Improving businesses with topic modeling

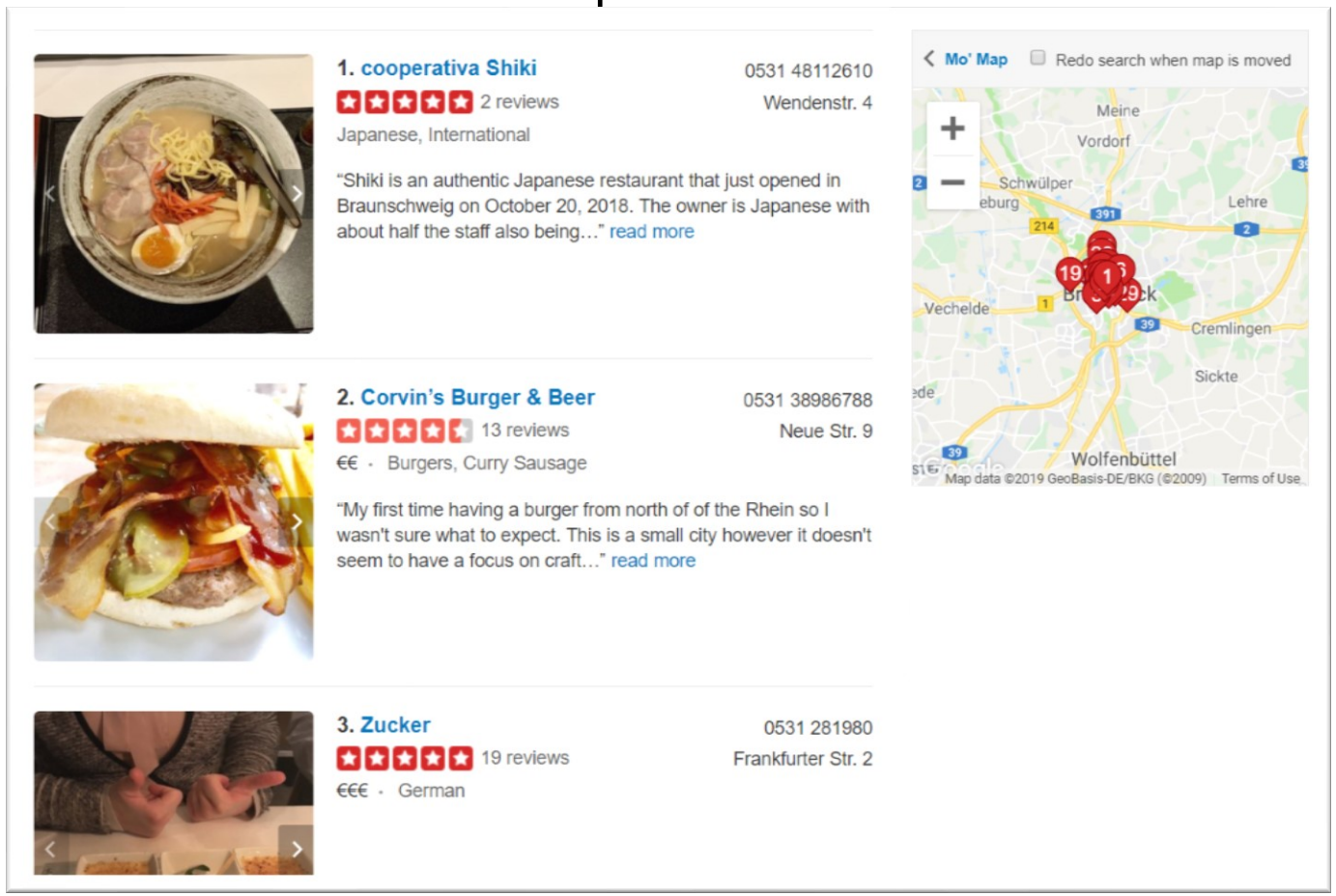
on the Yelp dataset

Elvira Pupka-Lipinski

Content

1. Yelp

- Crowd-sourced review forum
- Offers a recommendation portal for restaurants and shops
- Short form of **Yel**low **P**ages



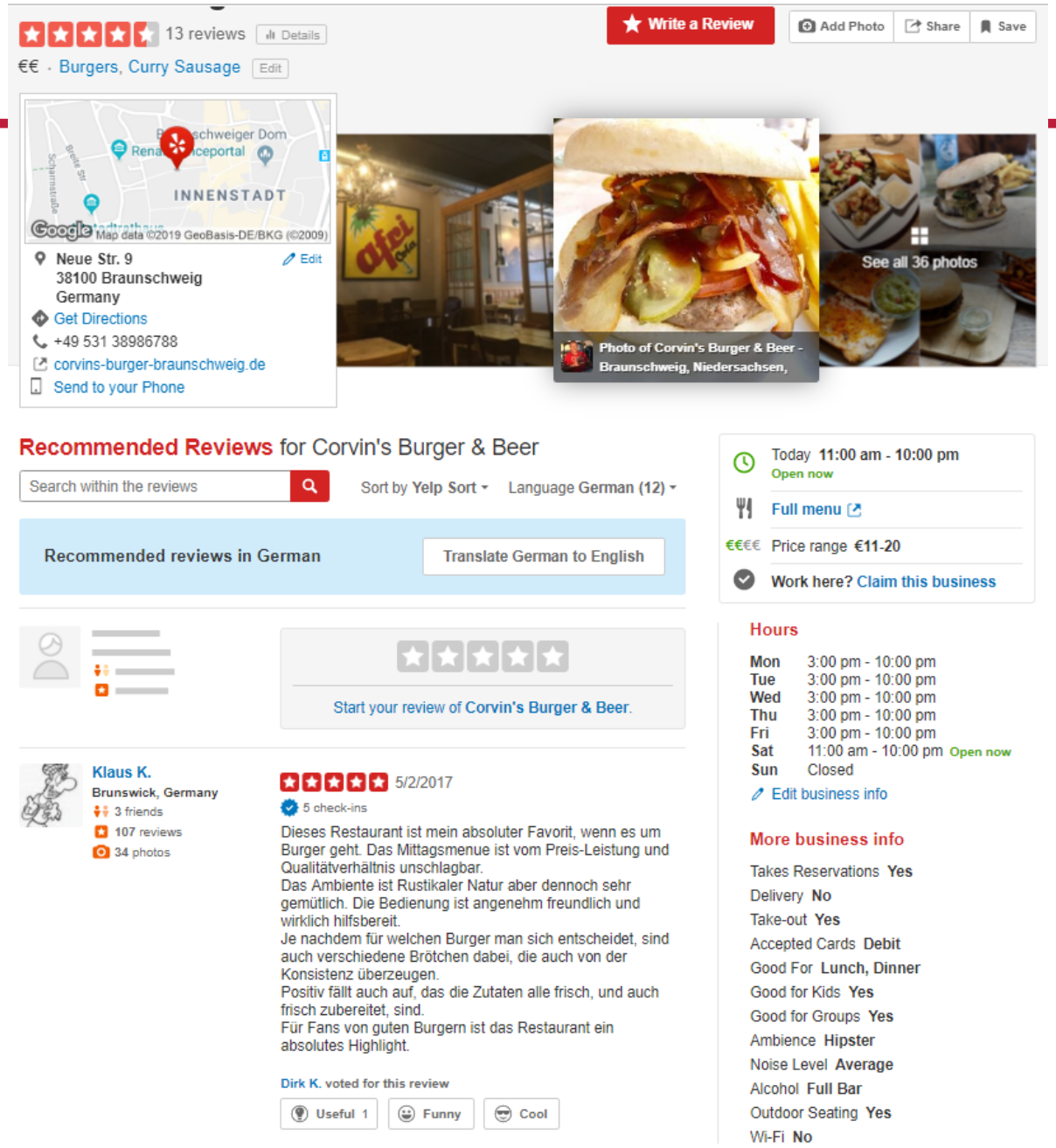
https://www.yelp.com/search?find_desc=Restaurants&find_loc=Braunschweig%2C+Niedersachsen%2C+Deutschland&ns=1

1. Yelp

- Yelp: crowd-sourced review forum
- Offers a recommendation portal for restaurants and shops
- Short form of **Yel**low **P**ages

