

Improving businesses with topic modeling on the Yelp dataset


Elvira Pupka-Lipinski

Content


1. Yelp
2. Possible tasks
3. Approach
4. Results
5. Discussion
6. Prospects

1. Yelp

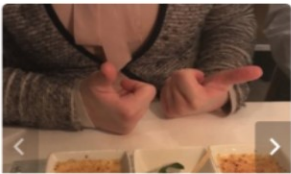
- Crowd-sourced review forum
- Offers a recommendation portal for restaurants and shops
- Short form of **Yellow Pages**



1. cooperativa Shiki
★★★★★ 2 reviews
Japanese, International
0531 48112610
Wendenstr. 4
"Shiki is an authentic Japanese restaurant that just opened in Braunschweig on October 20, 2018. The owner is Japanese with about half the staff also being..." [read more](#)




2. Corvin's Burger & Beer
★★★★★ 13 reviews
€€ - Burgers, Curry Sausage
0531 38986788
Neue Str. 9
"My first time having a burger from north of of the Rhein so I wasn't sure what to expect. This is a small city however it doesn't seem to have a focus on craft..." [read more](#)



3. Zucker
★★★★★ 19 reviews
€€€ - German
0531 281980
Frankfurter Str. 2

Mo' Map ☐ Redo search when map is moved



Map data ©2019 GeoBasis-DE/BKG (©2009) Terms of Use

1. Yelp

- Yelp: crowd-sourced review forum
- Offers a recommendation portal for restaurants and shops
- Short form of **Yellow Pages**

Corvin's Burger & Beer in Braunschweig

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

The screenshot displays the Yelp profile for Corvin's Burger & Beer. At the top, it shows a 4.5-star rating from 13 reviews, the address 'Neue Str. 9, 38100 Braunschweig, Germany', and contact information. A map on the left shows the location in the 'INNENSTADT' area. The main section features a 'Recommended Reviews' for Corvin's Burger & Beer, with a search bar and filters. Below this, a review by Klaus K. from Brunswick, Germany, dated 5/2/2017, is shown. The review is in German and describes the restaurant as a favorite for its burgers, atmosphere, and service. The review includes a 'Useful' button and a 'Funny' button. On the right side, there is a 'Hours' section showing the restaurant is open from 11:00 am to 10:00 pm on most days, and closed on Sunday. Below the hours is a 'More business info' section with details about reservations, delivery, take-out, accepted cards, good for lunch/dinner, good for kids, good for groups, ambiance (hipster), noise level (average), alcohol (full bar), outdoor seating (yes), and Wi-Fi (no).

★ ★ ★ ★ ★ 13 reviews [Details](#)

€€ · Burgers, Curry Sausage [Edit](#)

Neue Str. 9
38100 Braunschweig
Germany
[Get Directions](#)
+49 531 38986788
[corvins-burger-braunschweig.de](#)
[Send to your Phone](#)

Recommended Reviews for Corvin's Burger & Beer

Search within the reviews [Q](#) Sort by [Yelp Sort](#) Language [German \(12\)](#)

Recommended reviews in German [Translate German to English](#)

Klaus K.
Brunswick, Germany
3 friends
107 reviews
34 photos

★ ★ ★ ★ ★ 5/2/2017
5 check-ins

Dieses Restaurant ist mein absoluter Favorit, wenn es um Burger geht. Das Mittagsmenue ist vom Preis-Leistung und Qualitätverhältnis unschlagbar. Das Ambiente ist Rustikaler Natur aber dennoch sehr gemütlich. Die Bedienung ist angenehm freundlich und wirklich hilfsbereit. Je nachdem für welchen Burger man sich entscheidet, sind auch verschiedene Brötchen dabei, die auch von der Konsistenz überzeugen. Positiv fällt auch auf, das die Zutaten alle frisch, und auch frisch zubereitet, sind. Für Fans von guten Burgern ist das Restaurant ein absolutes Highlight.

Dirk K. voted for this review

[Useful 1](#) [Funny](#) [Cool](#)

Today 11:00 am - 10:00 pm
[Open now](#)

[Full menu](#)

Price range €11-20

[Work here? Claim this business](#)

Hours

Mon	3:00 pm - 10:00 pm
Tue	3:00 pm - 10:00 pm
Wed	3:00 pm - 10:00 pm
Thu	3:00 pm - 10:00 pm
Fri	3:00 pm - 10:00 pm
Sat	11:00 am - 10:00 pm Open now
Sun	Closed

[Edit business info](#)

More business info

Takes Reservations [Yes](#)
Delivery [No](#)
Take-out [Yes](#)
Accepted Cards [Debit](#)
Good For [Lunch, Dinner](#)
Good for Kids [Yes](#)
Good for Groups [Yes](#)
Ambience [Hipster](#)
Noise Level [Average](#)
Alcohol [Full Bar](#)
Outdoor Seating [Yes](#)
Wi-Fi [No](#)

1. Yelp

- Yelp: crowd-sourced review forum
- Offers a recommendation portal for restaurants and shops
- Short form of **Yellow Pages**

Corvin's Burger & Beer in Braunschweig

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

People also viewed



Duke Burger

★★★★☆ 25 reviews

€€ • Burgers
Mitte



Culinario

★★★★☆ 6 reviews

€€ • Burgers



Duke

★★★★☆ 2 reviews

Burgers
Oststadt



Black Button

Other Places Nearby

[Find more Burgers near Corvin's Burger & Beer](#)

[Find more Curry Sausage near Corvin's Burger & Beer](#)

Browse Nearby

 Restaurants

 Nightlife

 Shopping

... [Show all](#)

1. Yelp

- Yelp: crowd-sourced review forum
- Offers a recommendation portal for restaurants and shops
- Short form of **Yellow Pages**
- 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

