Extracting Information from Textual Data Using Natural Language Processing

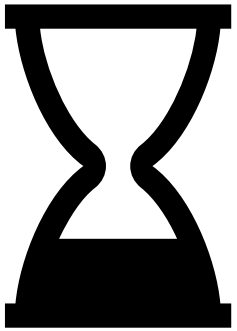
OUTLINE

1. Purpose/Business Question
2. Dataset(s)
3. Frequency-Based Methods
4. Topic Modeling
5. Question-Answering Task
6. Task Evaluation (Q&A)

BUSINESS PROBLEMS

Summarize several large text files.

Time-Consuming



Extraction of keywords, common themes, and needed information.

¡ Topic Modeling & Q&A Tasks !

THE DATASETS

Item 5: Initial Fees

- Initial license fee, franchise, and additional fees
- Required inventory prior to opening from approved distributor.
- Refunds of fees

Item 5_Answers

- **File Name –** The name of the text file corresponding to a franchise.
- **True Initial Fee –** The actual initial franchise fee for a specific franchise.
- **Alternative Answer –** Five (5) other accepted answers for the initial franchise fee but is not the actual fee.

NO MISSING DATA, ALL 50 ENTRIES

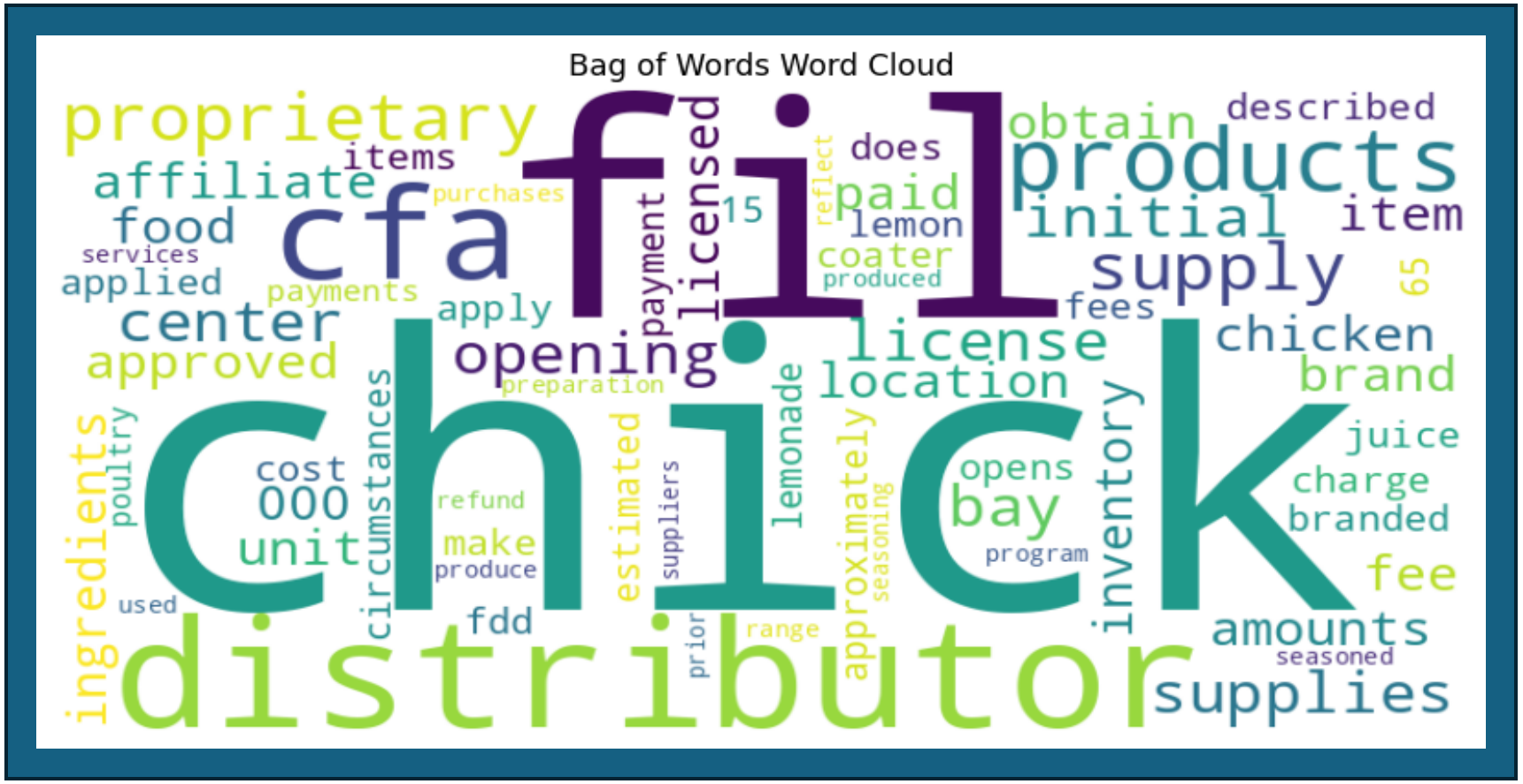
FREQUENCY-BASED METHODS

...for Item 5: Chick-fil-A

BAG OF WORDS (BoW)

Functions Used

```
CountVectorizer(stop_words='english')
bow_vectorizer.fit_transform(documents)
```