Corvin‘s Burger & Beer in Braunschweig

https://www.yelp.com/biz/corvins-burger-und-beer-braunschweig-2?osq=Restaurants

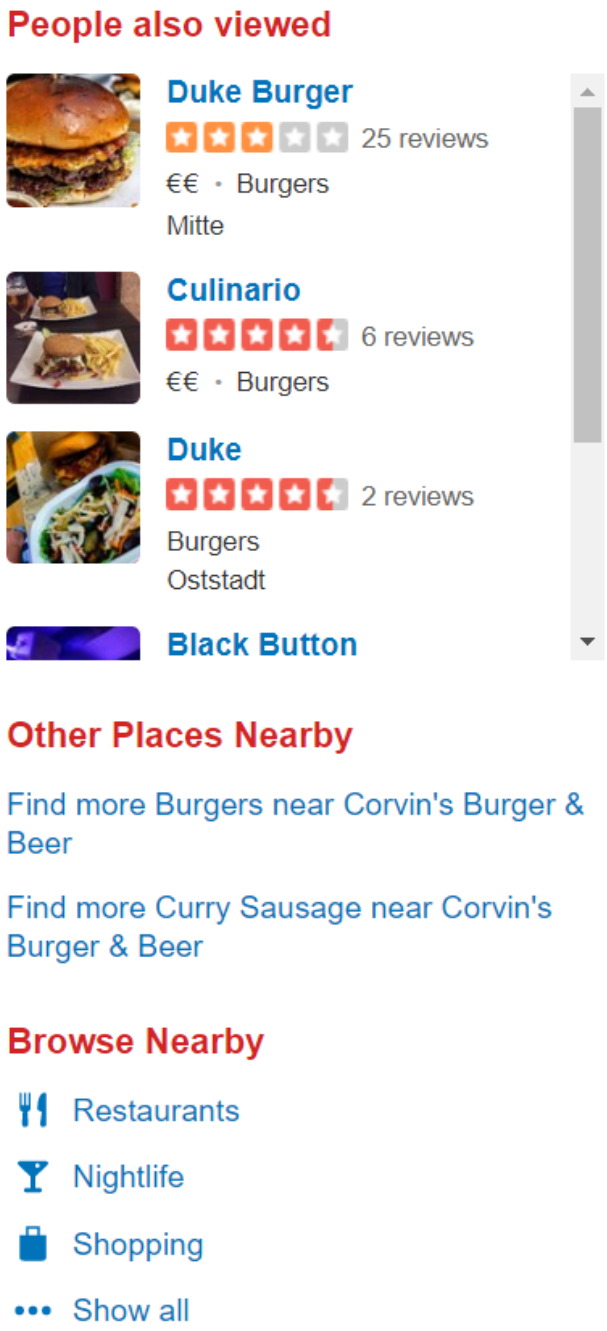


1. Yelp

- Yelp: crowd-sourced review forum
- Offers a recommendation portal for restaurants and shops
- Short form of **Yel**low **P**ages

Corvin‘s Burger & Beer in Braunschweig

https://www.yelp.com/biz/corvins-burger-und-beer-braunschweig-2?osq=Restaurants



1. Yelp

- Yelp: crowd-sourced review forum
- Offers a recommendation portal for restaurants and shops
- Short form of **Yel**low **P**ages
- Yelp offers their dataset for analyzing the data and sharing the discoveries (https://www.yelp.com/dataset/challenge)

2. Possible tasks

- Decision guidance for Users
  - Determination of the 3 best dishes in a Restaurant (already exists)
  - Suggestions for businesses due to users‘ previous patterns
    - Beneficial while traveling
  - Suggestions for business due to similar users (clustering)

- Determination of the probability of closure of a business
  - Not possible due to the lack of needed data

- Improvements for business
  - Opening new business? What do the citizen in Pittsburgh want? What is important to consider?
  - What can I as a business owner improve?

2. Possible tasks

- Decision guidance for Users
  - Determination of the 3 best dishes in a Restaurant (already exists)
  - Suggestions for businesses due to users‘ previous patterns
    - Beneficial while traveling
  - Suggestions for business due to similar users (Clustering)

- Determination of the probability of closure of a business
  - Not possible due to the lack of needed data

- Improvements for business
  - Opening new business? What do the citizen in Pittsburgh want? What is important to consider?
  - **What can I as a business owner improve?**

3. Approach: What can I as a business owner improve?

1. Building a pipeline allowing business owners to answer this question
2. Suggesting improvements for business ID: 'c0yPNU-BqS65u0vIKP7P0w‘ as an example

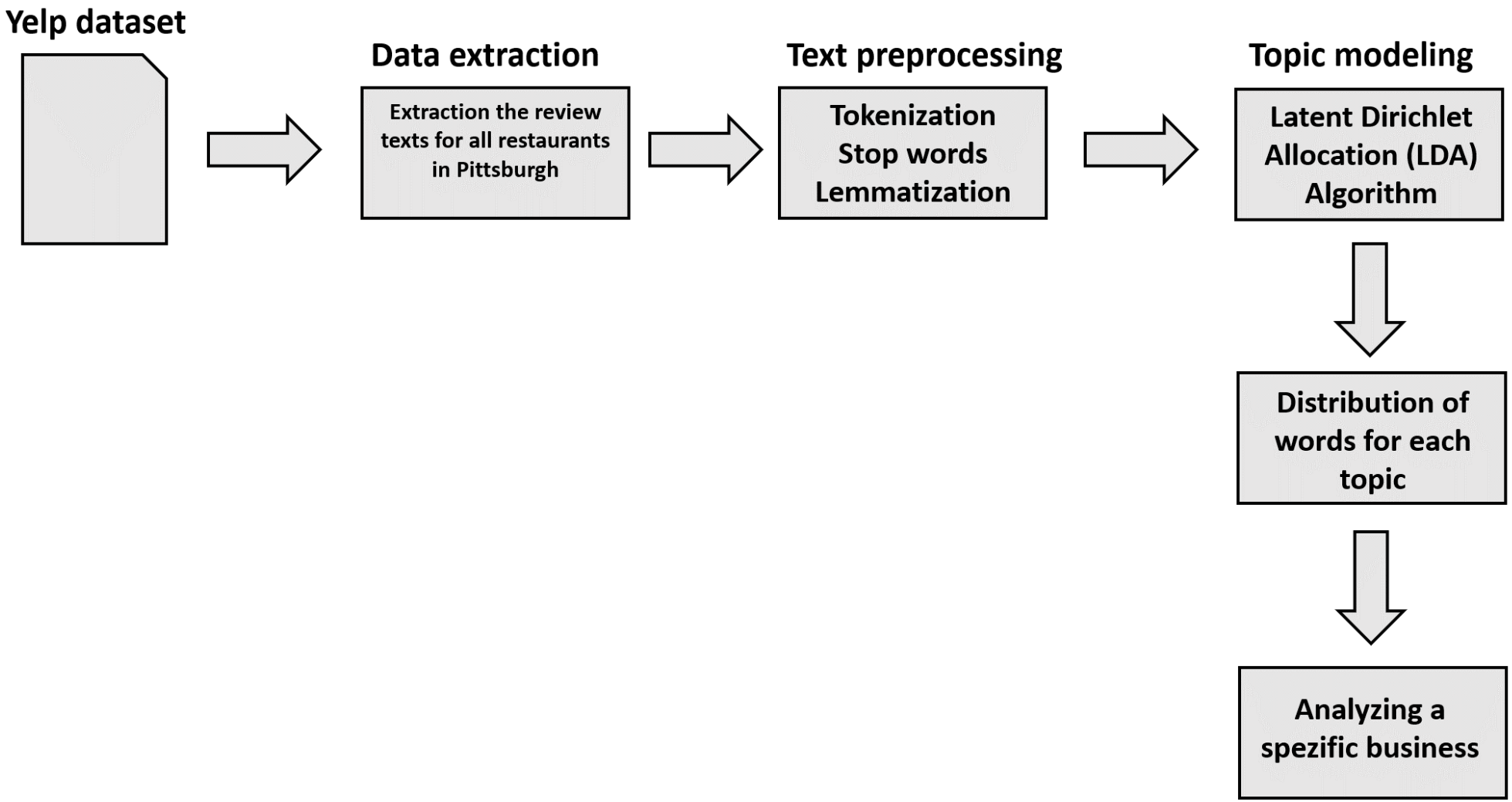
3. Business improvements

- Improvements for business
  - What can I as a business owner improve?
    - Idea: Extracting the topics the owner needs to improve
    - Approach: Topic modeling

