Project Title: Robots Reviewing Restaurants
(Project Proposal for CS 175, Winter 2016)

List of Team Members:
Shaun McThomas, 138288643, smcthoma@uci.edu
Sean King, 82425468, smking@uci.edu
Evangeline Smith, 80193366, evangels@uci.edu

1. Project Summary

The goal of this project is to create an algorithm that can synthesize a realistic Yelp review if given a business type (restaurant, bookstore, etc.) and a star rating.

2. Problem Description and Background

When attempting review a restaurant on Yelp, sometimes the hardest part is coming up with words. What if there was a computer system to do it for you? We intend to analyse the Yelp Data set using Stanford Natural Language parser in order to develop the "standard " structures for "good"(4 or 5 star rating), "bad"(1 or 2 star rating), or "neutral"(3 star rating) reviews. We also plan to gather and classify the words used to describe a business type by sentintent (i.e. in good review "tasty" would be classified as good, and in neutral review "ok" as neutral). From here, we will some probabilistic method to create computer generated review give business type.

How are you going to do gather and classify the words?

Some potentially helpful previous Research can be found at :
https://web.stanford.edu/~jurafsky/slp3/9.pdf ,
http://www.cs.princeton.edu/courses/archive/spr05/cos126/assignments/markov.html.

3. Data Sets

We will use the Yelp dataset from the Yelp Dataset Challenge(https://www.yelp.com/dataset_challenge) to train our program. This dataset contains JSON objects describing businesses, reviews, users, check-ins and tips. For our project, we will only use the business and review information.

4. Proposed Technical Approach

This project has two distinct components: processing of the Yelp review dataset in order to build a thorough knowledge base to draw on, and then synthesizing logical, reasonable-sounding reviews using this knowledge base. For the first half of this project, we will process our dataset of Yelp reviews, analyzing the sentiment of the review content based on star rating. We further connect keywords from the review to the type of business being reviewed.

We plan to approach the problem of correct syntax and correct word selection separately using a pair of markov chains.

The first chain will be used for syntax. It will not be associated with specific words; it will only recognize the parts-of-speech(POS) tags. The markov model will map n-grams of POS tags to the probability of successor POS tag n-grams. This POS markov chain will be used to generate the syntactically correct skeletons of our synthesized sentences.

The second markov chain will be associated with actual words. N-grams of words will be mapped to a collection of multiple bags of words(BOWs). Each N-gram will have a separate bag of words for each part-of-speech, further separated by sentiment and business type .

Using these chains together, a syntactically correct POS-skeleton for a sentence will be generated by the first markov chain. The second markov chain will then populate this skeleton with contextual words for each part of speech. For each POS-tag in the skeleton, the second chain will look at the preceding N-gram of words, and from that N-grams' bag of words for that POS-tag, pull out a word based on probability.

5. Experiments and Evaluation

We plan to evaluate our algorithm's performance with a user study. We will administer an online test to our friends (and possibly general public), where they will be given sets of reviews generated by different versions of the algorithm and asked to select the most realistic review. A particular version of the algorithm will be given a point every time one of it's synthesized reviews are selected. We also plan have controls tests. These test will place generated reviews side by side with actual user reviews from the yelp dataset, and the evaluators will choice the human created content.

Not sure what it means

6. Software

We plan on writing the majority of our project in Python(some parts may be written in Java), and we will use the following pieces of public software for our project:

- Numpy Library
- NLTK Library
- Stanford Parser (and associated Java-Python conversion tool)
- Yelp Acidemic Dataset Examples https://github.com/Yelp/dataset-examples

We also plan on developing the following to help with our project:

- A Markov Model specifically designed for POS-tag selection.
- A Markov Model specifically designed for word selection.
- An online voting system to be used for human-evaluation of our algorithm.

7. Milestones

- Weeks 5 and 6 :
  - Parse and tag user reviews from Yelp dataset.
  - Develop algorithm to generate markov chain for determining part-of-speech syntax.
- Weeks 7 and 8 :
  - Develop algorithm to generate markov chain for word mapping to POS skeleton.
  - Classify words as topic-words by their tf-idf.
    - Map these words to their respective topic, and replace them with generic <topicword> tokens for the word markov model.
  - Be able to generate rough syntactically/contextually correct sentences given a business type and sentiment level.
- Weeks 9 and 10:
  - Develop test for human evaluation.
  - Improve the realism of our natural language synthesis.

8. Individual Student Responsibilities

All Member: Will aide other member if they fall behind in these assigned task. Try to make each aspect of this project a team effort as much as possible.

Shaun McThomas: Write and test the parsing code for data set,write and test code for classifying words as topic-words by their tf-idf, will assist in doing experiments and interpreting results, and will assist in writing project reports.
Sean King: Write and test the markov chain models for creation of part-of-speech sentence skeleton and vocabulary selection. Will tie the formatted yelp data and markov chain sections together. Assist with testing and evaluation of algorithm performance, and project reports.
Evangeline Smith: Write and test the markov chain algorithm for mapping words to POS skeleton, write scripts for evaluating the accuracy of the topic-word classifying algorithm, will assist in writing project reports.

[Note these are just suggestions – These assigned responsibilities will change throughout project progress.]