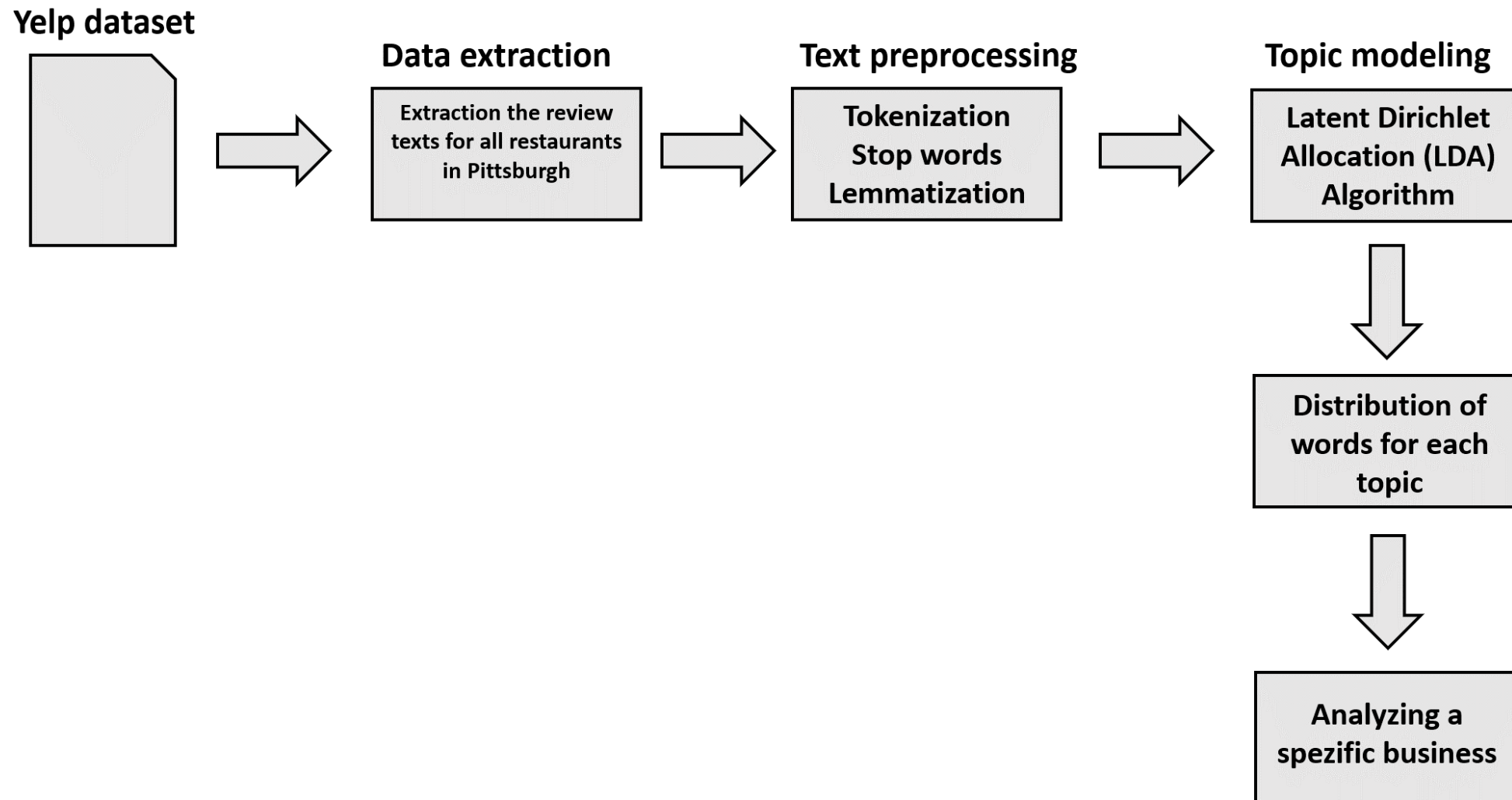
3. Business improvements

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

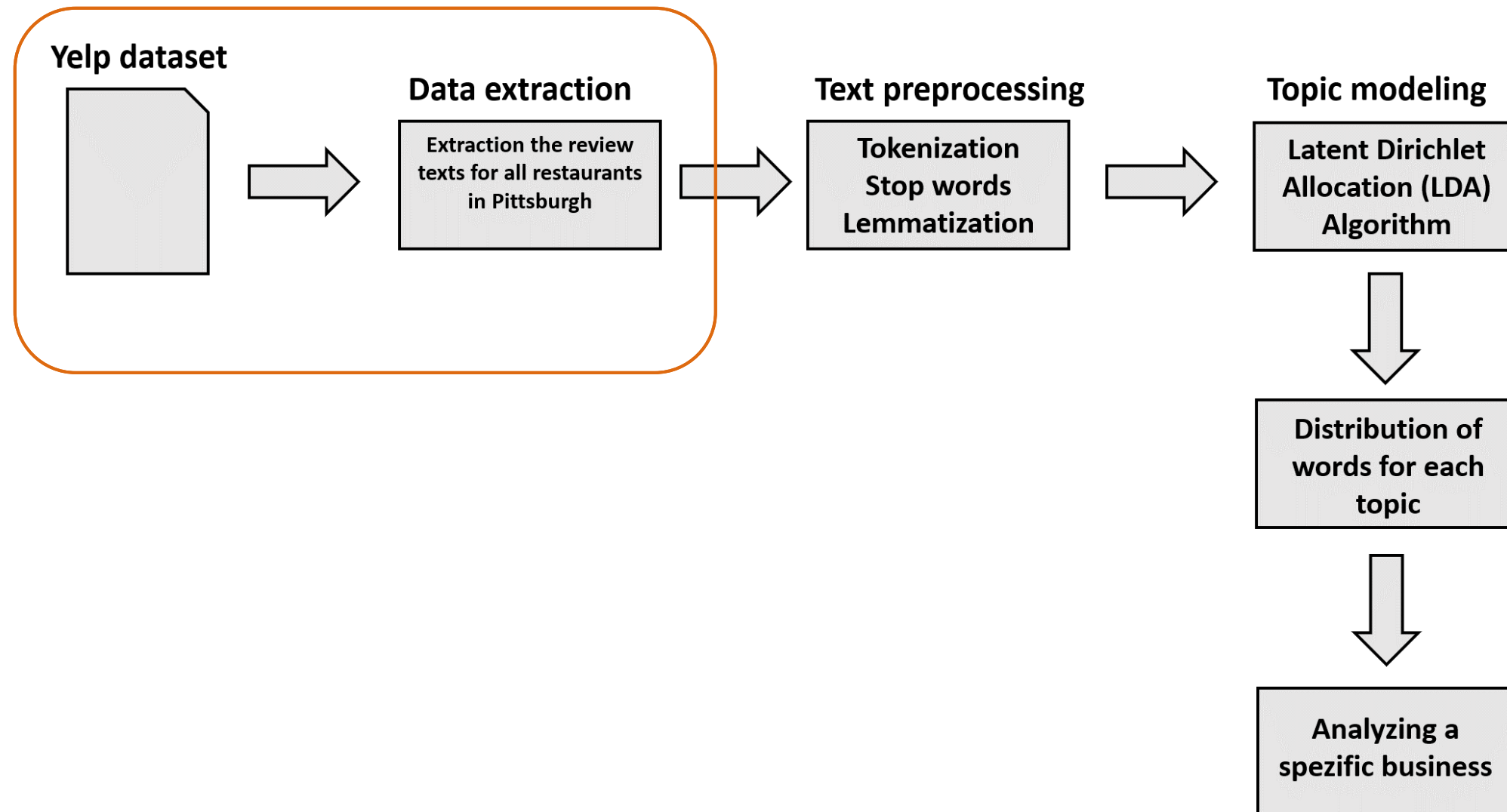
Topic	Avg Stars
Food	4.0
Service	2.0
Location	4.0

