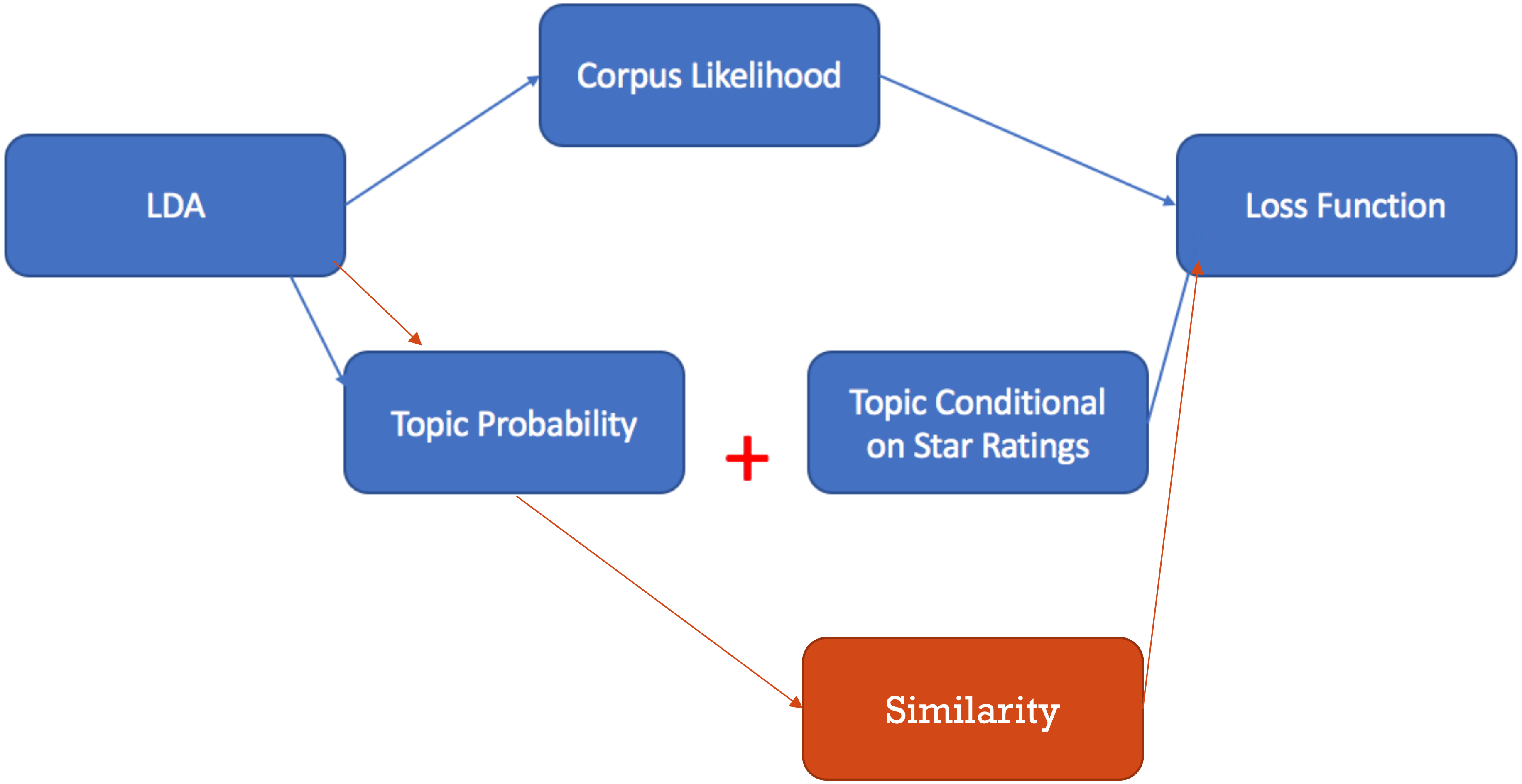
WEEK 5 PROGRESS REPORT

ROAD MAP



Corpus Likelihood
LDA
Loss Function
Topic Probability
+
Topic Conditional on Star Ratings
Similarity

PROGRESS

Week 5

- Implemented the Paper: Personalizing Yelp Star Rating:

  -----Add codeword after all the negative or positive words and manually label the topic with adjective (good or bad)

- Implemented the Paper: Hidden dimension of rating

Week 6

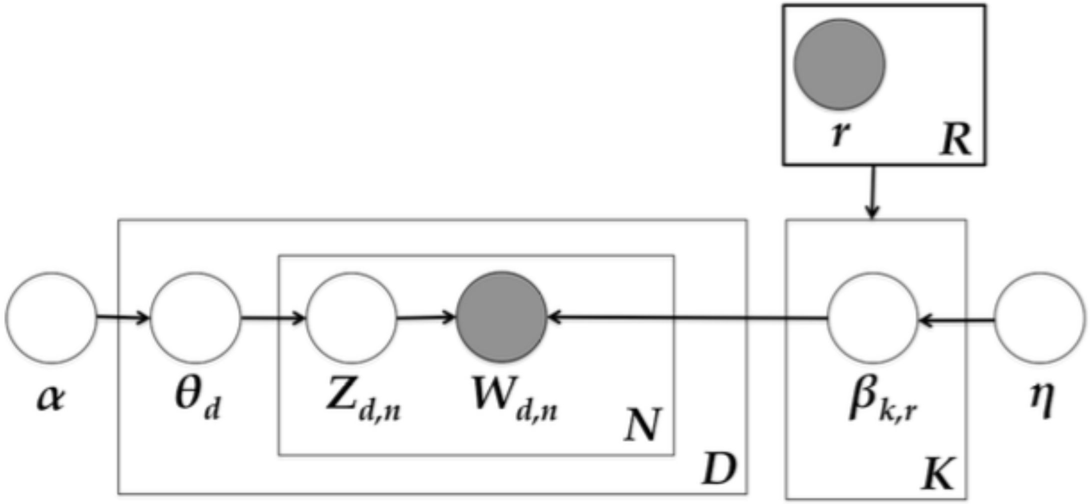
- Compared the results with traditional LDA
- Combine the results based on two papers above:

  ----Use the topic and preference we get from LDA to initialize rating dimensions and minimize the loss function by gradient descent.

