Rohini Vishwanathan
IST 332
Natural Language Processing

Lab Assignment #5

Google colab- https://colab.research.google.com/drive/1HwQRQPNR6sV8evxF9J9k6LC-gaoUv-lO?usp=sharing

1. **Prepare data:**
   Figure 1. Final cleaned review tokens without digits or punctuations.

```
Total reviews with digits or punctuations: 0
```

	Cleaned_Tokens
36144	stop saturday slow servic guy took order even ...
7823	review tasti food cool look bad servic recomme...
33222	roll roll littl small price still delici would...
43666	live year final tri spot love chicken littl sp...
62394	place must tri think tri everyth food offer li...

2. Dictionary and Corpus creation:
Figure 2. The printed dictionary size (len(id2word))

Step 2B: Build bigram model & dictionary

```
# Train a bigram model (appear >=60 times)
bigram = Phrases(texts, min_count=60)
bigram_phraser = Phraser(bigram)

# Apply bigram model to reviews
texts_bigrams = bigram_phraser[texts]

# Create Gensim dictionary using unigrams + bigrams
id2word = corpora.Dictionary(texts_bigrams)

# Remove very rare (<60 docs) and very frequent (>50% docs) terms
id2word.filter_extremes(no_below=60, no_above=0.5)

# Display dictionary size
print("Dictionary size (number of terms):", len(id2word))
print("Example bigrammed review:\n", texts_bigrams[0][:20])
```

```
Dictionary size (number of terms): 514
Example bigrammed review:
 ['best', 'pizza', 'found', 'alway', 'got', 'great', 'deal', 'offer', 'staff', 'friend', 'make', 'feel', 'welcom', 'make', 'feel', 'extra'
```

Figure 3. The BOW corpus preview or summary

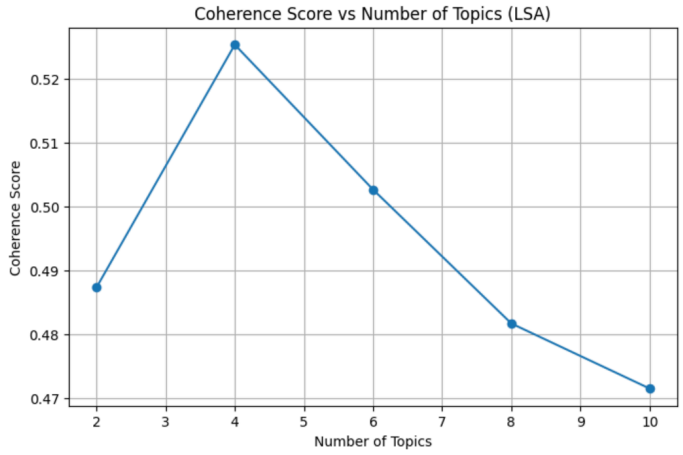
Step 2C: Create corpus (term_id, frequency)

```
corpus = [id2word.doc2bow(text) for text in texts_bigrams]

print("Example BoW for first review:\n", corpus[0][:10])
print("Total number of documents in corpus:", len(corpus))
```

```
Example BoW for first review:
 [(0, 1), (1, 1), (2, 1), (3, 1), (4, 2), (5, 1), (6, 1), (7, 1), (8, 1), (9, 1)]
Total number of documents in corpus: 63158
```

Step 3: LSA Model (BOW+ Bigrams)
Figure 4. Coherence Score vs. Number of Topics for LSA (BOW).

```
Training LSA model with 2 topics ...
Coherence Score for 2 topics: 0.4874

Training LSA model with 4 topics ...
Coherence Score for 4 topics: 0.5254

Training LSA model with 6 topics ...
Coherence Score for 6 topics: 0.5027

Training LSA model with 8 topics ...
Coherence Score for 8 topics: 0.4817

Training LSA model with 10 topics ...
Coherence Score for 10 topics: 0.4715
```



The best LSA model using BOW achieved a coherence of 0.5254 with 4 topics. Topics were labeled as Food Quality, Service, Customer Experience, and Pricing.
The 4-topic LSA model (BOW + bigrams) achieved the highest coherence score (0.5254).