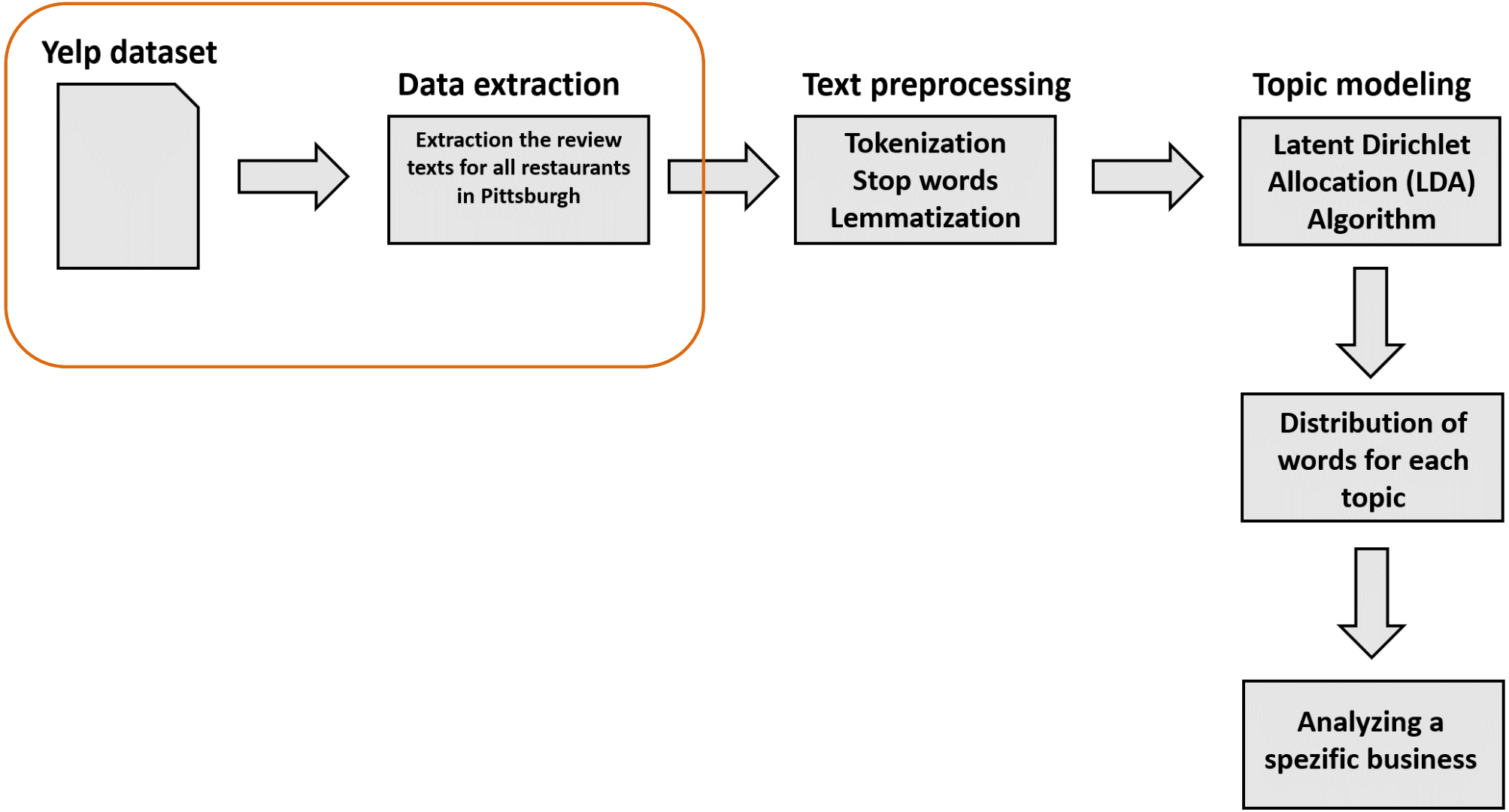
Topic	Avg Stars
Food	4.0
Service	2.0
Location	4.0

→ **Improvement in service is needed**

3. Workflow


Yelp dataset
Data extraction
Extraction the review texts for all restaurants in Pittsburgh
Text preprocessing
Tokenization
Stop words
Lemmatization
Topic modeling
Latent Dirichlet Allocation (LDA) Algorithm
Distribution of words for each topic
Analyzing a spezific business

3. Workflow


Yelp dataset
Data extraction
Extraction the review texts for all restaurants in Pittsburgh
Text preprocessing
Tokenization
Stop words
Lemmatization
Topic modeling
Latent Dirichlet Allocation (LDA) Algorithm
Distribution of words for each topic
Analyzing a spezific business

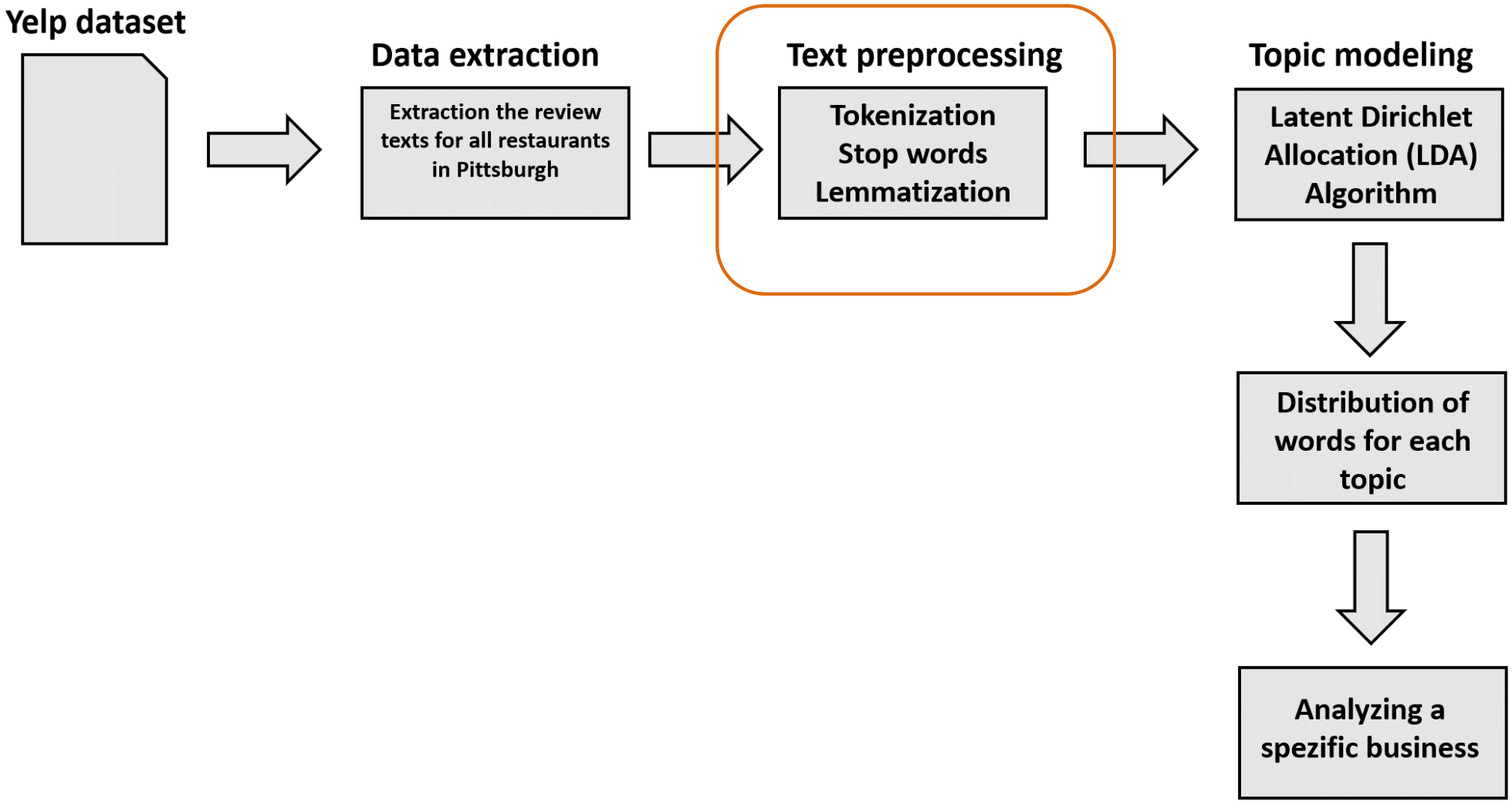
3. Yelp dataset – Data Extraction

- Datasets
  - **business.json:** business data, location, attributes and categories (business_id)
  - **review.json:** full review data such as text, date and stars (review_id, user_id, business_id)
  - **user.json**: user information such as the average star rating and the users' friends (user_id)
  - **checkin.json:** checkin data: the visits' date and business (business_id)
  - **tip.json:** short text on a business (business_id, user_id)
  - **photo.json:** caption and classification (photo_id, business_id)

3. Yelp dataset – Data Extraction

- Datasets
  - **business.json:** business data, location, attributes and categories (business_id)
  - **review.json:** full review data such as text, date and stars (review_id, user_id, business_id)
  - **user.json**: user information such as the average star rating and the users' friends (user_id)
  - **checkin.json:** checkin data: the visits' date and business (business_id)
  - **tip.json:** short text on a business (business_id, user_id)
  - **photo.json:** caption and classification (photo_id, business_id)

3. Workflow


Yelp dataset
Data extraction
Extraction the review texts for all restaurants in Pittsburgh
Text preprocessing
Tokenization
Stop words
Lemmatization
Topic modeling
Latent Dirichlet Allocation (LDA) Algorithm
Distribution of words for each topic
Analyzing a spezific business

3. Text preprocessing

“I definitely enjoyed my meal at avenue b, but with a meal that comes at a hefty price, I don't know if it's worth another trip. If I'm paying that much, the food BETTER be mind blowing.”

