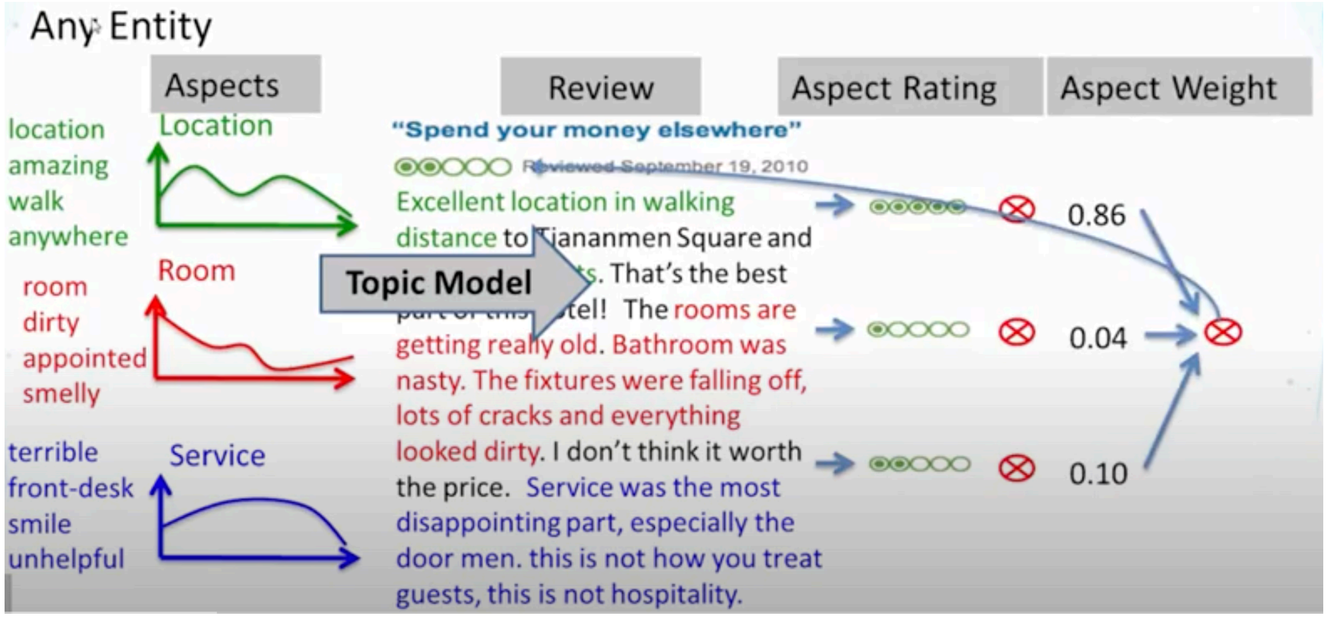
Reproducing a Paper: Latent Aspect Rating Analysis without keyword supervision

Rakesh Patnaik (rakeshp2@illinois.edu) 12/13/2020

Visual depiction of the task



- Input
  - Review texts
  - Overall rating
  - Assumed aspects in the review (Location, Room, Service etc)
- Output
  - Latent aspects (Topic model used to extract text from review corresponding to a topic)
  - Rating associated to each latent aspect
  - Weight associated to each latent aspect
- Validation
  - Mean squared error from ground truth overall rating.

Stages in the process

- Pre-processing (preprocessing_Sec5_1.py)
  - Lowercase
  - Remove punctuation characters
  - Remove stop words
  - Lemmatize
- Processing and Analyzing (Main.py)
  - Model topics based on "Service", "Cleanliness", "Overall", "Value", "Location", "Rooms", "Sleep Quality"
  - Identify words that correlate to model topics
  - Use regression to identify topic rating to maximize probability to ground truth latent ratings
  - Use regression to identify topic weights to maximize probability to ground truth overall rating
  - Calculate mean squared error to ground truth ratings
  - Output results to results/results.txt and MSE to stdout.

How to run the code

- git clone https://github.com/rakesh-patnaik/CourseProject.git
- cd CourseProject
- python3 -m venv env
- source env/bin/activate
- pip install --upgrade pip
- python -m pip install -r requirements.txt
- python -m nltk.downloader stopwords
- python -m nltk.downloader punkt
- python -m nltk.downloader wordnet
- python preprocessing_Sec5_1.py
- python Main.py

Results

- Results will be output to results/results.txt
- Mean Squared Error will be output to stdout
  - (env) rakesh@Rakeshs-MacBook-Pro-4.local:~/work/uiuc-mcsds/cs410-fall2020/CourseProject$ python preprocessing_Sec5_1.py
    (env) rakesh@Rakeshs-MacBook-Pro-4.local:~/work/uiuc-mcsds/cs410-fall2020/CourseProject$ python Main.py
    Total reviews: 183
    MSE: 2.99805326964421