→ Improvement in service is needed

3. Workflow



3. Workflow



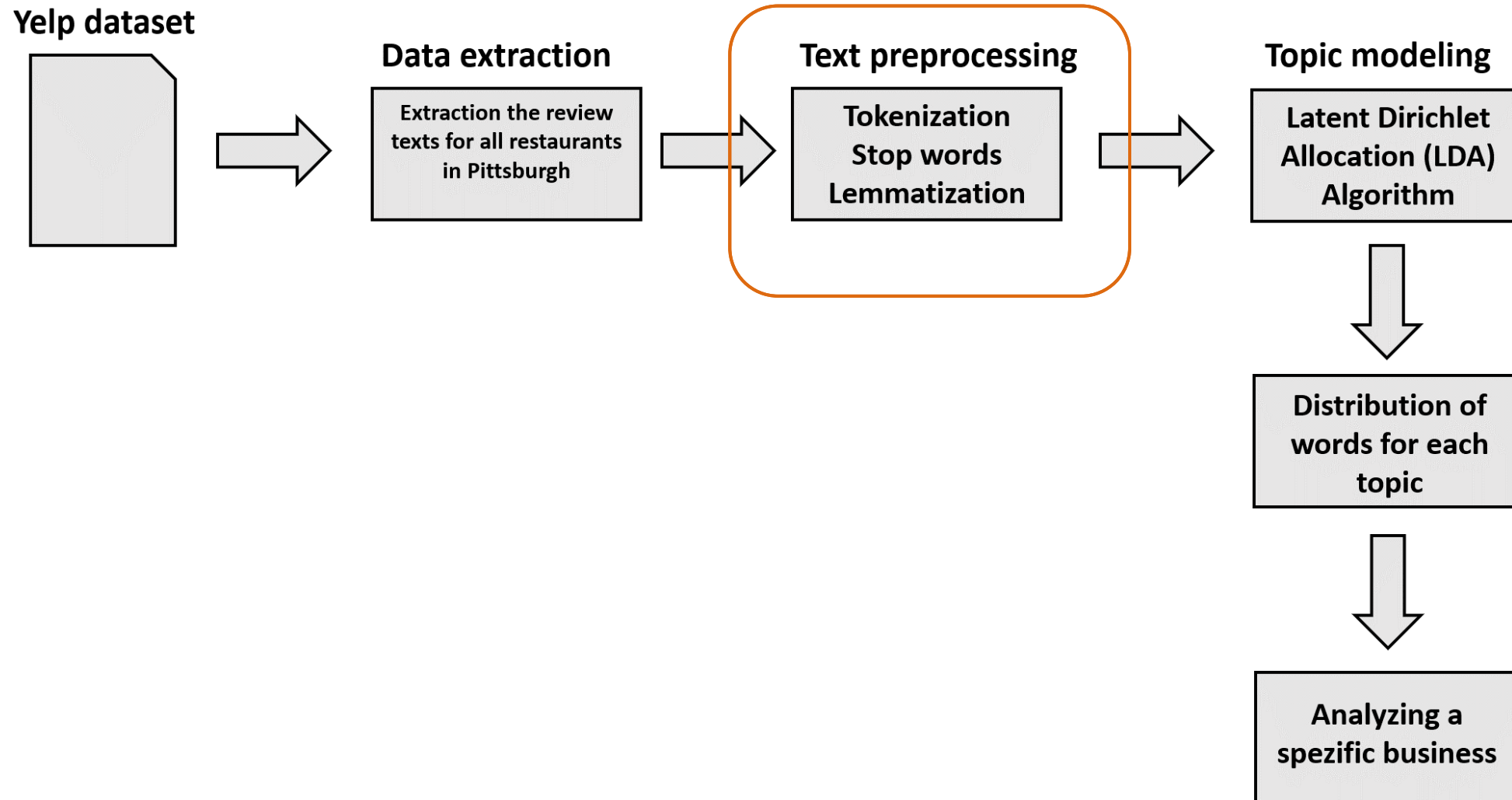
3. Yelp dataset – Data Extraction

- 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



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

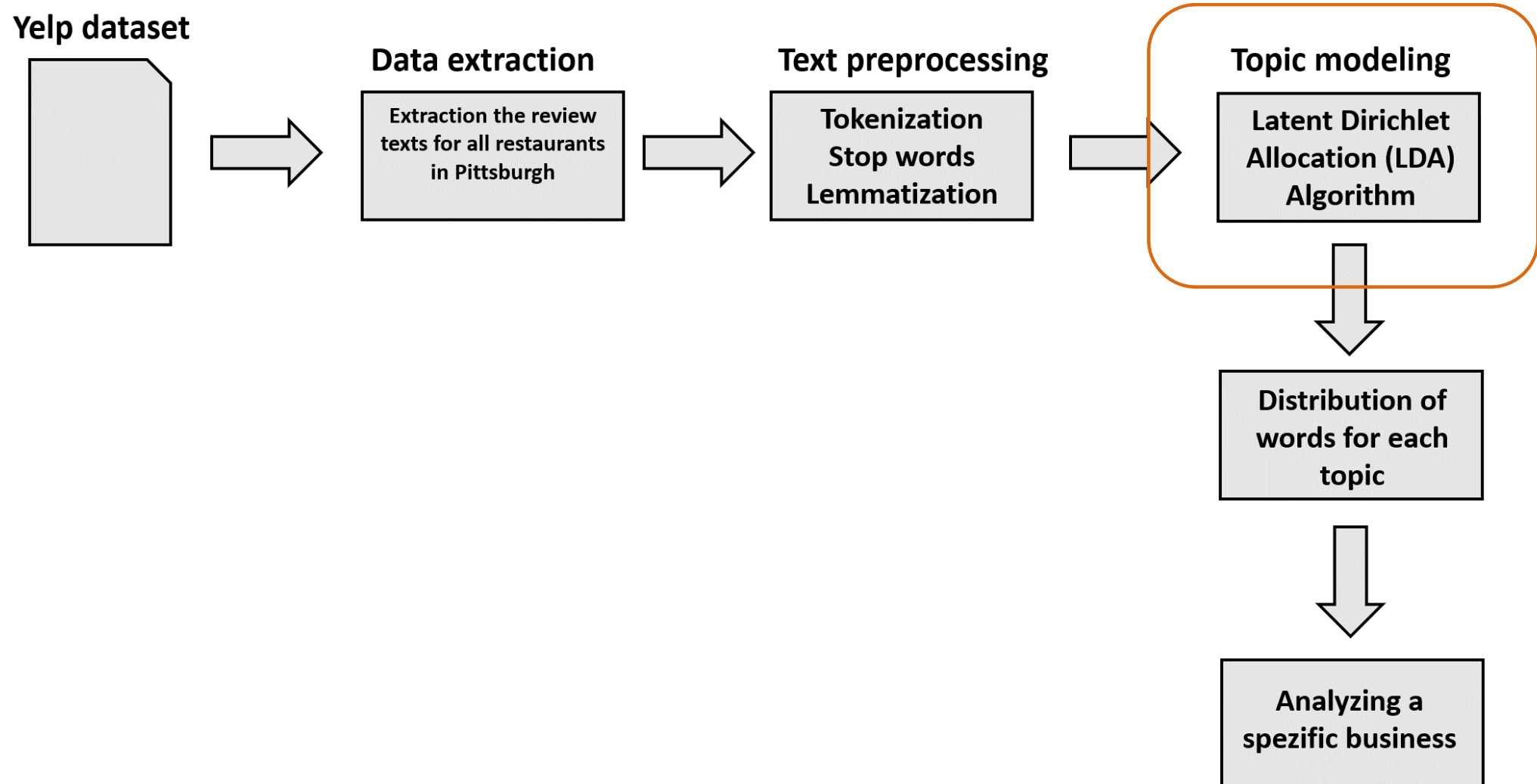
- **Stop words:**

→ ['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

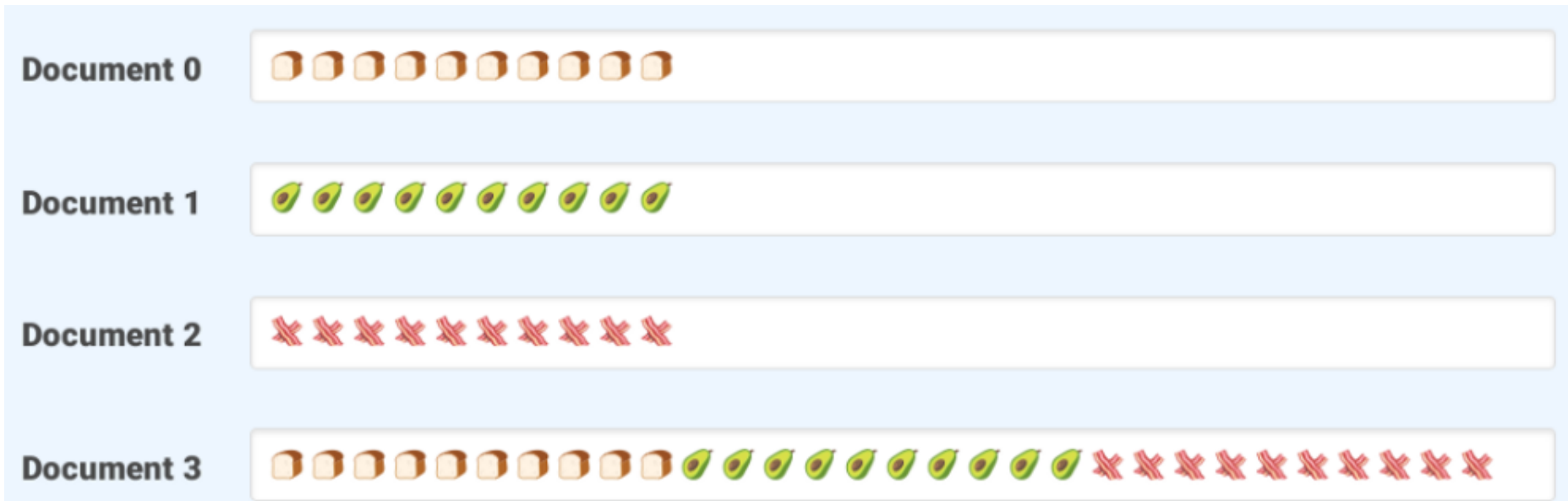