- **Lowering all characters:**

→ “i definitely enjoyed my meal at avenue b, but with a meal that comes at a hefty price, i don’t know if it's worth another trip. if i'm paying that much, the food better be mind blowing.”

- **Removing all punctuation:**

→ “i definitely enjoyed my meal at avenue b but with a meal that comes at a hefty price i don't know if it worth another trip if i paying that much the food better be mind blowing”

- **Tokenization:** Tokenization is the process of splitting the given text into smaller pieces called tokens.

→ ['i', 'definitely', 'enjoyed', 'my', 'meal', 'at', 'avenue', 'b', 'but', 'with', 'a', 'meal', 'that', 'comes', 'at', 'a', 'hefty', 'price', 'i', 'do', "n't", 'know', 'if', 'it', 'worth', 'another', 'trip', 'if', 'i', 'paying', 'that', 'much', 'the', 'food', 'better', 'be', 'mind', 'blowing’]

3. Text preprocessing

- **Tokenization:** Tokenization is the process of splitting the given text into smaller pieces called tokens.

→['i', 'definitely', 'enjoyed', 'my', 'meal', 'at', 'avenue', 'b', 'but', 'with', 'a', 'meal', 'that', 'comes', 'at', 'a', 'hefty', 'price', 'i', 'do', "n't", 'know', 'if', 'it', 'worth', 'another', 'trip', 'if', 'i', 'paying', 'that', 'much', 'the', 'food', 'better', 'be', 'mind', 'blowing‘]

- **Removing stop words:**

→['definitely’, 'enjoyed’, 'meal’, 'avenue’, 'b’, 'meal’, 'comes’, 'hefty’, 'price’, 'n't’, 'know’, 'worth’, 'trip’, 'paying’, 'food’, 'mind’, 'blowing']

3. Text preprocessing

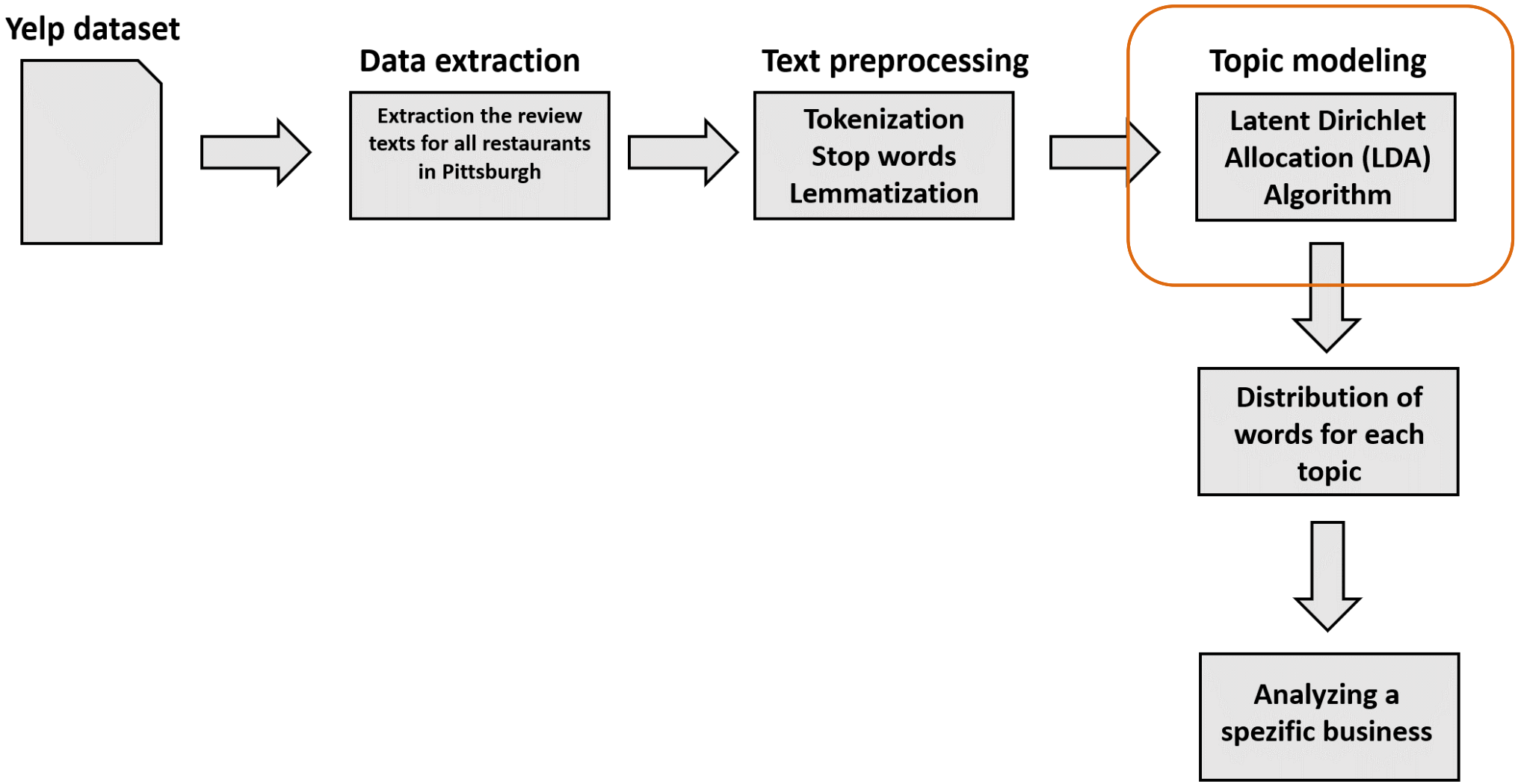
- **Stop words:**

→['definitely', 'enjoyed', 'meal', 'avenue', 'b', 'meal', 'comes', 'hefty', 'price', 'n't', 'know', 'worth', 'trip', 'paying', 'food', 'mind', 'blowing']

- **Lemmatization**: reduce inflectional forms to a common base form

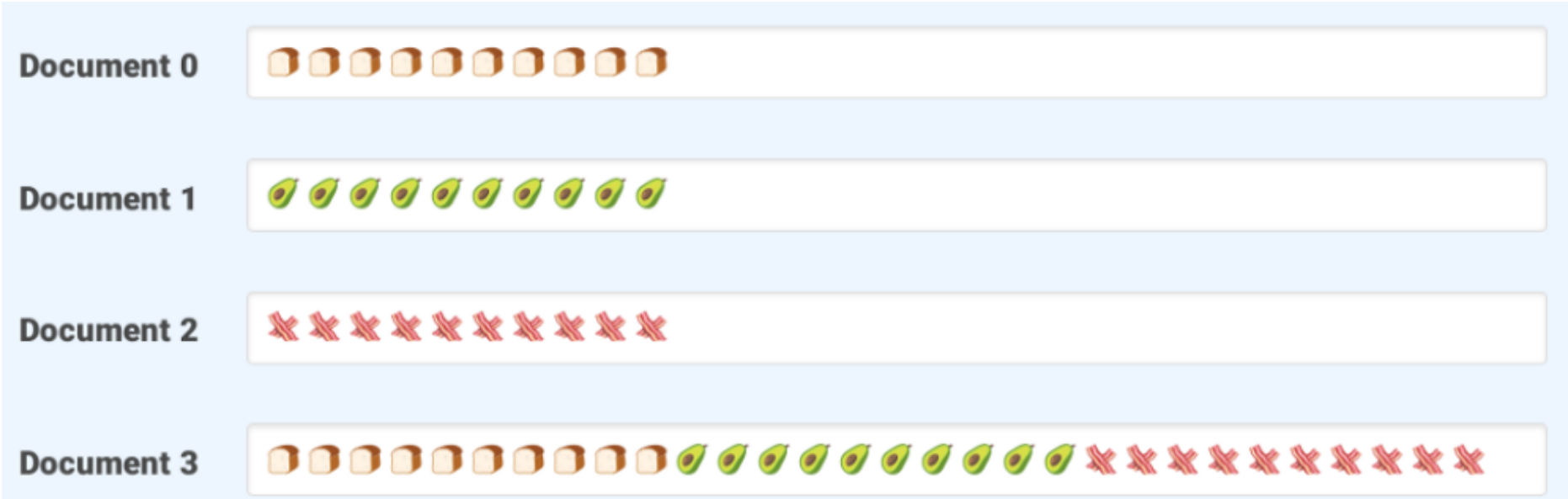
→['definitely', 'enjoy', 'meal', 'avenue', 'b', 'meal', 'come', 'hefty', 'price', 'n't', 'know', 'worth', 'trip' 'pay', 'food', 'mind', 'blow']