PAPER 1: PERSONALIZING YELP STAR RATING

- Motivation: Traditional topic modeling lacks methods of incorporating star ratings or semantic analysis in the generative process
- Method: Modified LDA – term distributions of topics are conditional on star ratings.

PAPER 1 METHOD



- Then a more appropriate LDA would model the conditional dependence between a rating *r* and *bk*.
- The way to implement the method: codeword

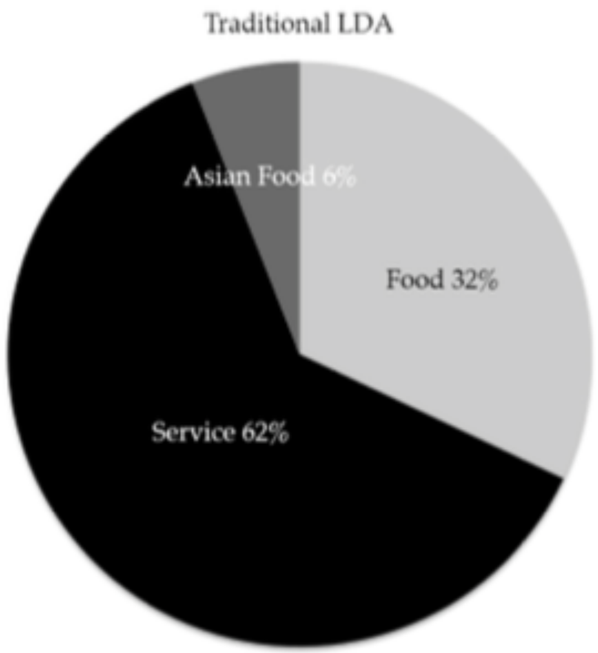
CODEWORD

- Find a dictionary of negative and positive stemmed words respectively
- Modify the corpus to include a *codeword*, "GOODREVIEW" or "BADREVIEW," after each positive or negative word, respectively

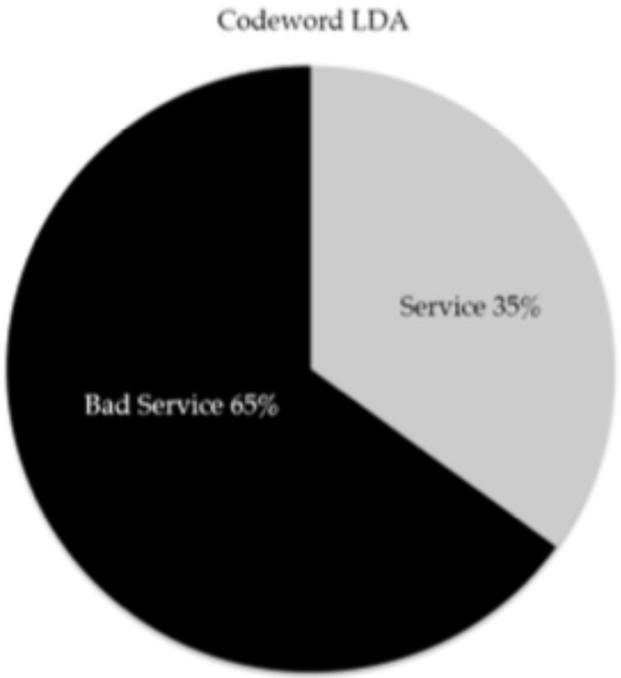
- *awesome car mainten famili servic honest fair priced*

- *GOODREVIEW car mainten famili servic honest GOODREVIEW fair GOODREVIEW priced*

COMPARE



Traditional LDA
Asian Food 6%
Food 32%
Service 62%



Codeword LDA
Service 35%
Bad Service 65%

PAPER 2: HIDDEN DIMENSIONS

$$f(\mathcal{T}|\Theta, \Phi, \kappa, z) = \sum_{r_{u,i} \in \mathcal{T}} \underbrace{(rec(u,i) - r_{u,i})^2}_{\text{rating error}} - \mu \underbrace{l(\mathcal{T}|\theta, \phi, z)}_{\text{corpus likelihood}}.$$

SIMILARITY

- A model that links LDA with constraints derived from document relative similarities. Specifically, in our model, the constraints act as a regulorization term of the log likelihood of LDA.

$$\mathfrak{L}(\boldsymbol{\lambda}) = \underbrace{\log p(\boldsymbol{w}|\boldsymbol{\lambda}, \boldsymbol{\beta})}_{\text{log likelihood}} + \underbrace{\log p(\boldsymbol{\lambda}|(\mathbf{0}, \sigma^2 \mathbf{I}))}_{\text{prior}} - \underbrace{\eta \sum_{i=1}^{T} \mathcal{L}_i(d_i, d_i^+, d_i^-)}_{\text{hinge loss}}$$

TRANSFORMATION BETWEEN TOPIC AND RATING

$$\theta_{i,k} = \frac{\exp(\kappa\gamma_{i,k})}{\sum_{k'} \exp(\kappa\gamma_{i,k'})}.$$

By linking the two, we hope that if a product exhibits a certain property (high$\theta_{i,k}$), this will correspond to a particular topic being discussed (high$\gamma_{i,k}$).