→ ['definitely', 'enjoy', 'meal', 'avenue', 'b', 'meal', 'come', 'hefty', 'price', 'n't', 'know', 'worth', 'trip', 'pay', 'food', 'mind', 'blow']

3. Workflow



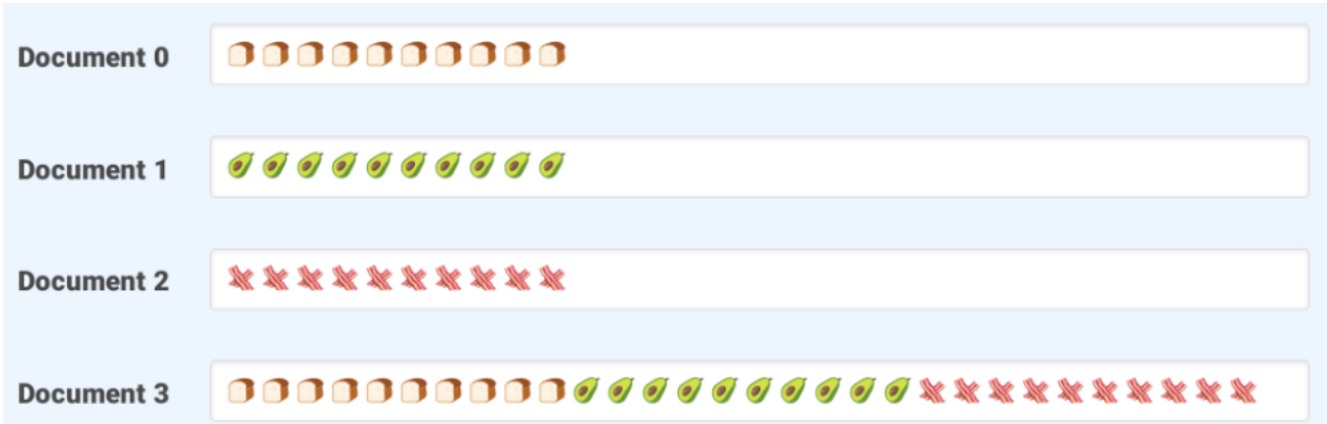
3. Approach: Latent Dirichlet Allocation

- 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 Assumption 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 Assumption 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	