3. Workflow


Yelp dataset
Data extraction
Extraction the review texts for all restaurants in Pittsburgh
Text preprocessing
Tokenization
Stop words
Lemmatization
Topic modeling
Latent Dirichlet Allocation (LDA) Algorithm
Distribution of words for each topic
Analyzing a spezific business

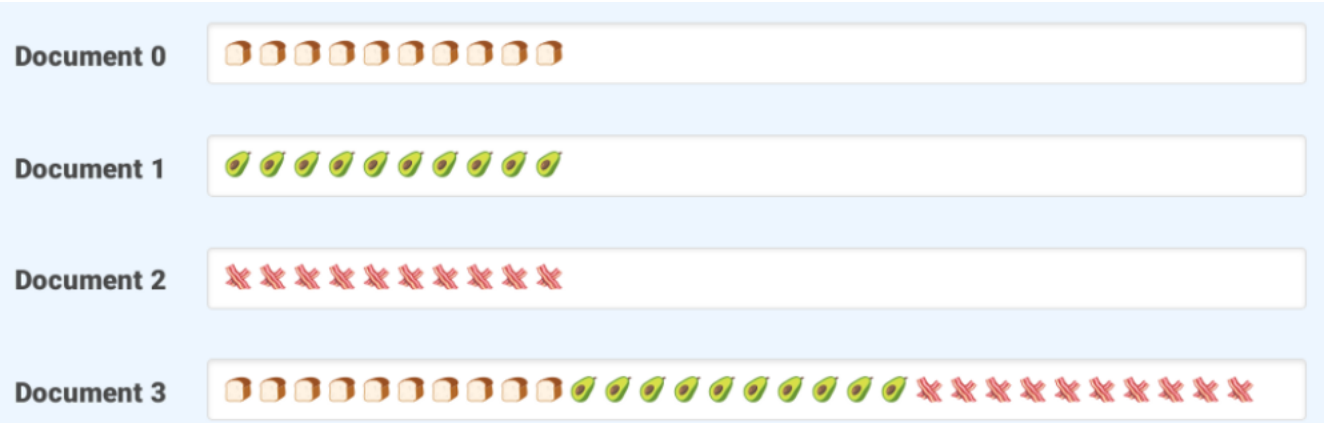
3. Approach: Latent Dirichlet Allocation

- Generative statistical model for topic discovery
- Unsupervised learning



https://medium.com/@lettier/how-does-lda-work-ill-explain-using-emoji-108abf40fa7d

3. Approach: Latent Dirichlet Allocation



Word vs.Topic Matrix (θ)

	Topic 0	Topic 1	Topic 2
	0.000	0.000	0.999
	0.999	0.000	0.000
	0.000	0.999	0.000
	Topic 0	Topic 1	Topic 2

Document vs.Topic Matrix (ϕ)

	Topic 0	Topic 1	Topic 2
Document 0	0.030	0.030	0.939
Document 1	0.939	0.030	0.030
Document 2	0.030	0.939	0.030
Document 3	0.333	0.333	0.333
	Topic 0	Topic 1	Topic 2

https://medium.com/@lettier/how-does-lda-work-ill-explain-using-emoji-108abf40fa7d

3. Approach: Latent Dirichlet Allocation

- LDA's assumption on how documents are generated:

1. Determine a unique set of words, determine amount of documents and the amount of words per document, determine amount of topics, determine α & β
2. Calculation the probability of each word per topic (Word vs.Topic Matrix (θ))
   - drawing a Dirichlet distribution for each topic
   - hyperparameter β: controls the distribution of words in topics

Word vs.Topic Matrix (θ)

	Topic 0	Topic 1	Topic 2
🍞	0.000	0.000	0.999
🥑	0.999	0.000	0.000
🌶	0.000	0.999	0.000
	Topic 0	Topic 1	Topic 2

https://medium.com/@lettier/how-does-lda-work-ill-explain-using-emoji-108abf40fa7d

3. Approach: Latent Dirichlet Allocation

- LDA's Assumtion on how documents are generated:

3. Calculation the probability of each Topic per Document (Document vs.Topic Matrix (ϕ))
   - drawing a Dirichlet distribution for each document
   - hyperparameter α: controls the mixture of topics for any given document

Document vs.Topic Matrix (ϕ)

	Topic 0	Topic 1	Topic 2
Document 0	0.030	0.030	0.939
Document 1	0.939	0.030	0.030
Document 2	0.030	0.939	0.030
Document 3	0.333	0.333	0.333
	Topic 0	Topic 1	Topic 2

https://medium.com/@lettier/how-does-lda-work-ill-explain-using-emoji-108abf40fa7d

3. Approach: Latent Dirichlet Allocation

- LDA's Assumtion on how documents are generated:

4. Building the actual document

	Topic 0	Topic 1	Topic 2
🍞	0.000	0.000	0.999
🥑	0.999	0.000	0.000
🥓	0.000	0.999	0.000
	Topic 0	Topic 1	Topic 2

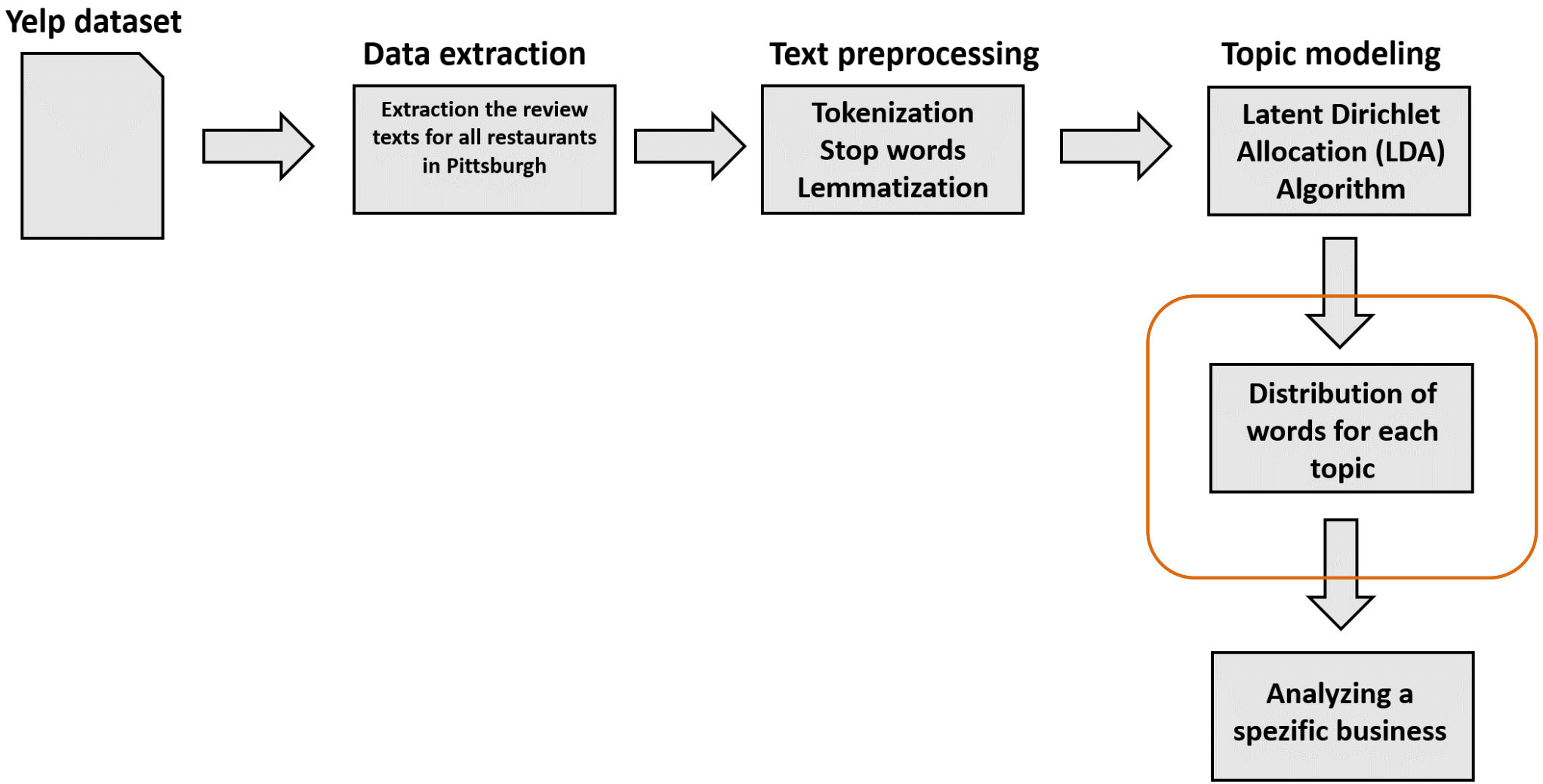
	Topic 0	Topic 1	Topic 2
Document 0	0.030	0.030	0.939
Document 1	0.939	0.030	0.030
Document 2	0.030	0.939	0.030
Document 3	0.333	0.333	0.333
	Topic 0	Topic 1	Topic 2

https://medium.com/@lettier/how-does-lda-work-ill-explain-using-emoji-108abf40fa7d