The topics extracted align well with core restaurant service dimensions:

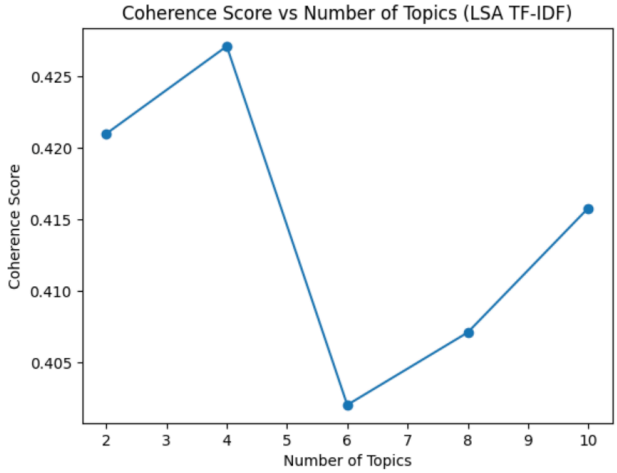
- Topic 0 – Food Quality & Taste: Focuses on food quality, ingredients, and dining satisfaction.
- Topic 1 – Service Speed & Order Accuracy: Highlights waiting time, order accuracy, and service pace.
- Topic 2 – Customer Experience & Staff Behavior: Captures staff friendliness, service attitude, and overall experience.
- Topic 3 – Pricing & Restaurant Atmosphere: Reflects menu variety, pricing, and perception of value.

All topics were interpretable and could be labeled because the dataset was domain-specific and yielded semantically coherent clusters. In more diverse corpora, some topics might remain unlabeled if dominated by mixed or ambiguous keywords.

This model provides interpretable insights into restaurant reviews and reveals what aspects customers emphasize most when discussing their dining experiences. Although some overlap between “order” and “food” occurs due to the linear nature of LSA, each topic still represents a distinct dimension of restaurant service quality.

Step 4: LSA model (TF-IDF + Bigrams)
Figure 5. Coherence Score vs. Number of Topics for LSA (TF-IDF).

```
Training LSA (TF-IDF) model with 2 topics ...
Coherence Score for 2 topics: 0.421
Training LSA (TF-IDF) model with 4 topics ...
Coherence Score for 4 topics: 0.4271
Training LSA (TF-IDF) model with 6 topics ...
Coherence Score for 6 topics: 0.402
Training LSA (TF-IDF) model with 8 topics ...
Coherence Score for 8 topics: 0.4071
Training LSA (TF-IDF) model with 10 topics ...
Coherence Score for 10 topics: 0.4158
```



Comparison and Interpretation: BOW vs TF-IDF LSA Models

Model Performance Comparison
Both LSA models (using Bag-of-Words and TF-IDF) identified 4 coherent topics, reflecting consistent semantic structure within the restaurant review corpus. The BOW-based LSA achieved a slightly higher coherence score (0.5254) compared to the TF-IDF-based LSA (0.4296). This indicates that the BOW model captured clearer topic boundaries when raw term frequencies were used. However, the TF-IDF model still produced interpretable and meaningful topics related to restaurant service dimensions.

Interpretation of Topics
All topics were interpretable and could be labeled because the dataset was domain-specific and yielded semantically coherent clusters. In more diverse corpora, some topics might remain unlabeled if dominated by mixed or ambiguous keywords.
The BOW-based LSA emphasized broader themes such as food quality, order accuracy, staff behavior, and pricing, while the TF-IDF model surfaced more content-specific nuances—such as pizza preparation, waiting time, and menu variety. TF-IDF weighting helped highlight distinctive terms that appear less frequently but carry higher importance in describing restaurant experiences.

Conclusion
While both models provide valuable insights, the BOW-based LSA performs slightly better in coherence and general interpretability. The TF-IDF-based LSA adds complementary value by emphasizing rare but informative keywords, making it especially useful for finer-grained analysis of customer sentiment and operational quality.

Step 5: LDA Model (BOW + Bigrams)

Model Configuration:
Each model was trained using the specified parameters — chunksize = 1500, passes = 20, iterations = 400, gamma_threshold = 0.001, alpha = 'auto', eta = 'auto'.
Topic numbers were tested from 4 to 18 in increments of 2.