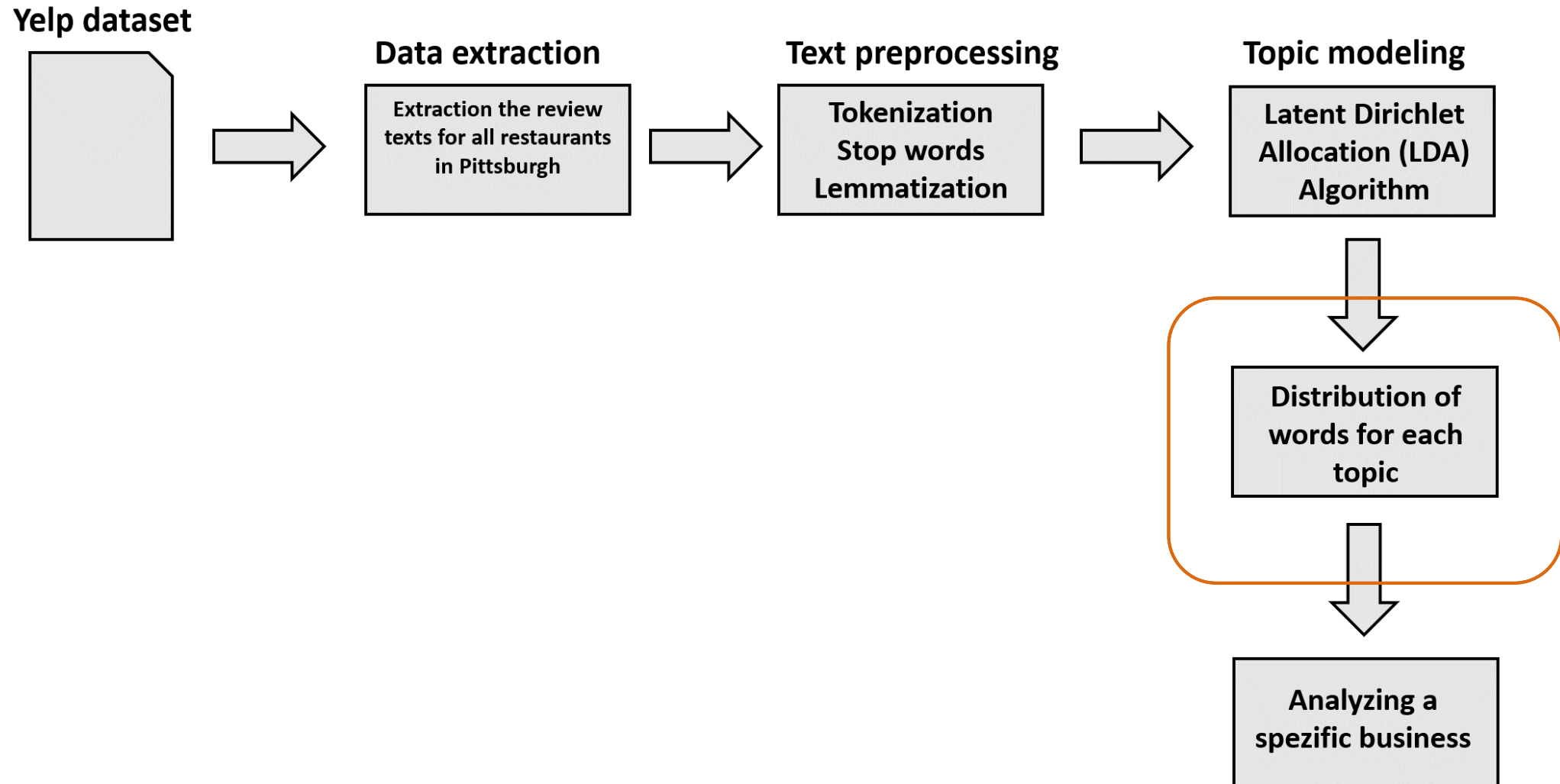
	Topic 0	Topic 1	Topic 2
Document 0	0.030	0.030	0.939
Document 1	0.939	0.030	0.030
Document 2	0.030	0.939	0.030
Document 3	0.333	0.333	0.333
Topic 0	Topic 1	Topic 2	

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

3. How does LDA works for topic modeling?

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

3. Workflow



4. Results

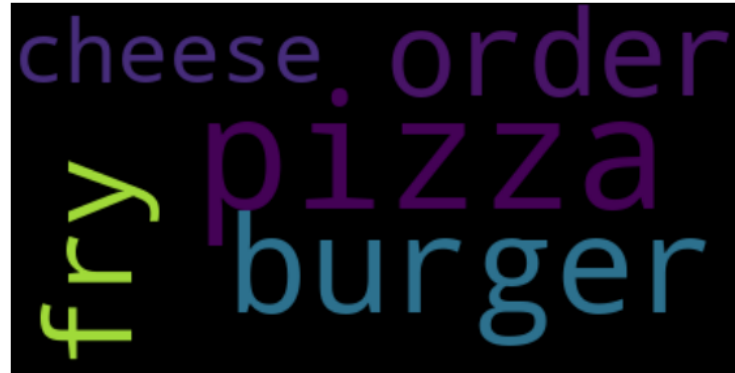
- Top 10 words in topics for Pittsburgh and Restaurants:
 - Topic 0: ['food', 'place', 'service', 'try', 'restaurant', 'price', 'chicken', 'time', 'come', 'like']
 - Topic 1: ['pizza', 'burger', 'order', 'fry', 'cheese', 'like', 'sauce', 'wing', 'eat', 'place']
 - Topic 2: ['taco', 'dish', 'delicious', '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