3. How does LDA works for topic modeling?

- $p(Topics, \theta, \phi \mid Documents)$
  - variables depend on each other
  - NP-hard problem
  - → approximation of p using variance inference
    - distribution $q(Topics, \theta, \phi \mid Documents)$
    - minimizing Kullback-Leibler divergence

3. Workflow


Yelp dataset
Data extraction
Extraction the review texts for all restaurants in Pittsburgh
Text preprocessing
Tokenization
Stop words
Lemmatization
Topic modeling
Latent Dirichlet Allocation (LDA) Algorithm
Distribution of words for each topic
Analyzing a spezific business

4. Results

- Top 10 words in topics for Pittsburgh and Restaurants:

  - Topic 0: ['food', 'place', 'service', 'try', 'restaurant', 'price', 'chicken', 'time', 'come', 'like']
  - Topic 1: ['pizza', 'burger', 'order', 'fry', 'cheese', 'like', 'sauce', 'wing', 'eat', 'place']
  - Topic 2: ['taco', 'dish', 'delicious', 'order', 'flavor', 'sauce', 'thai', 'restaurant', 'meal', 'dessert‘]
  - Topic 3: ['food', 'time', 'order', 'place', 'come', 'service', 'table', 'wait', 'drink', 'bar']
  - Topic 4: ['place', 'sandwich', 'beer', 'pittsburgh', 'coffee', 'love', 'like', 'breakfast', 'brunch', 'selection‘]