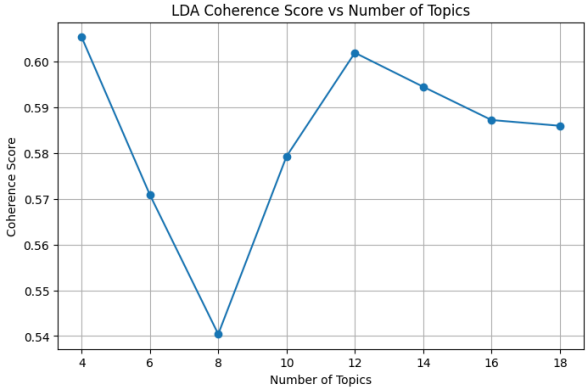
Coherence Score Results:

# Topics	4	6	8	10	12	14	16	18
Coherence	**0.6054**	0.5709	0.5404	0.5793	0.6019	0.5945	0.5872	0.5860

Best Model: 4 topics (Highest coherence = 0.6054)

Figure 6. Coherence Score vs. Number of Topics for LDA.

```
Training LDA model with 4 topics ...
Coherence Score for 4 topics: 0.6054

Training LDA model with 6 topics ...
Coherence Score for 6 topics: 0.5709

Training LDA model with 8 topics ...
Coherence Score for 8 topics: 0.5404

Training LDA model with 10 topics ...
Coherence Score for 10 topics: 0.5793

Training LDA model with 12 topics ...
Coherence Score for 12 topics: 0.6019

Training LDA model with 14 topics ...
Coherence Score for 14 topics: 0.5945

Training LDA model with 16 topics ...
Coherence Score for 16 topics: 0.5872

Training LDA model with 18 topics ...
Coherence Score for 18 topics: 0.586
```



Topic Model Output and Labeling

Topic #	Top Keywords (abridged)	Assigned Label
0	good, place, food, great, like, friend, staff	Overall Experience & Atmosphere

1	pizza, salad, cheese, sandwich, burger, chicken	Food Quality & Menu Variety
2	order, time, wait, get, would, one	Order Accuracy & Waiting Time
3	chicken, dish, sauce, rice, roll, sushi	Cuisine Specialties & Preparation

All topics were interpretable and could be labeled because the dataset is domain-specific; words cluster naturally around restaurant-related themes. In more diverse corpora, unlabeled or ambiguous topics might emerge.

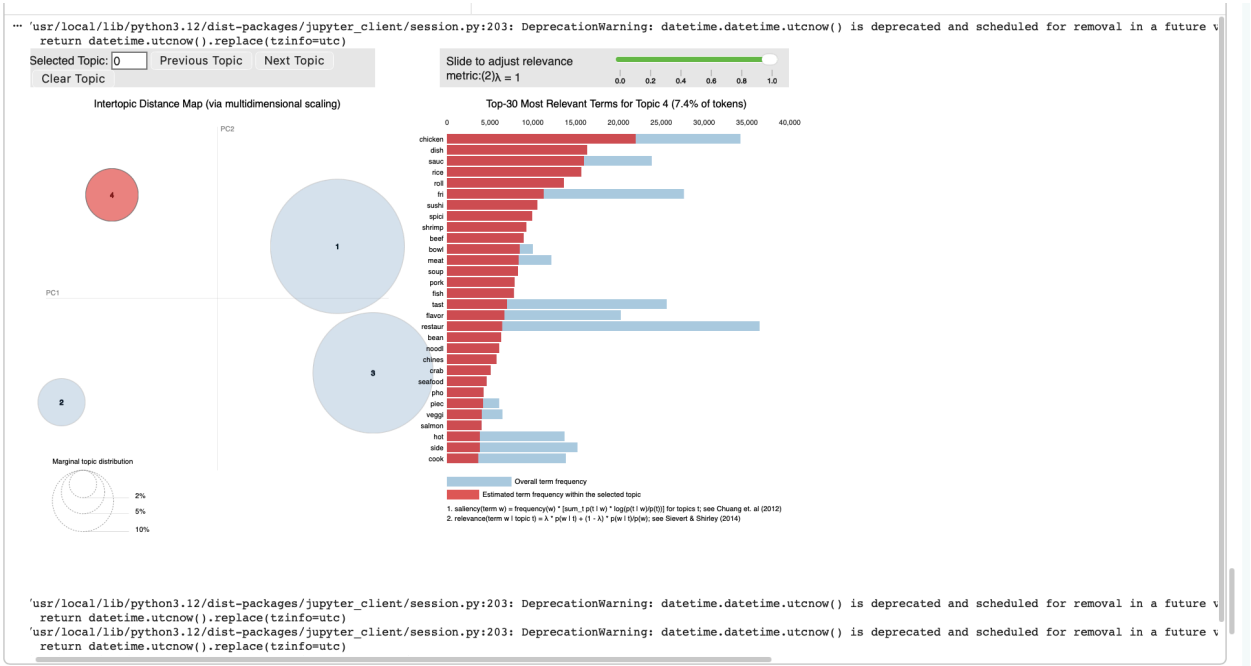
Interpretation and Discussion
The 4-topic LDA model produced the clearest, most coherent representation of restaurant review themes.