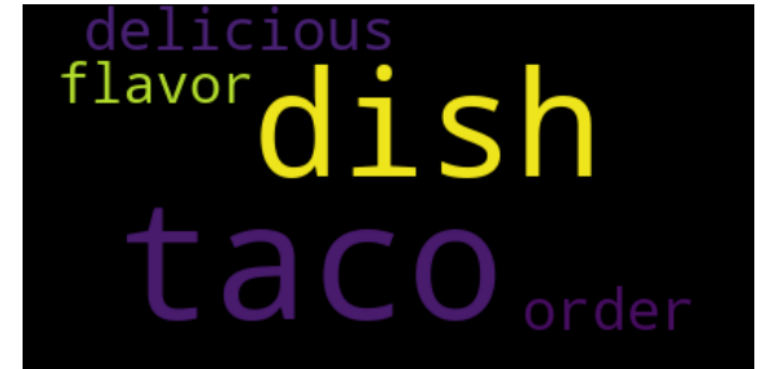
Top words of Topic 0



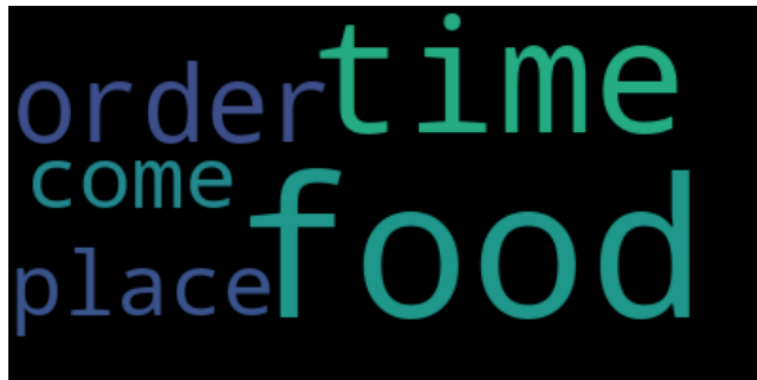
Top words of Topic 1



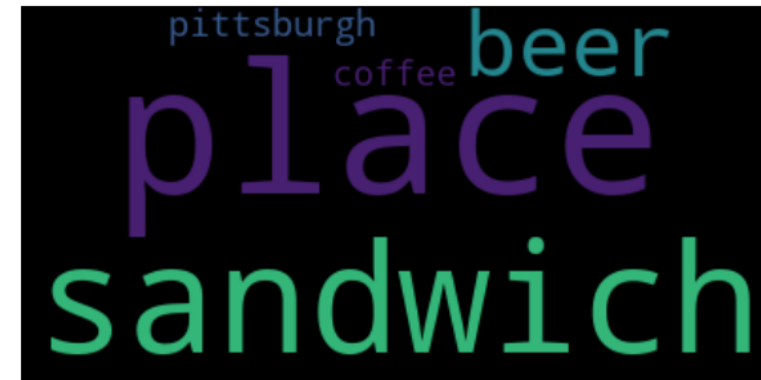
Top words of Topic 2



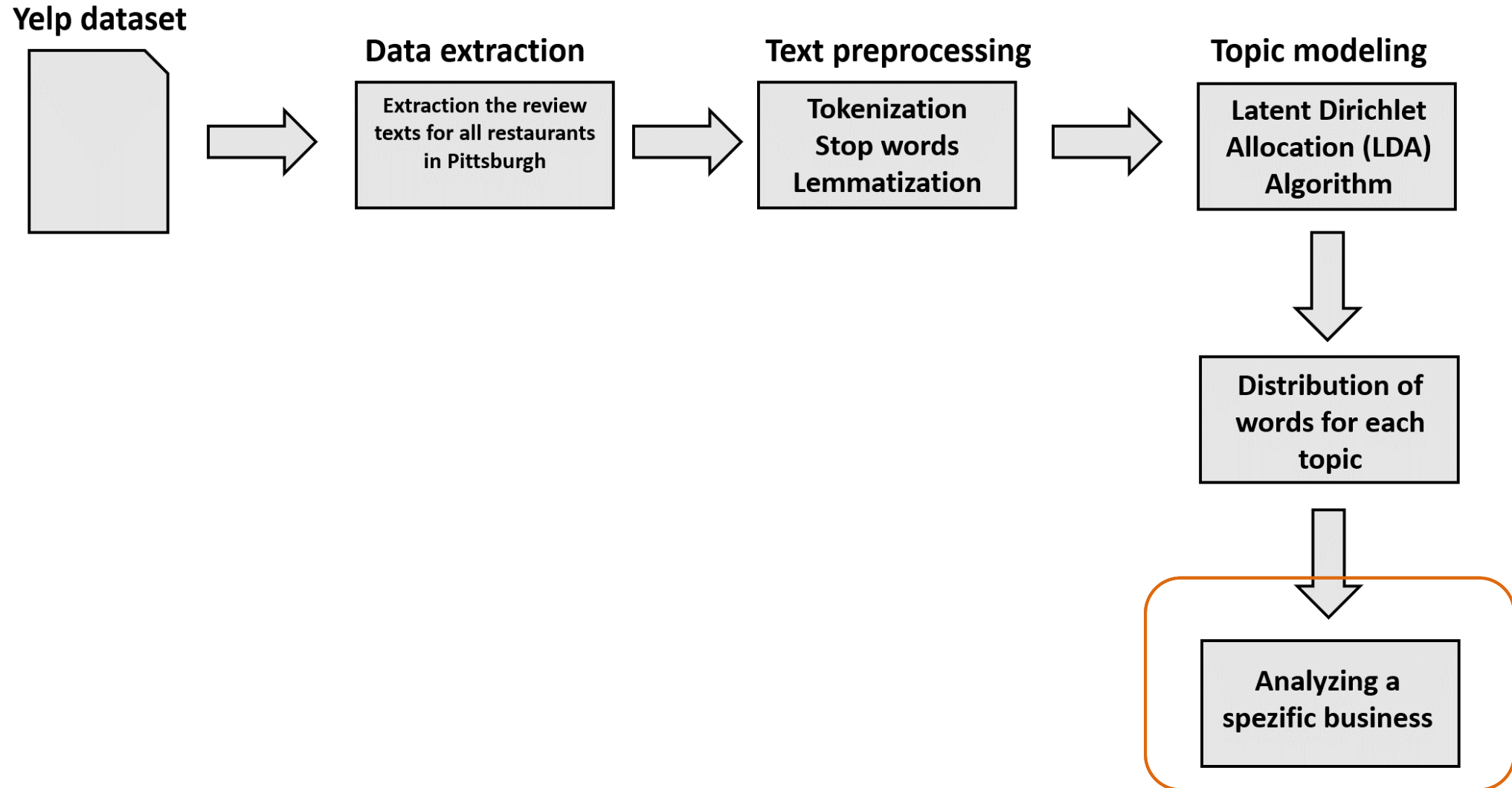
Top words of Topic 3



Top words of Topic 4



4. Workflow



4. Results

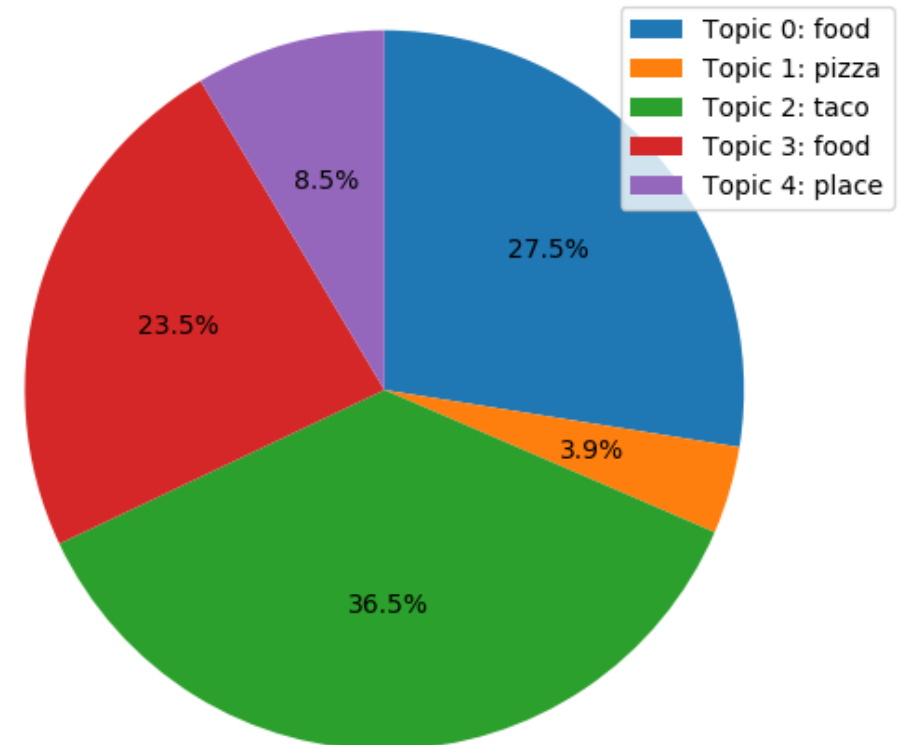
- Business ID: 'c0yPNU-BqS65u0vIKP7P0w'