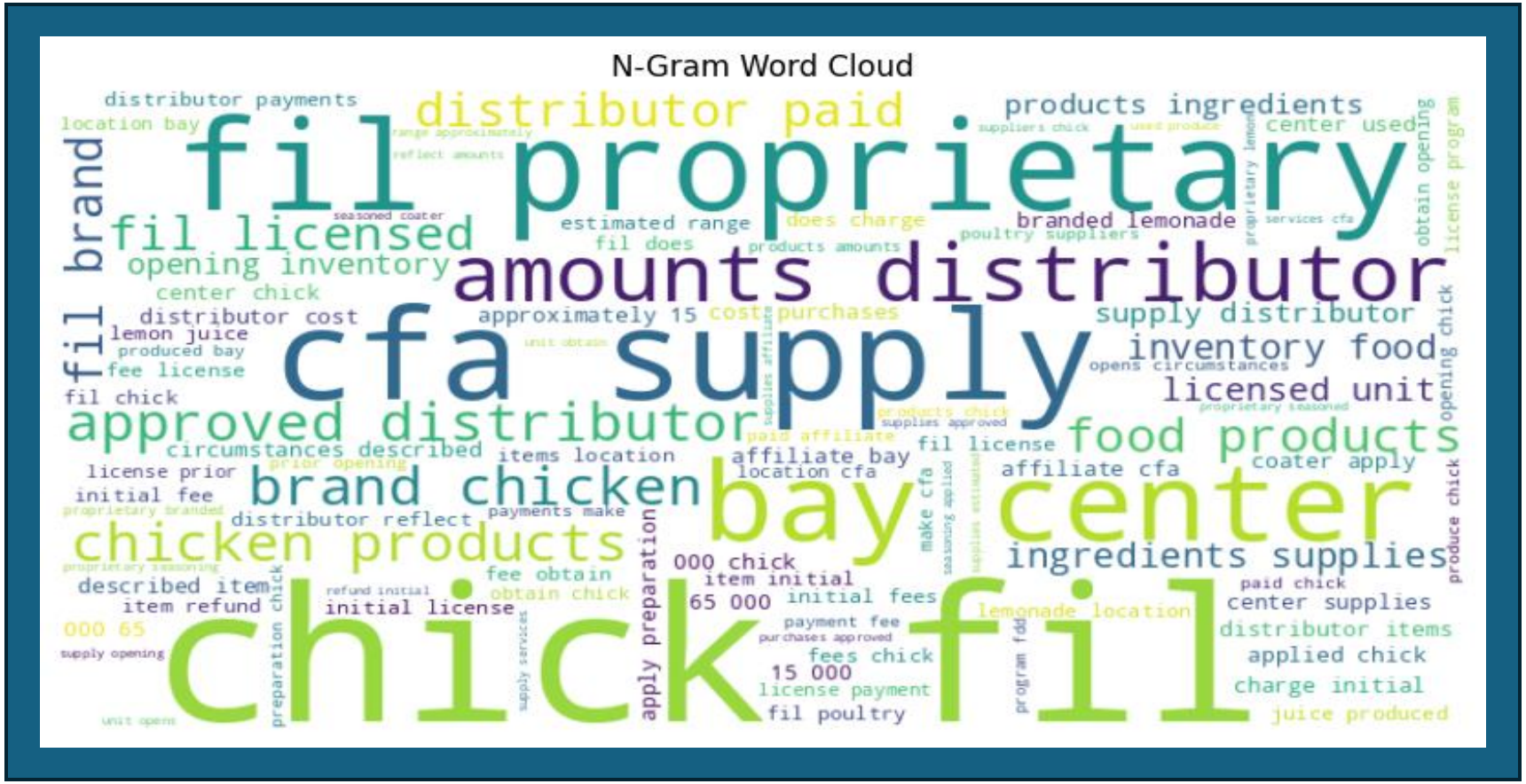


- **Ranked by word frequency**
  - **most common words:** “chick: and “fil”
- **Consistent with context of text**

N-GRAM

Functions Used

```
CountVectorizer(ngram_range=(2,2), stop_words='english')
n_gram_vectorizer.fit_transform(documents)
```

