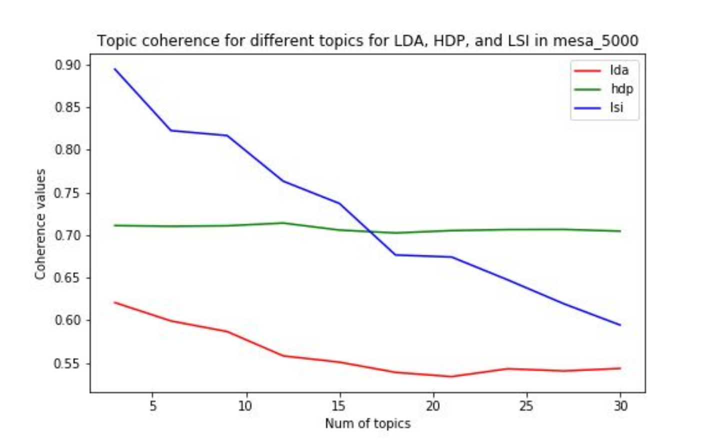
LSI (Latent Semantic Indexing)

TOPIC MODELS

- Parameter tuning
  - Number of topics
  - Number of words
  - Number of passes and iterations
- Trade-off: LSI and HDP outperforms LDA in topic coherence score BUT LDA is easier to visualize and interpret
- Similarity between two restaurants
  - KL divergence

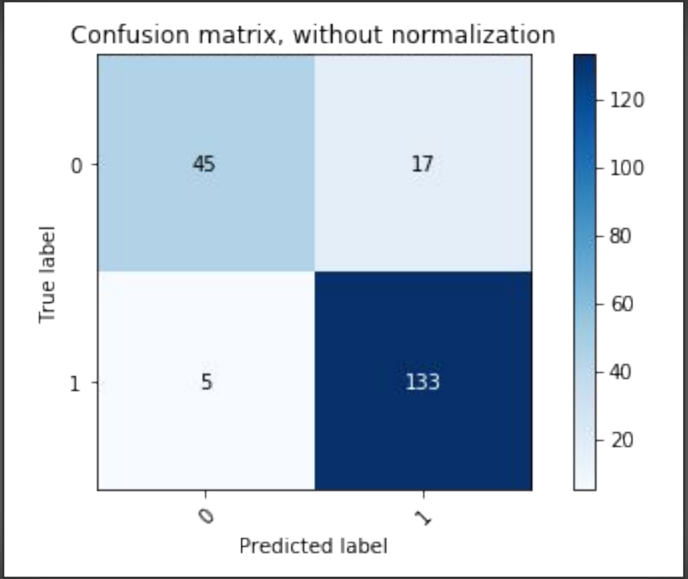
TOPIC MODELS

- Performance Metrics for topic models
    - Topic coherence score
    - Visualization
    - Manually compare results



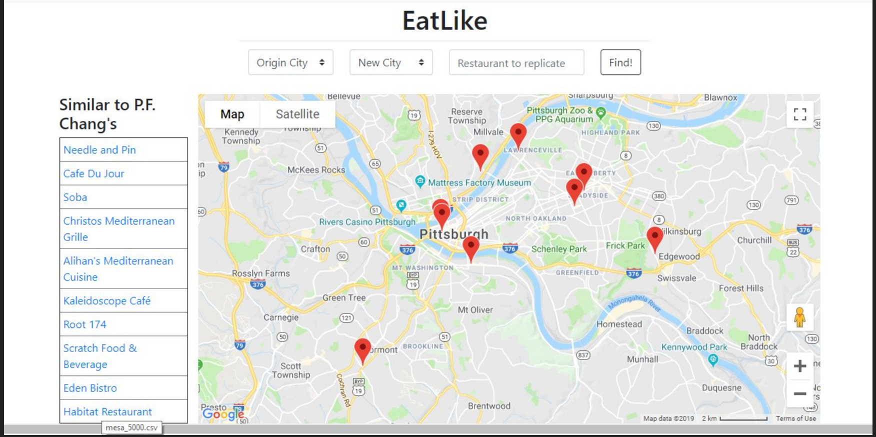
SENTIMENT ANALYSIS

- Sentence-level sentiment classification
- Uses Vader sentiment analyzer
- Extracts positive and negative sentences



RESULTS

- Topic Models
  - LDA model: 0.55 coherence score when num of topics = 6
  - HDP model: 0.7 coherence score when num of topics = 12
- Applied Sentiment Analysis
  - Precision: 88.6%
  - Recall: 96.4%
  - Accuracy: 89.0%
- Successful web app development



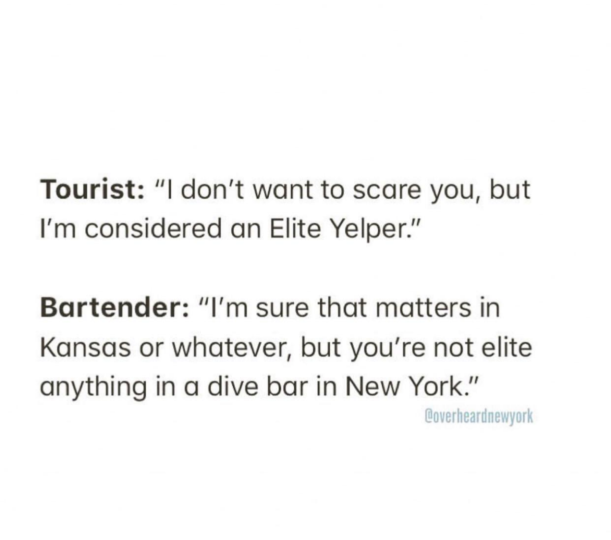
EatLike

Similar to P.F. Chang's

- Needle and Pin
- Cafe Du Jour
- Soba
- Christos Mediterranean Grille
- Alihan's Mediterranean Cuisine
- Kaleidoscope Café
- Root 174
- Scratch Food & Beverage
- Eden Bistro
- Habitat Restaurant