- name: Avenue B
- city: Pittsburgh
- stars: 4.0
- review_count: 228
- categories: American (New), Restaurants

- Prediction:

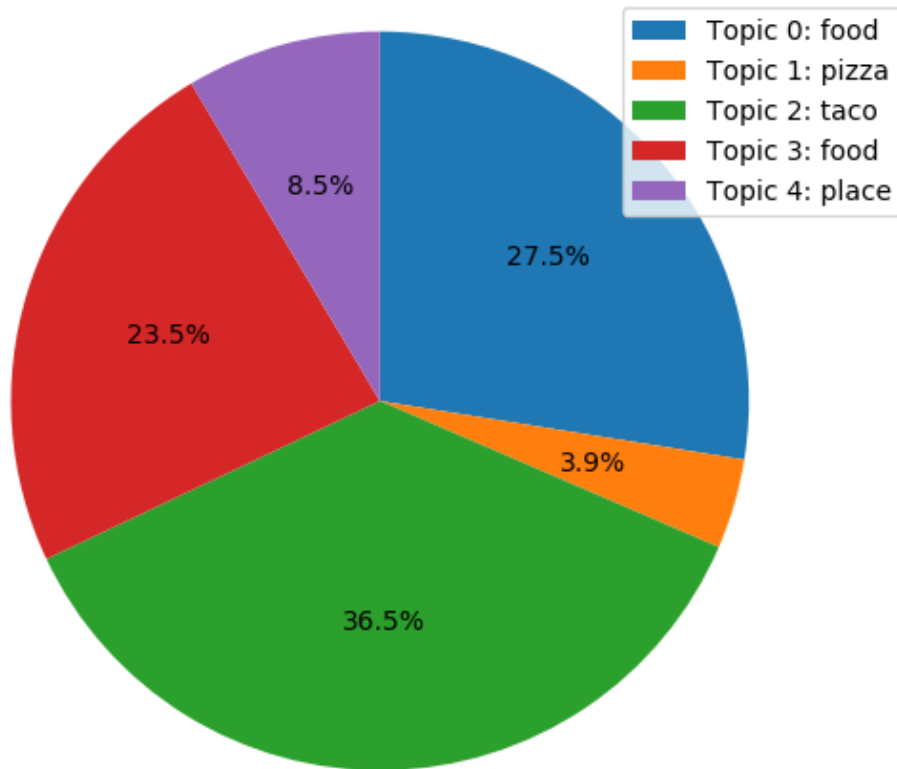
	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
Review 1	0.884	0.029	0.029	0.029	0.029
Review 2	0.216	0.121	0.658	0.003	0.003
Review 3	0.0029	0.299	0.317	0.325	0.056
...

Distribution of topics for business ID: c0yPNU-BqS65u0vIKP7P0w



4. Results

Distribution of topics for business ID: c0yPNU-BqS65u0vIKP7P0w



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

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

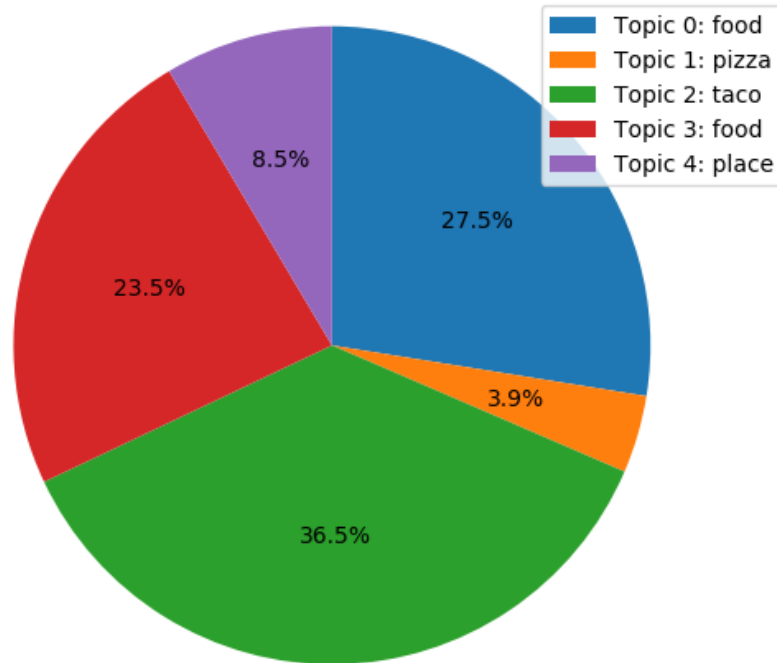
Topic 2: ['taco', 'dish', 'delicious', 'order', 'flavor', 'sauce', 'thai', 'restaurant', 'meal', 'dessert']