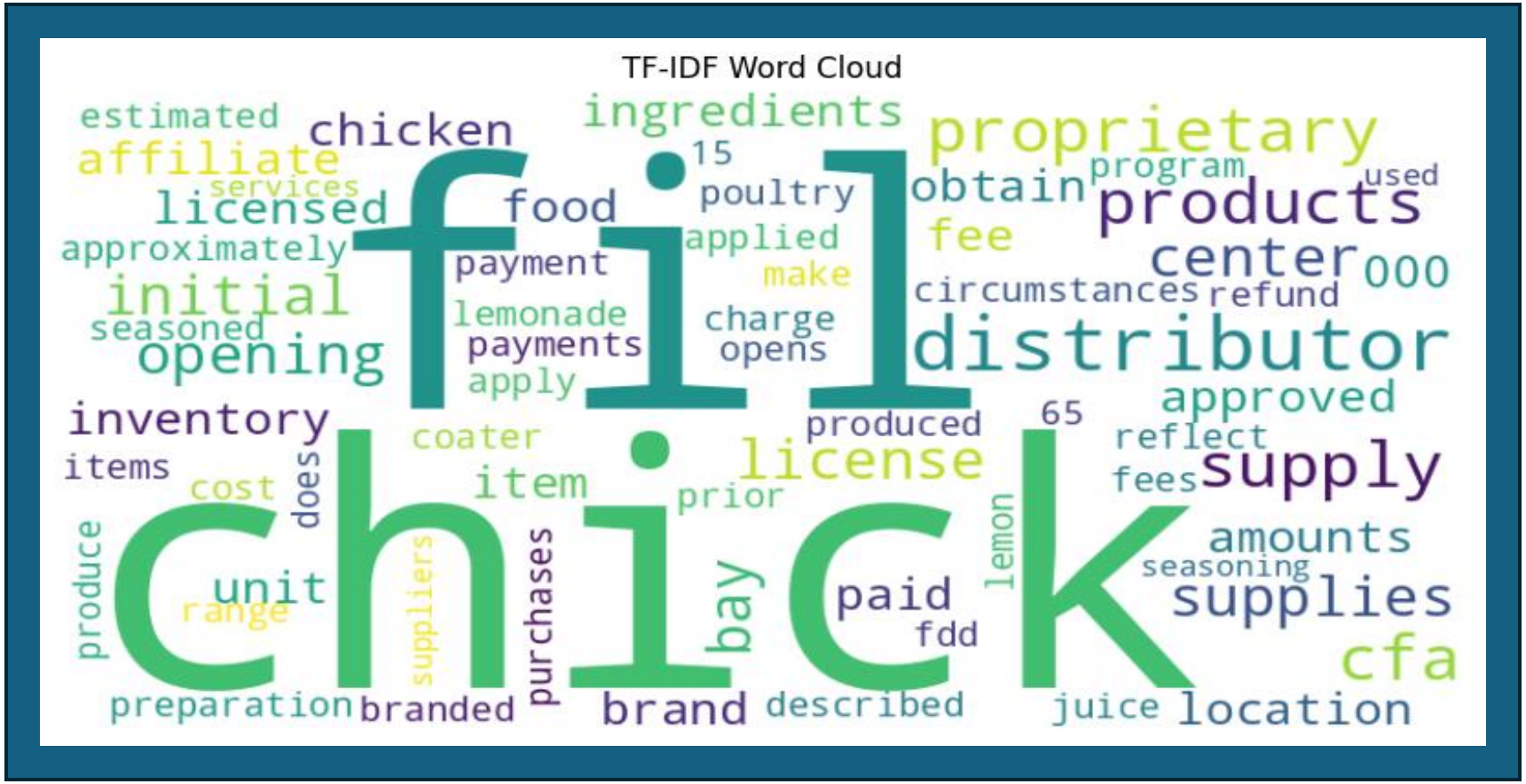


- **common 2-word phrases (bigrams)**
- **more context (& denser cloud)**
- **Considers word order and frequency**

Term Frequency-Inverse Document Frequency (TF-IDF)

Functions Used

```
TfidfVectorizer(stop_words='english’)
tfidf_vectorizer.fit_transform(documents)
```

