Improving businesses with Topic Modelling

Yelp Dataset

Elvira Pupka-Lipinski

Content

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

1. Insight into Yelp Dataset

- Yelp: crowd-sourced review forum

- 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)

1. 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?

1. 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?**

1. Business improvements

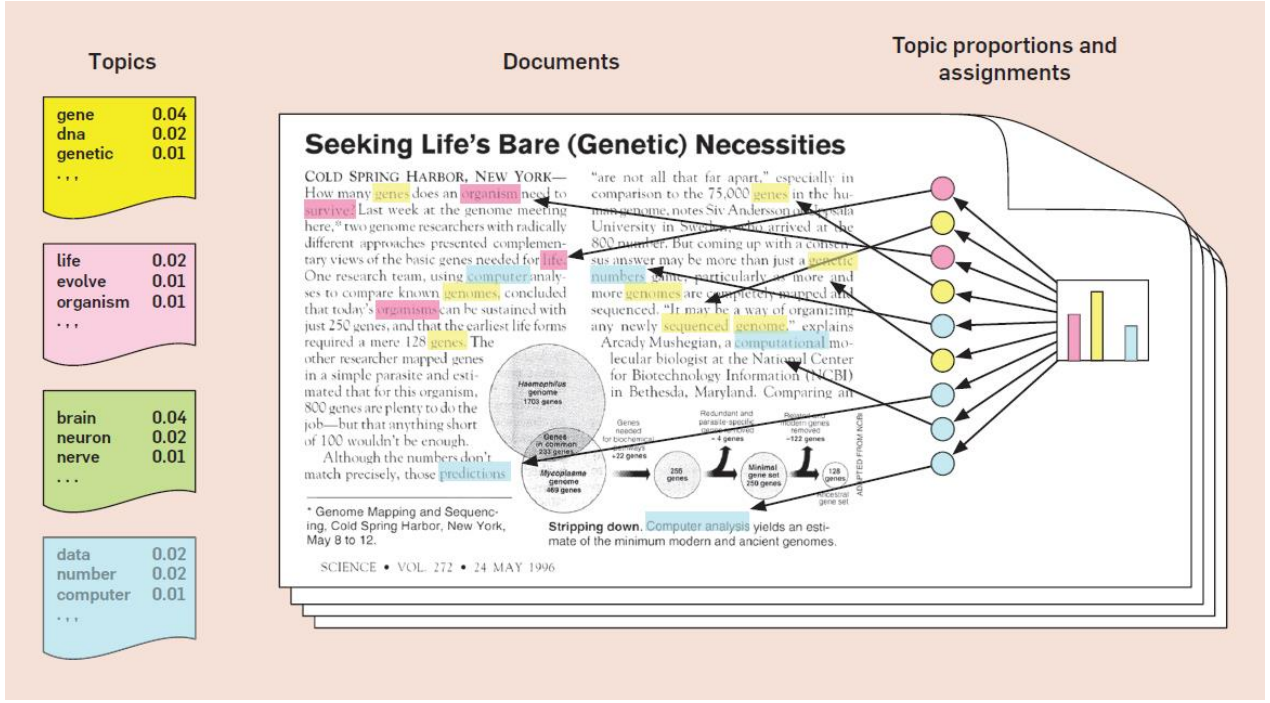
- Improvements for business
  - **What can I as a business owner improve?**
    - **Idea: Extracting the topics the owner needs to improve**
    - **Approach: Topic Modelling**

Topic	Avg Stars
Food	4.0
Service	2.0
Location	4.0

→ Improvement in Service is needed

Approach:Latent Dirichlet Allocation

- Unsupervised learning
- Variables to determine number of topics



https://medium.com/@connectwithghosh/topic-modelling-with-latent-dirichlet-allocation-lda-in-pyspark-2cb3ebd5678e

- Alternatives: NMF, Autoencoder

Results

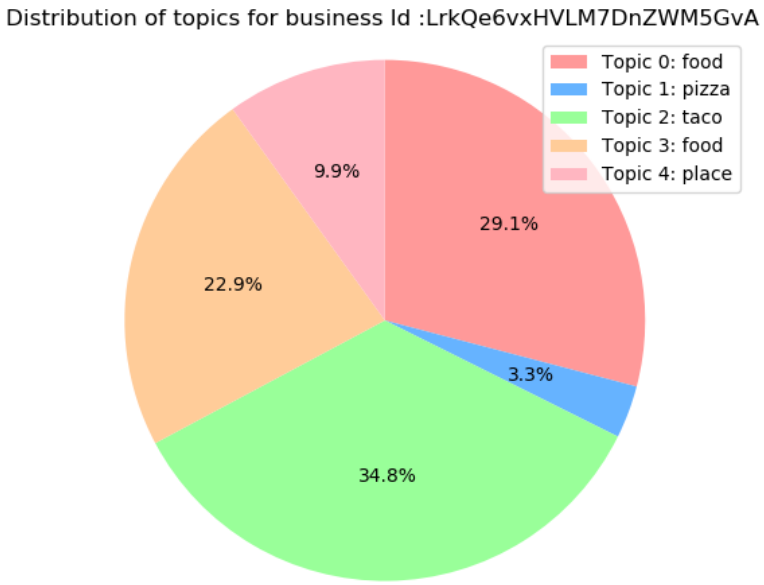
- Top words in topics

  - 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‘]