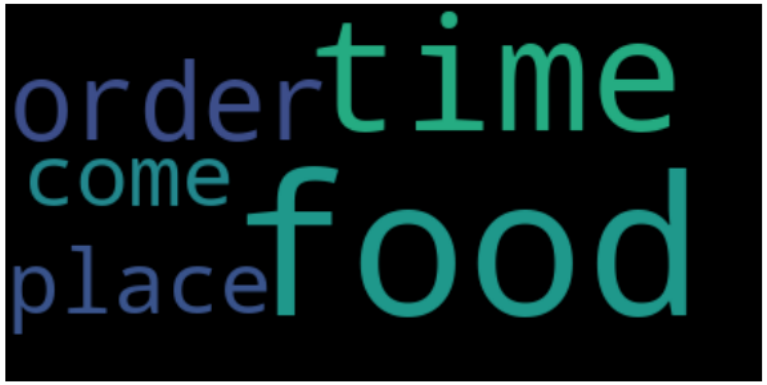
4. Results

Top words of Topic 0

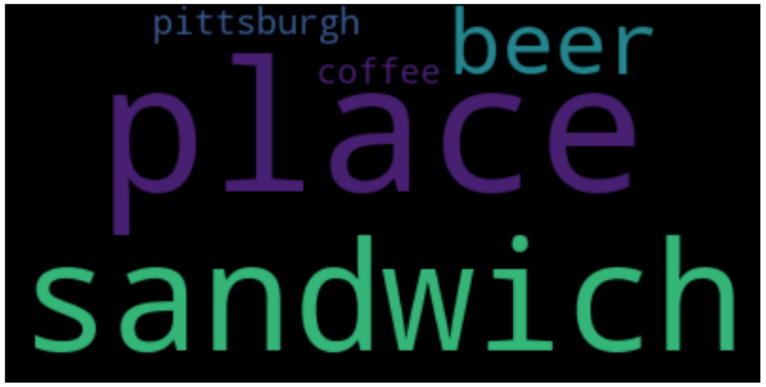
Top words of Topic 1

Top words of Topic 2

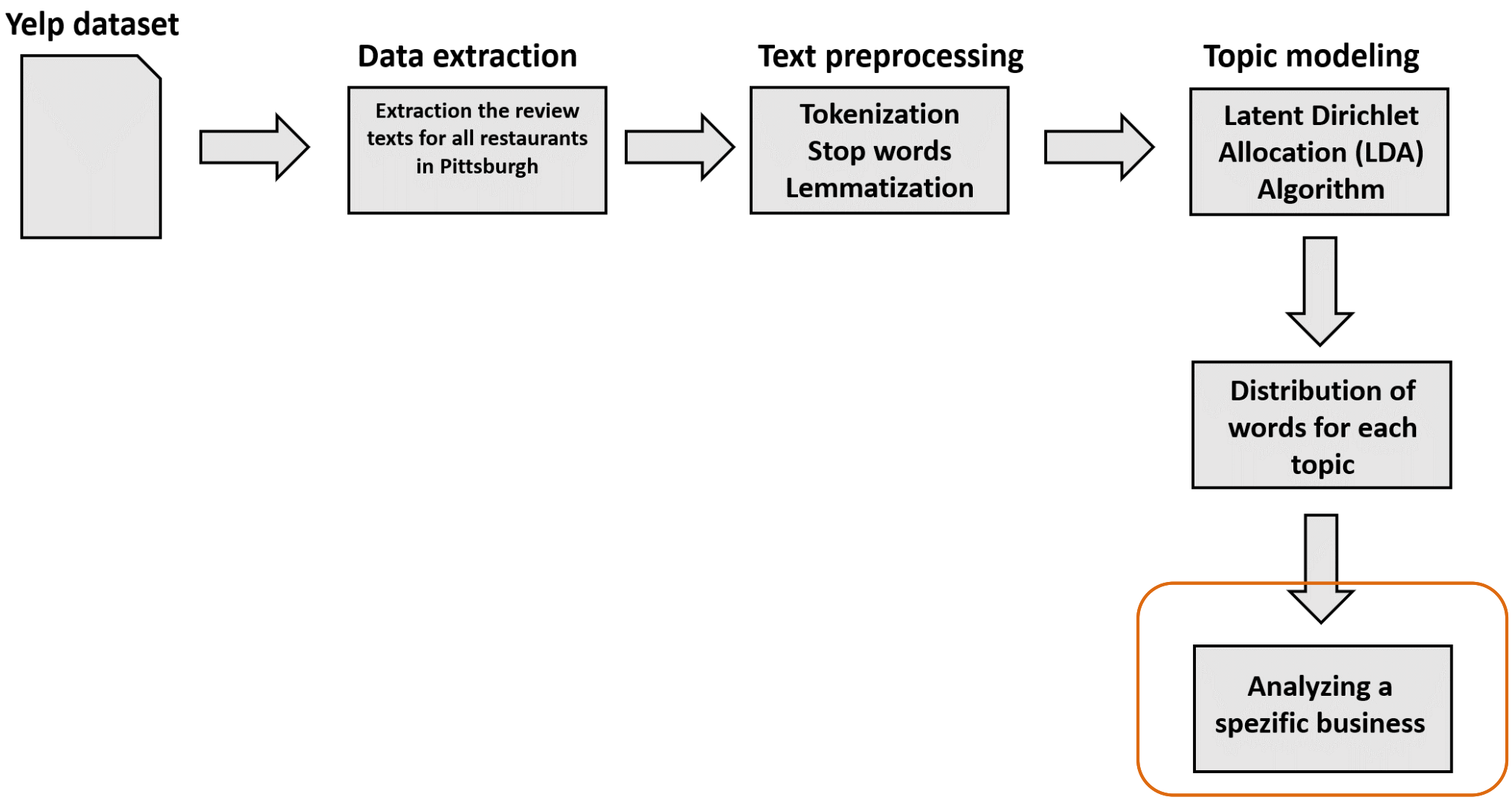
Top words of Topic 3



Top words of Topic 4



4. Workflow

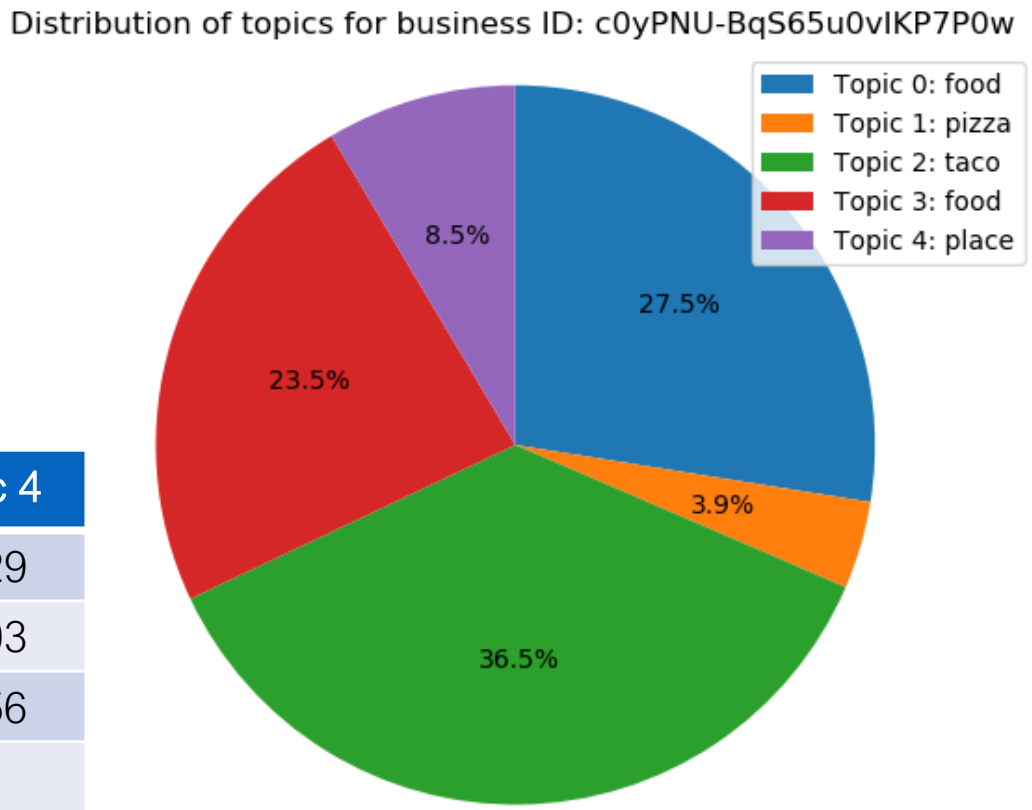

Yelp dataset
Data extraction
Extraction the review texts for all restaurants in Pittsburgh
Text preprocessing
Tokenization
Stop words
Lemmatization
Topic modeling
Latent Dirichlet Allocation (LDA) Algorithm
Distribution of words for each topic
Analyzing a spezific business

4. Results

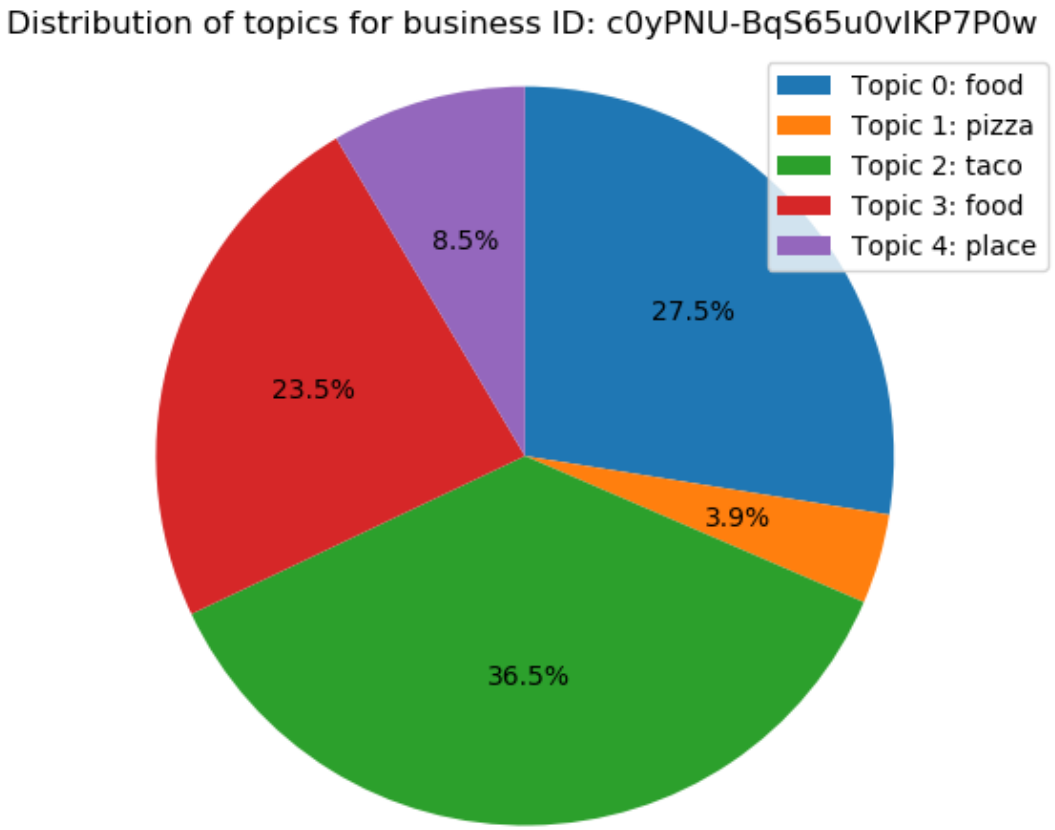
- Business ID: 'c0yPNU-BqS65u0vIKP7P0w‘
  - name: Avenue B
  - city: Pittsburgh
  - stars: 4.0
  - review_count: 228
  - categories: American (New), Restaurants

- Prediction:

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
Review 1	0.884	0.029	0.029	0.029	0.029
Review 2	0.216	0.121	0.658	0.003	0.003
Review 3	0.0029	0.299	0.317	0.325	0.056
…	…	…	…	…	…



4. Results



Topic 0: ['food', 'place', 'service', 'try', 'restaurant', 'price', 'chicken', 'time', 'come', 'like']

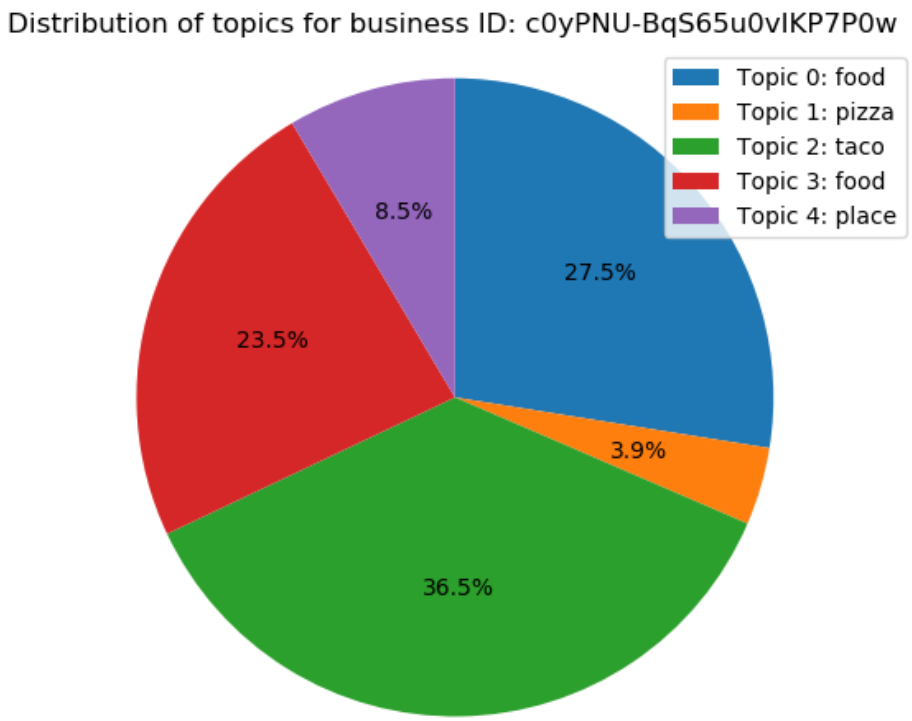
Topic 1: ['pizza', 'burger', 'order', 'fry', 'cheese', 'like', 'sauce', 'wing', 'eat', 'place']

Topic 2: ['taco', 'dish', 'delicious', 'order', 'flavor', 'sauce', 'thai', 'restaurant', 'meal', 'dessert']

Topic 3: ['food', 'time', 'order', 'place', 'come', 'service', 'table', 'wait', 'drink', 'bar']

Topic 4: ['place', 'sandwich', 'beer', 'pittsburgh', 'coffee', 'love', 'like', 'breakfast', 'brunch', 'selection']

4. Results



Topic 0: ['food', 'place', 'service', 'try', 'restaurant', 'price', 'chicken', 'time', 'come', 'like']

- **Menu?**

Topic 1: ['pizza', 'burger', 'order', 'fry', 'cheese', 'like', 'sauce', 'wing', 'eat', 'place']

- **Fast food**

Topic 2: ['taco', 'dish', 'delicious', 'order', 'flavor', 'sauce', ‘’thai', 'restaurant', 'meal', 'dessert‘]