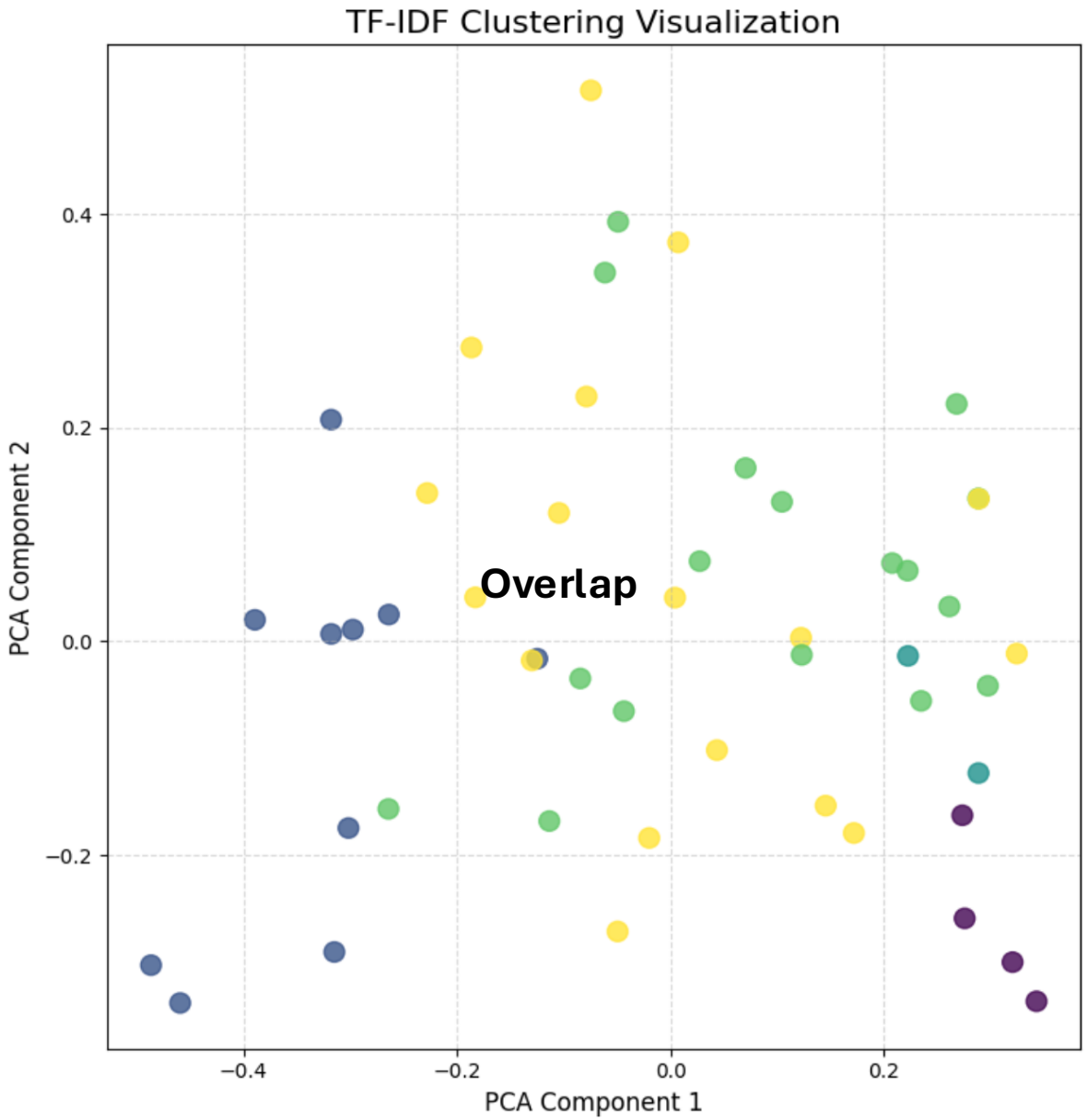


- **Ranked by TF-IDF score:** relevancy of a word or phrase to a document
  - **highest:** “chick,” “fil”
- **keywords unique to dataset**
- **similar to BoW word cloud**

TOPIC MODELING

...for all 50 ”Item 5” Files

FREQUENCY-BASED METHOD: TF-IDF



Original Topic Assignment

Franchise Agreements & Restaurant Businesses = “franchise” & “restaurant”

Restaurant Development & Franchise Targets = “development” & “target”