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

Distribution of topics for business ID: c0yPNU-BqS65u0vIKP7P0w



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**

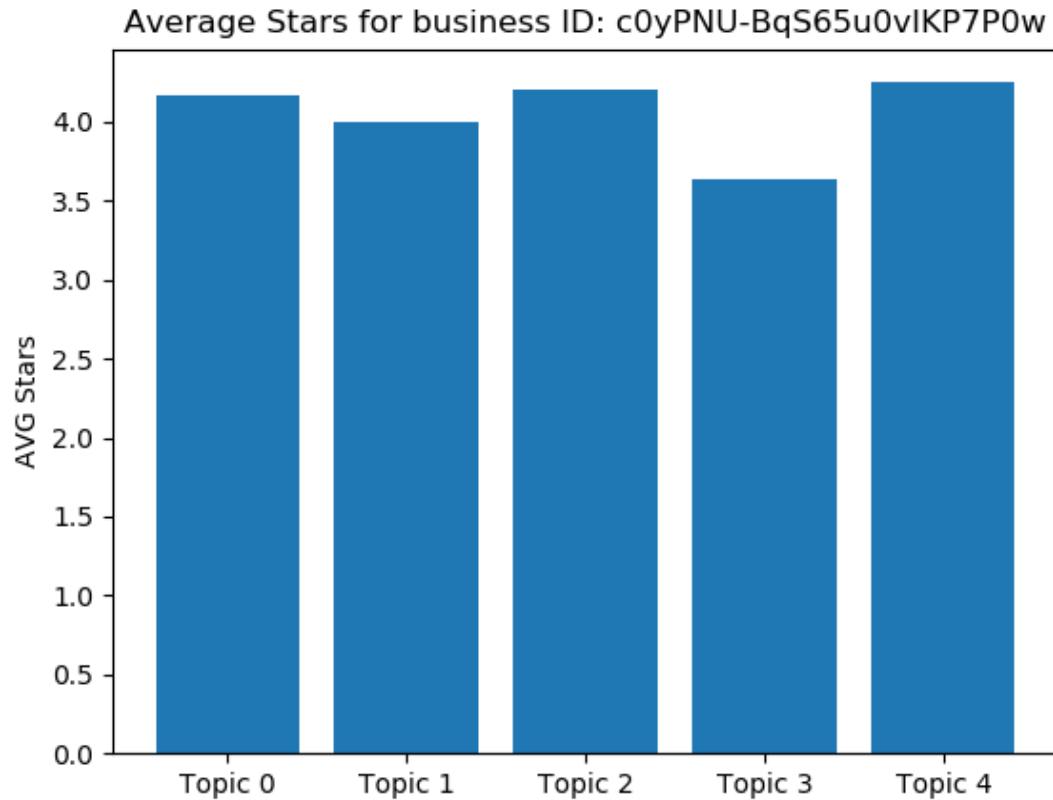
Topic 3: ['food', 'time', 'order', 'place', 'come', 'service', 'table', 'wait', 'drink', 'bar']

- **Restaurant / Diner?**

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

- **Café**

4. Results



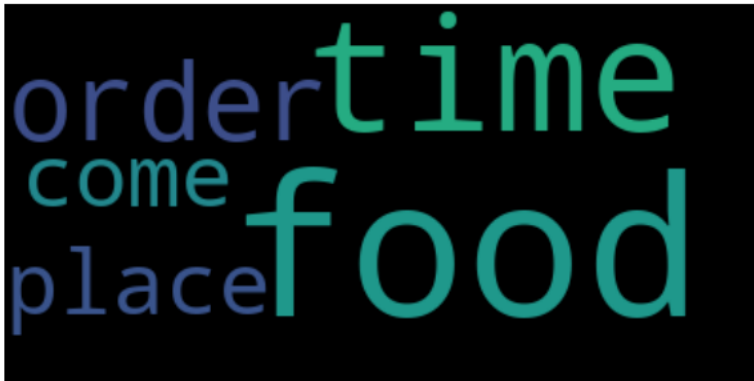
→ Less average stars for Topic 3 ('food', 'time', 'order', 'place', 'come', 'service', 'table', 'wait', 'drink', 'bar') Restaurant / Diner

5. Discussion

Top words of Topic 0



Top words of Topic 3

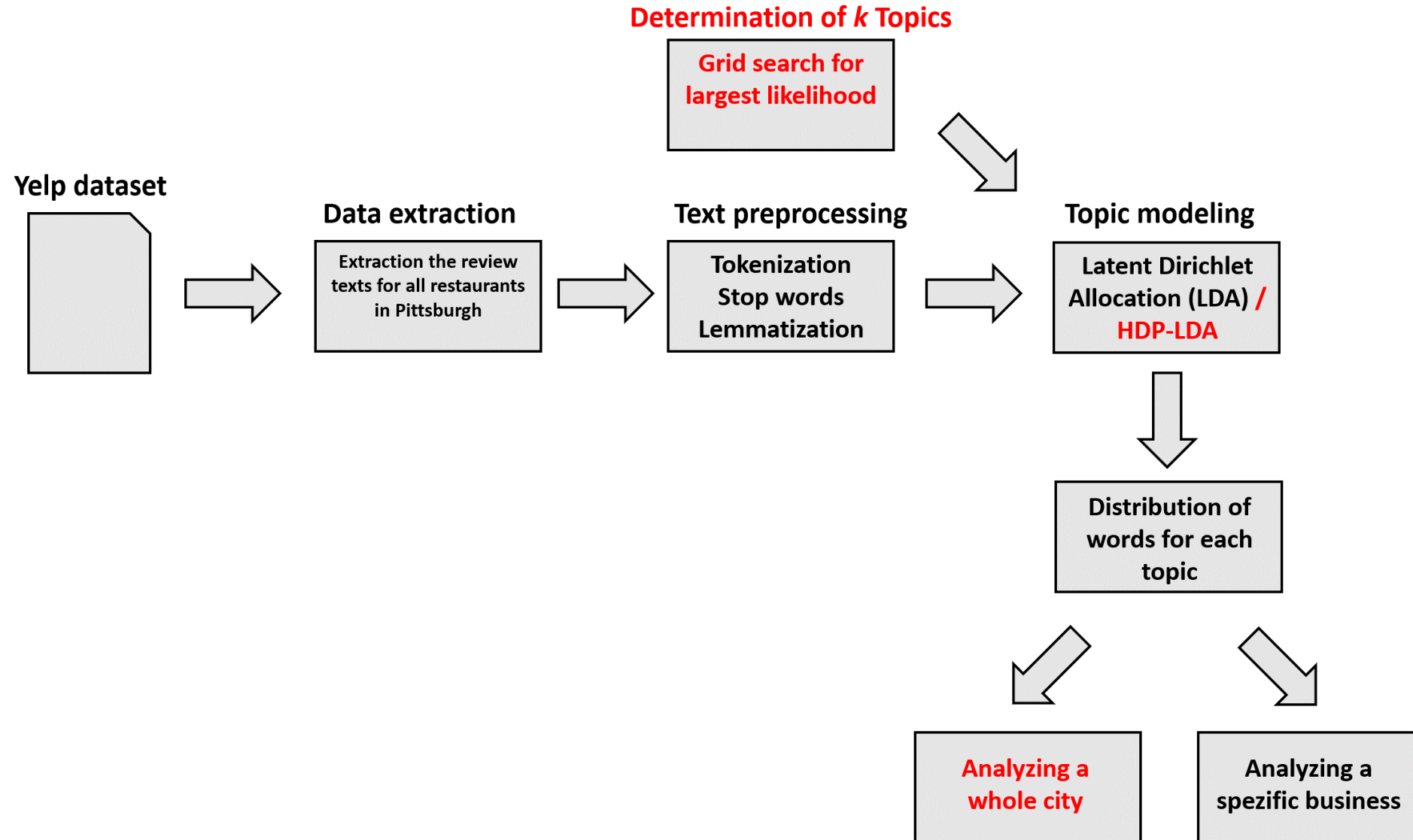


- Working pipeline implemented
- Topic 0 and 3 are pretty similar
 - 5 topics for LDA are not optimal
 - Default values for α and β
 - Only used 1-grams

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



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