CHALLENGES

- Using an unsupervised model
- No real metric to evaluate results → deploy model
- Subject to human interpretation
- Finding good Yelp jokes and other words that rhyme with Yelp (besides “help”)



Tourist: “I don’t want to scare you, but I’m considered an Elite Yelper.”

Bartender: “I’m sure that matters in Kansas or whatever, but you’re not elite anything in a dive bar in New York.”

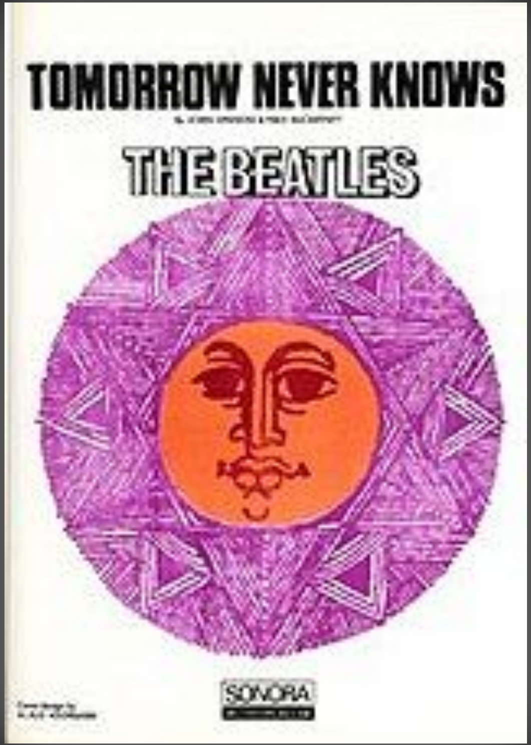
@overheardnewyork

ETHICS

- **Selection bias (non-response)**
  - Unbalanced representation of a restaurant’s ratings due to location
- **Implicit bias**
  - Linguistics (non-universal slang and/or phrases)
- **Group Attribution Bias**
  - Viewing ratings from friends or neighbors
  - Skewed inside/outside cultural preferences
- **Reporting Bias**
  - People only writing if they *really* did not like/liked a restaurant

NEXT STEPS AND FUTURE WORK

- Measure neighborhood development via restaurants
- Create a mobile app version
- Use more learning tools
  - Word cloud
- Use a database
- Deploy A/B testing
  - Showing variations of a page



THANK YOU!

yelp®

RESOURCES

SLDC Models- an overview: https://medium.com/existek/sdlc-models-explained-agile-waterfall-v-shaped-iterative-spiral-e3f012f390c5

What are Review Highlights?: https://www.yelp-support.com/article/What-are-Review-Highlights?l=en_US

An Exploratory Data Analysis (EDA) for Text Data:https://towardsdatascience.com/a-complete-exploratory-data-analysis-and-visualization-for-text-data-29fb1b96fb6a

Topic Modeling: https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/
https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/
https://www.kdnuggets.com/2016/07/text-mining-101-topic-modeling.html

Yelp Dataset Challenge Winner (sample):
https://www.yelp.com/html/pdf/YelpDatasetChallengeWinner_PersonalizingRatings.pdf?fbclid=IwAR292yTyZ4CV3zp3YVBEDeGzJ6RMszoBfGmiabGAM16JDirmBLA3vtKb_zw

Yelp Dataset Examples: https://github.com/Yelp/dataset-examples

Relevant Papers:
https://www.yelp.com/html/pdf/YelpDatasetChallengeWinner_PersonalizingRatings.pdf?fbclid=IwAR0ef70_Bn1qgoO7vCQhokeBrM8w1_6Vbqm5-7OMQOiek6-XS0p6504ZVI8

Interesting read on how Yelp data can impact others: https://www.hbs.edu/faculty/Publication%20Files/18-077_a0e9e3c7-eceb-4685-8d72-21e0f518b3f3.pdf

LDA and Document Similarity: https://www.kaggle.com/ktattan/lda-and-document-similarity

LDA Building a Missing Feature with Bars: https://towardsdatascience.com/using-lda-to-build-a-missing-yelp-feature-43436e575d65
Preprocessed Text: https://orange3-text.readthedocs.io/en/latest/widgets/preprocesstext.html

Sentiment Analysis and Applications: https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

Topic Modeling and Latent Dirichlet Allocation: https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24

Beatles font: https://fontmeme.com/the-beatles-font/

QUESTIONS?

- Team
- Timeline
- Machine Learning Diagram
- Why topic models
- Topic Models 1 2 3 4
- Sentiment Analysis
- Demo
- Results
- Challenges
- Ethics
- Next Steps & Future Work

We got by with a little yelp from our friends- the Scripps AMLI squad, Josh, Ju, David S., David B., Liza, Sidnie, Winston, Abel, and Shu. Thank you!!