- **Mexican and Thai food**

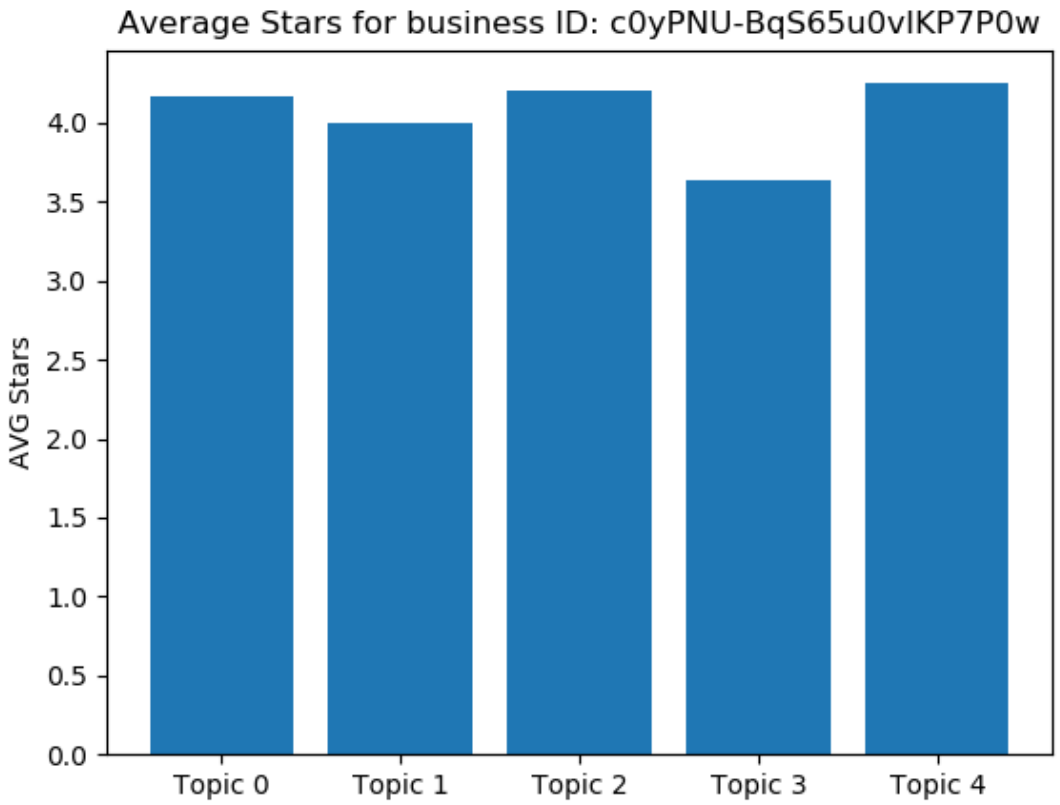
Topic 3: ['food', 'time', 'order', 'place', 'come', 'service', 'table', 'wait', 'drink', 'bar']

- **Restaurant / Diner?**

Topic 4: ['place', 'sandwich', 'beer', 'pittsburgh', 'coffee', 'love', 'like', 'breakfast', 'brunch', 'selection‘]

- **Café**

4. Results

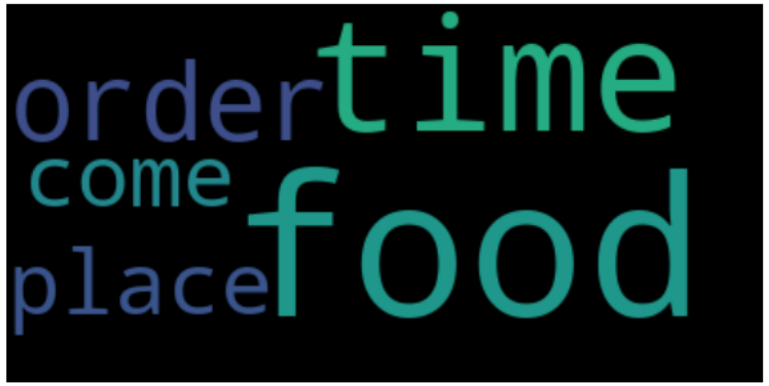


→Less average stars for Topic 3 ('food', 'time', 'order', 'place', 'come', 'service', 'table', 'wait', 'drink', 'bar') Restaurant / Diner

5. Discussion

Top words of Topic 0

Top words of Topic 3



- Working pipeline implemented
- Topic 0 and 3 are pretty similar
  - 5 topics for LDA are not optimal
  - Default values for α and β
  - Only used 1-grams

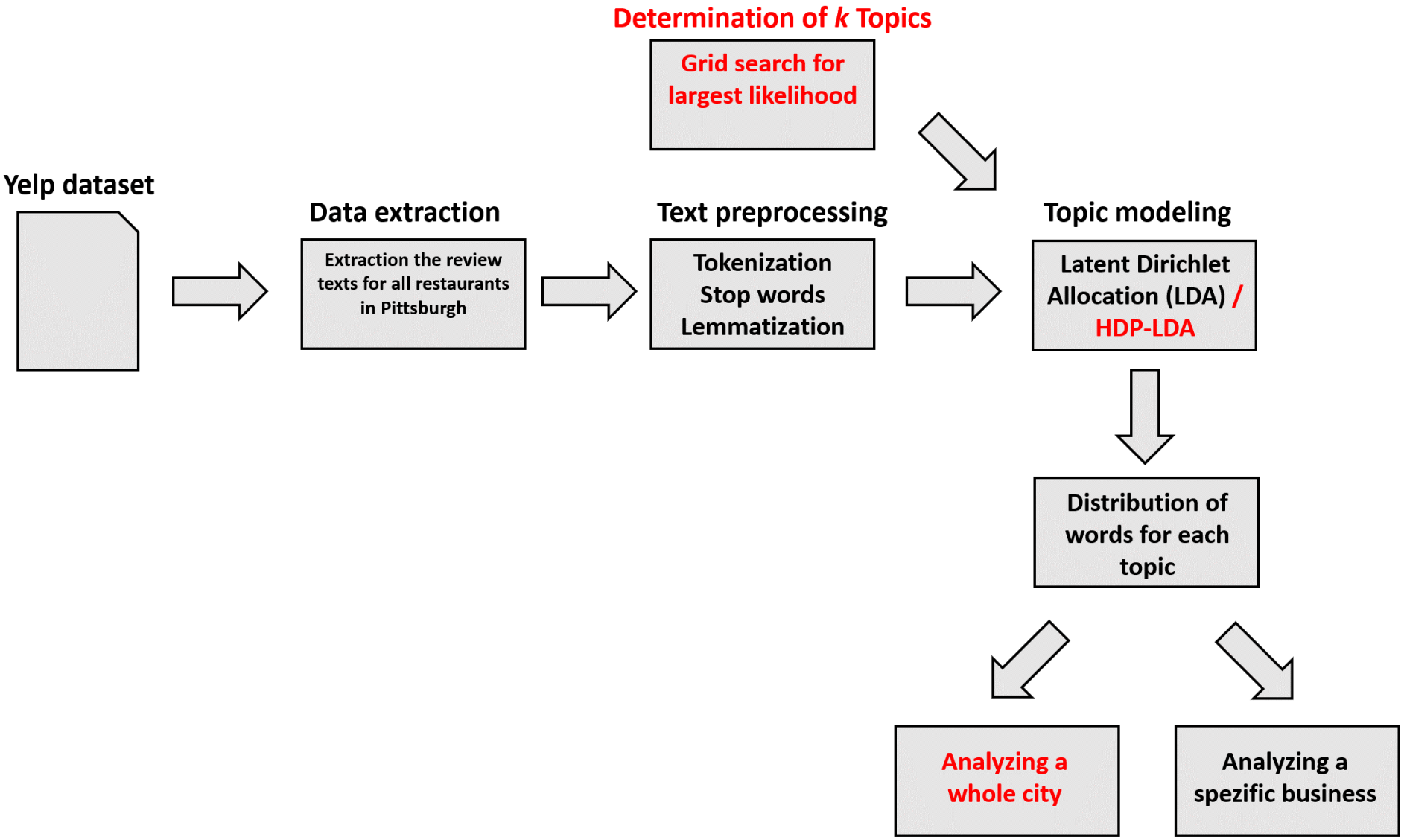
6. Prospects

- Determine α and β
  - Grid search

- Determine the number of topics (k)
  - Grid search
  - HDP-LDA

- Expanding code to answer the question: What do the citizen in Pittsburgh want? What is important to consider when opening a new business?
  - Analysis of all businesses in the area

6. Prospects


Determination of *k* Topics
Grid search for largest likelihood
Yelp dataset
Data extraction
Extraction the review texts for all restaurants in Pittsburgh
Text preprocessing
Tokenization
Stop words
Lemmatization
Topic modeling
Latent Dirichlet Allocation (LDA) / HDP-LDA
Distribution of words for each topic
Analyzing a whole city
Analyzing a spezific business

6. Prospects

- Determine α and β
  - Grid search
- Determine the number of topics (k)
  - Grid search
  - HDP-LDA
- Expanding code to answer the question: What do the citizen in Pittsburgh want? What is important to consider when opening a new business?
  - Analysis of all businesses in the area
- Evaluation of the results
  - Score: approximate log-likelihood
- Using a database for storage

Thank you for your attention