- Topic 0 – Overall Experience & Atmosphere: captures general satisfaction and staff friendliness.
- Topic 1 – Food Quality & Menu Variety: focuses on popular dishes and product diversity.
- Topic 2 – Order Accuracy & Waiting Time: represents service efficiency and timeliness.
- Topic 3 – Cuisine Specialties & Preparation: relates to specific cuisines and meal preparation.

These topics align with major restaurant service dimensions—food quality, service efficiency, staff experience, and ambiance—making the LDA model highly relevant for business insight.

Interactive Visualization
The interactive visualization generated using pyLDAvis.gensim_models displays each topic as a circle in a 2-D semantic space.
Minimal overlap among the circles indicates clear topic separation.
The right-side keyword bars rank terms by importance, confirming that food and service-related terms dominate the corpus.

Figure 7. Interactive pyLDAvis visualization for 4-topic LDA model



Step 6: Model Comparison and Selection

Model Overview
Three topic-modeling approaches were developed and evaluated on the restaurant review corpus:

Model	Feature Type	# Topics	Coherence Score	Interpretability	Key Focus
LSA (BOW)	Bag of Words + Bigrams	4	0.5254	Moderate	Food quality & service
LSA (TF-IDF)	TF-IDF weighted terms	4	0.4296	Low–moderate	Menu variety, delivery accuracy
LDA (BOW + Bigrams)	Probabilistic BOW	4	**0.6054 (best)**	High	Comprehensive restaurant themes (food, service, experience)

Model Selection
The LDA model (4 topics) is selected as the best overall model.
It achieves the highest coherence score (0.6054) and produces the most distinct and interpretable topic clusters, as confirmed by the interactive pyLDAvis visualization.
While the LSA (BOW) model performed reasonably well, its topics partially overlapped and tended to merge multiple aspects of reviews.
The LSA (TF-IDF) model emphasized frequent terms but reduced semantic diversity, leading to less coherent and more redundant topics.
In contrast, the LDA model not only achieved superior quantitative coherence but also demonstrated clearer semantic separation across topics such as food quality, service speed, and customer experience.

Why LDA is the Best Model

1. **Higher Coherence & Topic Purity** → LDA captured more semantically consistent word groupings, indicating stronger internal cohesion of themes.
2. **Probabilistic Interpretation** → Unlike LSA, which is purely algebraic, LDA models the probability distribution of words within topics and topics within documents, yielding richer contextual insights.
3. **Business Relevance** → The LDA topics align directly with restaurant operations (food, service, staff, ambiance), making the findings actionable for management analysis.
4. **Visualization Support** → The *pyLDAvis* interface confirmed minimal overlap and intuitive interpretability—an important factor for communicating results to non-technical stakeholders.

Business Insights from the Best Model
The final LDA model reveals four primary dimensions of restaurant performance based on customer reviews:

Topic #	Theme Label
0 – Food Quality & Taste	Mentions of *“good,” “place,” “food,” “great,” “flavor”* indicate satisfaction with food quality and taste consistency.

1 – Menu Variety & Ingredients	Frequent terms like *“pizza,” “cheese,” “salad,” “sandwich”* highlight focus on product diversity and ingredient freshness.
2 – Service Speed & Order Experience	Words such as *“order,” “time,” “wait,” “get,” “would”* capture customer perceptions of promptness and order accuracy.
3 – Dining Experience & Staff Behavior	References to *“chicken,” “rice,” “roll,” “sauce,” “friend,” “staff”* emphasize meal presentation, interaction, and overall atmosphere.

Collectively, these topics provide actionable guidance for restaurants to maintain menu quality, streamline order fulfillment, and enhance the in-store experience.

Conclusion
Among the three models, the LDA model (4 topics) stands out as the most coherent, interpretable, and business-relevant representation of the restaurant review corpus.
Its clear topic boundaries and rich interpretability make it the ideal choice for understanding customer sentiment and identifying key operational improvement areas within the restaurant industry.