Extracting Information from Text Data Using Natural Language Processing

Group 6

Pamela Alvarado-Zarate
Vishwanath Garlapalli
Mark Hite
Harshal Kamble
Venkatesh Subramony

Introduction

Frequency-based methods and embedded-based methods were utilized for topic modeling to extract keywords and/or topics from the agreements and terms of a franchise in which a new licensee must comply with. Such agreements and terms are separated as “Item 5: Initial Fees” and “Item 12: Territory” text files.

The purpose of topic modeling is to extract common themes or topics from a given text, where in this case, across several large text files from different franchises. Each topic, whether that be “Non-Food Franchise” or “Food Franchises” as examples, meet a certain criterion. The criteria can include the frequency and/or presence of specific words. Furthermore, word cloud visualizations display a word summary, where word frequency or relevancy is ranked by text size. These visuals provide a quick view into the text, which can be helpful in keyword searching.

Searching for information across several files can be a mundane and time-consuming task. A question-Answering model is a quick and alternative method to manual file inspection and analysis. As the name suggests, question-answering models answer the user’s questions based on the context of the dataset provided. In this case, the model is deployed to reply with the initial fees of a new licensee operating a franchise or determine if their expansion proposal will be approved based on the agreements stated in Item 5 and Item 12.

The Dataset

The chosen text file “Chick-Fil-A Item 5” for word cloud visualization outlines the requirements and expectations of the licensee and the initial fees once a Chick-Fil-A license is obtained. Each paragraph in the text file corresponds to a guideline or rule set by Chick-Fil-A. A few requirements to list, as stated from the file, include inventory before opening, choosing or partnering with an approved distributor, and the due payment for inventory purchases. Expectations from the licensee include using a specific supplier (i.e., Bay Center) for specific products, maintaining inventory, and making on-time payments.

All English stopwords were removed from the original “Chick-Fil-A Item 5” list or dataset to be used for all frequency-based methods.

All 40 files of Item 5 text files were used to apply a frequency-based and embeddings-based method, separately. Across all Item 5 text files, the texts’ context is similar to the “Chick-Fil-A Item 5” file. Each food franchise adds additional agreements and terms that are dependent on their discretion.

To assess the accuracy of the Question Answering model, actual values from “Item 5_Answers” were compared with the predicted values.

Five (5) topics were defined via K-Means clustering for the frequency-based and embeddings-based methods.

Item 5_Answers

The “Item 5_Answers” dataset consists of seven (7) features with at most 50 entries each. The data type for all features is “object.” The features are the following:

- **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.
    - Answers are separated into 5 columns; each column with one answer.

Frequency-Based Methods

Bag of Words (BoW)

The A Bag of Words (BoW) word cloud displayed all the words stated in the text file. Each word was ranked based on its occurrence in the text with the rank displayed as text size in the word cloud. The three (3) most common words found in the text were “chick,” “fil,” and “distributor.” All the other words describe the type of inventory and the transactions between business and licensee. The common and following words are consistent with the context of the text: Chick-Fil-A licensee choosing a distributor for inventory. The BoW word cloud only accounts for word frequency, not word order or meaning.