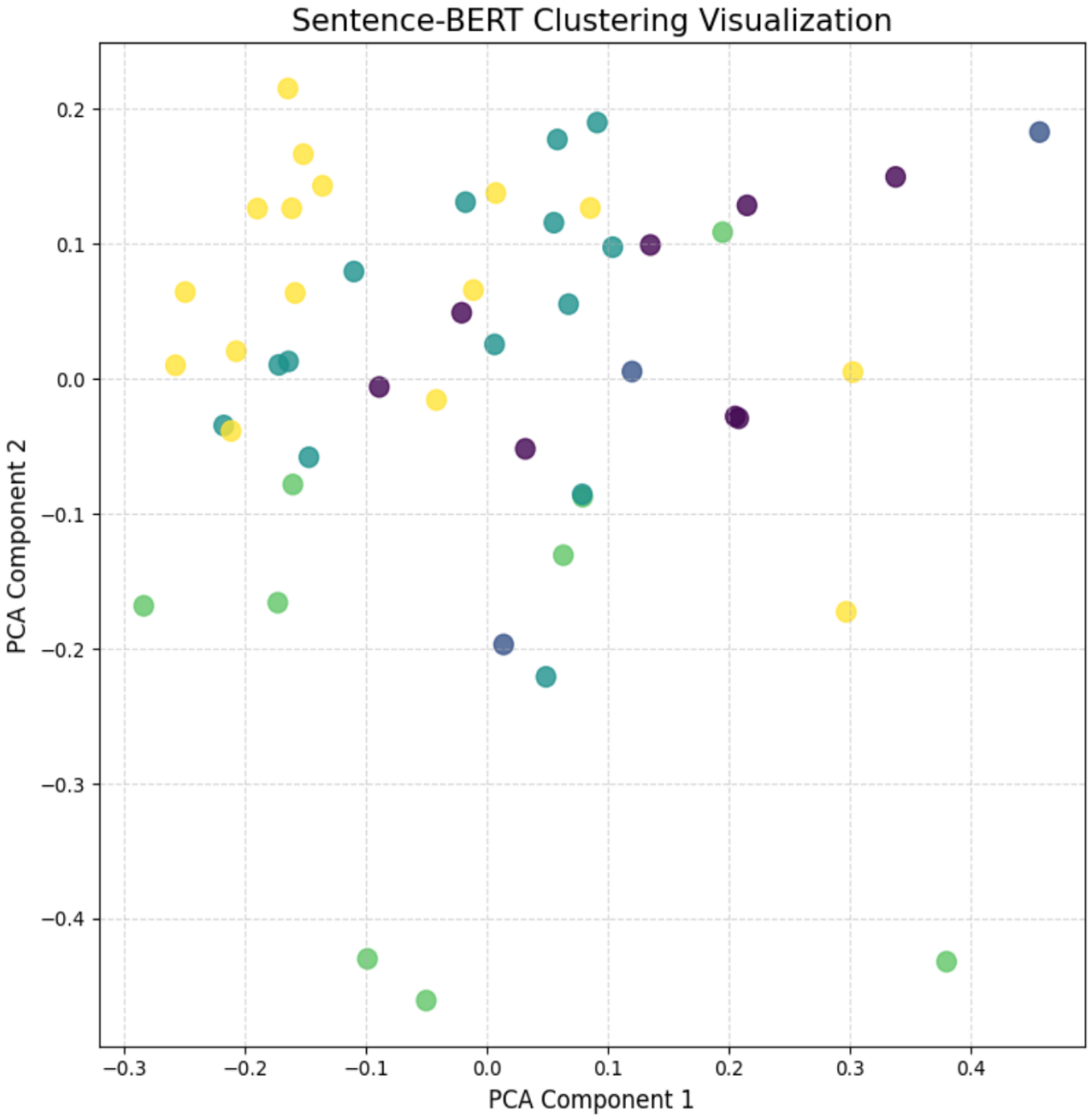
Retail & Service Franchises = “store” & “products”

Franchise Development & Agreements = “fee” & “agreement”

Restaurant Development & Incentives = “traditional” & “incentive”

Cluster	Food Franchise(s)	Top Terms
Franchise Agreements & Restaurant Businesses	McDonald's, KFC, Circle K, Wendy's, Pizza Hut, REMAX, Dunkin', Tim Hortons, Dairy Queen, Chili's, Berkshire Hathaway, Denny's, Hardee's, Zaxbys, Carl's Jr., Midas, Homevesters	franchise, fee, initial, restaurant, agreement, development, pay, training, franchised, shop
Franchise Agreements & Restaurant Businesses	Burger King, Popeyes Louisiana Kitchen	mtra, restaurant, development, franchise, target, popeyes, tra, franchised, deposit, restaurants
Retail & Service Franchise	Ace Hardware, Chick-fil-A, Domino's, Keller Williams, The UPS Store, Servpro, Valvoline Instant Oil Change, Snap-On Tools, Anytime Fitness, Jiffy Lube	center, fee, franchise, initial, store, license, products, viocf, supply, chickfila
Franchise Agreements & Restaurant Businesses	Subway, Taco Bell, Panera Bread, Little Caesars, Papa John's, Applebees, Express Employment Professionals, Jack In The Box, Buffalo Wild Wings, Planet Fitness, IHOP, Five Guys, Culver's, Wingstop, Jersey Mikes, Paris Baguette, Home Instead	franchise, fee, development, agreement, initial, restaurant, area, fees, program, purchase
Franchise Agreements & Restaurant Businesses	Sonic Drive-In, Arby's, Jimmy John's, Baskin Robbins	restaurant, development, nro, agreement, incentive, traditional, franchise, fee, restaurants, deeper