Results

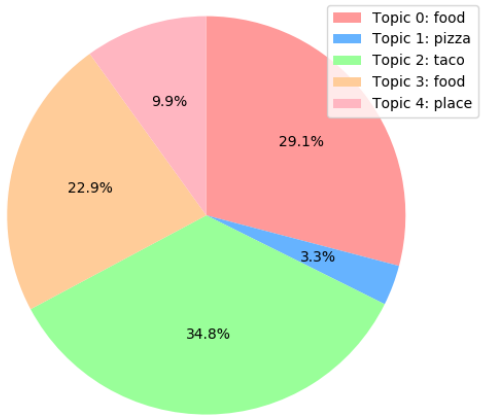
- Business ID: 'c0yPNU-BqS65u0vIKP7P0w', ,
  - name: Avenue B
  - city: Pittsburgh,
  - stars: 4.0
  - review_count: 228
  - categories: American (New), Restaurants
- Prediction:

	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
…	…	…	…	…	…



Results

Distribution of topics for business Id :LrkQe6vxHVLM7DnZWM5GvA



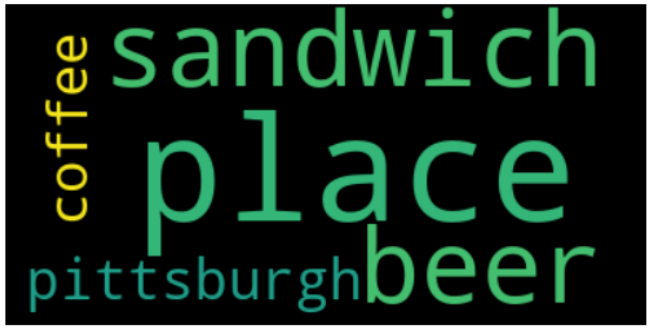
Top words of Topic 0 for businss ID: LrkQe6vxHVLM7DnZWM5GvA

Top words of Topic 1 for businss ID: LrkQe6vxHVLM7DnZWM5GvA

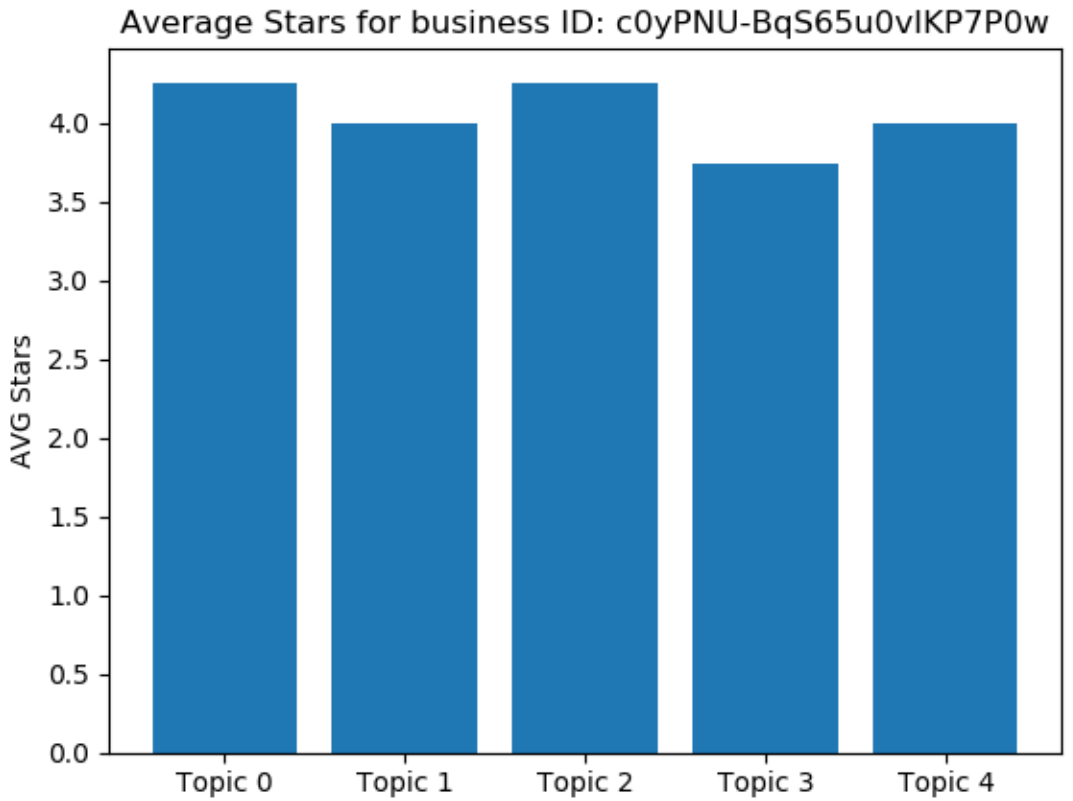
Top words of Topic 2 for businss ID: LrkQe6vxHVLM7DnZWM5GvA

Top words of Topic 3 for businss ID: LrkQe6vxHVLM7DnZWM5GvA

Top words of Topic 4 for businss ID: LrkQe6vxHVLM7DnZWM5GvA



Results


Average Stars for business ID: c0yPNU-BqS65u0vIKP7P0w
AVG Stars
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Topic 0
Topic 1
Topic 2
Topic 3
Topic 4

Prospects

- Determine the number of topics (k) for LDA with an Algorithm
- Re-training LDA with more data
- Changing 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