EMBEDDING-BASED METHOD: Sentence-BERT



Original Topic Assignment

Diverse Franchise Businesses = “McDonald’s” & “KFC”

Retail & Automotive Franchises = “Ace Hardware” & “Midas”

Food & Beverage Franchises = “Chick-fil-A” & “Pizza Hut”

Convenience Stores & Fitness Franchises = “Circle K” & “Planet Fitness”

Food & Service Franchises = “Tim Hortons” & “Jiffy Lube”

Cluster	Franchise(s)	Key Observations
Diverse Franchise Businesses	McDonald's, KFC, Burger King, Subway, Wendy's, REMAX, Keller Williams, Popeyes Louisiana Kitchen, Applebees, Express Employment Professionals, Culver's, Home Instead, Carl's Jr., Homevesters	Food & Non-Food Franchises
Retail & Automotive Franchises	Ace Hardware, The UPS Store, Midas	Exclusively Non-Food Franchises
Food & Beverage Franchises	Chick-fil-A, Domino's, Taco Bell, Pizza Hut, Dairy Queen, Little Caesars, Papa John's, Arby's, Chili's, Jack In The Box, IHOP, Denny's, Wingstop, Hardee's, Jimmy John's, Zaxbys	Major Dine-In Food Franchises
Convenience Stores & Fitness Franchises	Circle K, Dunkin', Panera Bread, Buffalo Wild Wings, Planet Fitness, Five Guys, Jersey Mikes, Paris Baguette, Valvoline Instant Oil Change	(Dine-In) Food, Non-Food, Gym Franchises
Food & Service Franchises	Tim Hortons, Sonic Drive-In, Berkshire Hathaway, Servpro, Baskin Robbins, Snap-On Tools, Anytime Fitness, Jiffy Lube	Exclusively Food & Service

QUESTION-ANSWERING TASK

...for all 50 ”Item 5” Files

PROCESS

Using RoBERTa as pre-trained model

- Trained on a larger corpus than BERT

Tokenization (>4 tokens) and segment identifications of question-and-answer texts.
Conducts numerical extraction (i.e., currency ($))

Returns probability score and answer text.

Ask RoBERTa 4 questions:

1. 'What is the initial fee?'
2. 'What is the initial fee in dollars?'
3. **'What is the initial franchise fee in dollars?'**
4. 'What is the initial nonrefundable franchise fee in dollars (standard)?'

Selects alternative with best performance.

Results

Evaluation Metrics for Each Question

Question	EM Score	F1 Score	Manual Scoring*
1	0.66	0.66	0.79
2	0.66	0.66	0.77
3	0.76	0.76	0.80
4	0.62	0.62	0.68

***including partial credit**

Item 5: Predicted vs Actual Initial Fees

File Name	True Initial Fee	Alternative Answer(s)	What is the initial fee?	What is the initial fee in dollars?	What is the initial franchise fee in dollars?	What is the initial nonrefundable franchise fee in dollars (standard)?
KFC…	$45,000	-	$45,000	$45,000	$45,000	$45,000
Burger King…	$50,000	$25,000 (conditional FSS) $15,000 (shorter period)	$25,000	$50,000	$50,000	$50,000
Chick-Fil-A…	$5,000	-	$5,000	$5,000	$5,000	five thousand dollars