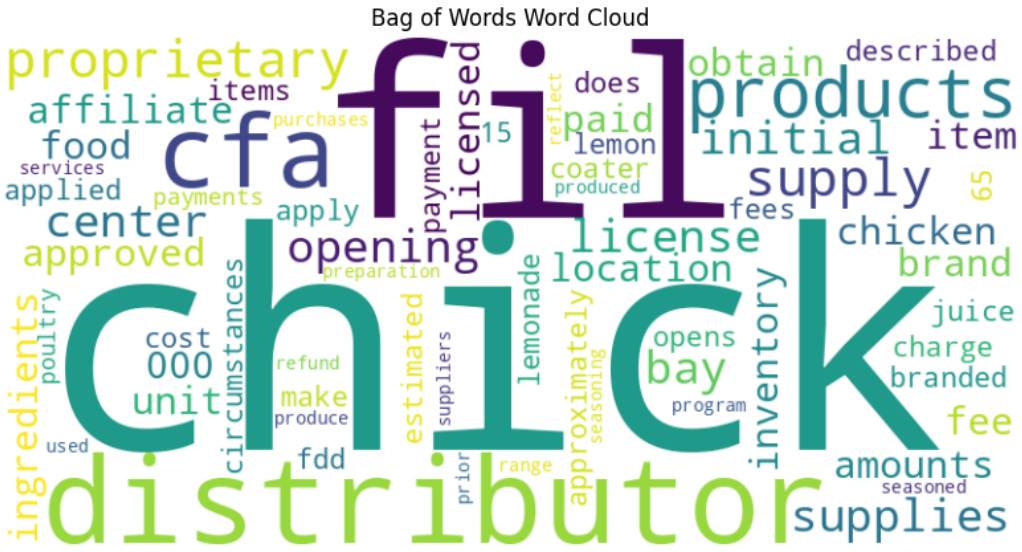

Figure 1. Bag of Words (BoW) word cloud of Chick-Fil-A Item 5 dataset. Word frequency is determined by text size.

N-Gram (Bigram)

The n-gram, specifically a bigram, word cloud displayed unique pairs of consecutive words or two-word phrases. Each phrase was ranked based on its occurence in the text with the rank displayed as text size in the word cloud. The three (3) most common phrases found in the text were “chick fil,” “cfa supply,” and “fil proprietary.” The displayed words remain consistent with the text’s context. In comparison to the BoW word cloud, the bigram word cloud is denser, can be used to extract more context, and accounts for word frequency, order, and uniqueness.

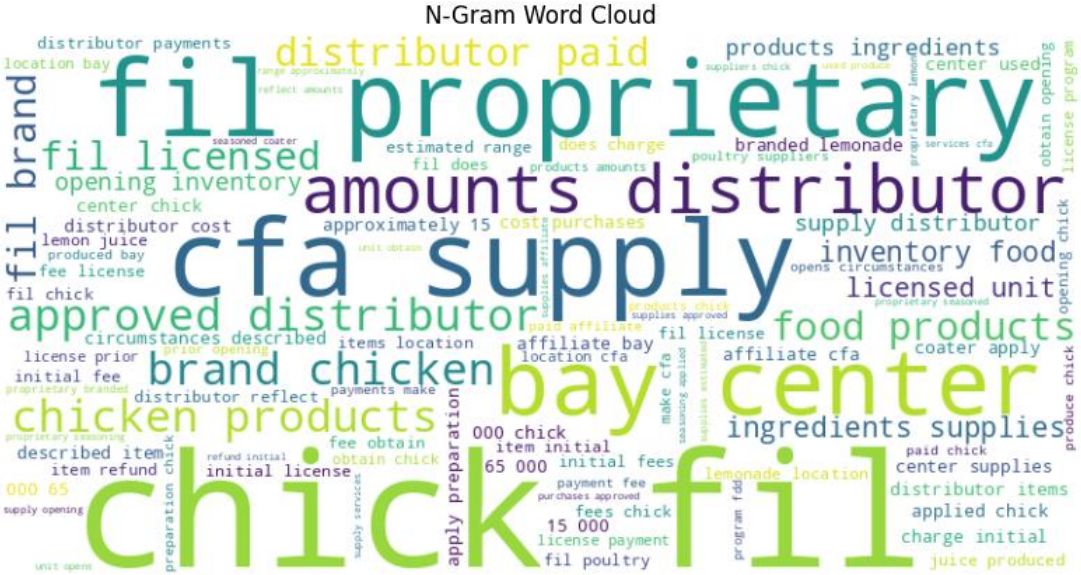

Figure 1. Bag of Words (BoW) word cloud of Chick-Fil-A Item 5 dataset. Phrase frequency is determined by text size.

Term Frequency-Inverse Document Frequency (TF-IDF)

The term frequency-inverse document frequency (TF-IDF) word cloud displayed all words relevant to the text. Each word was ranked based off its TF-IDF score or its relevancy to the text with the rank displayed as text size in the word cloud. The most relevant words found in the text were “chick” and “fil,” which are truly unique to the text as it concerns with the franchise Chick-Fil-A.

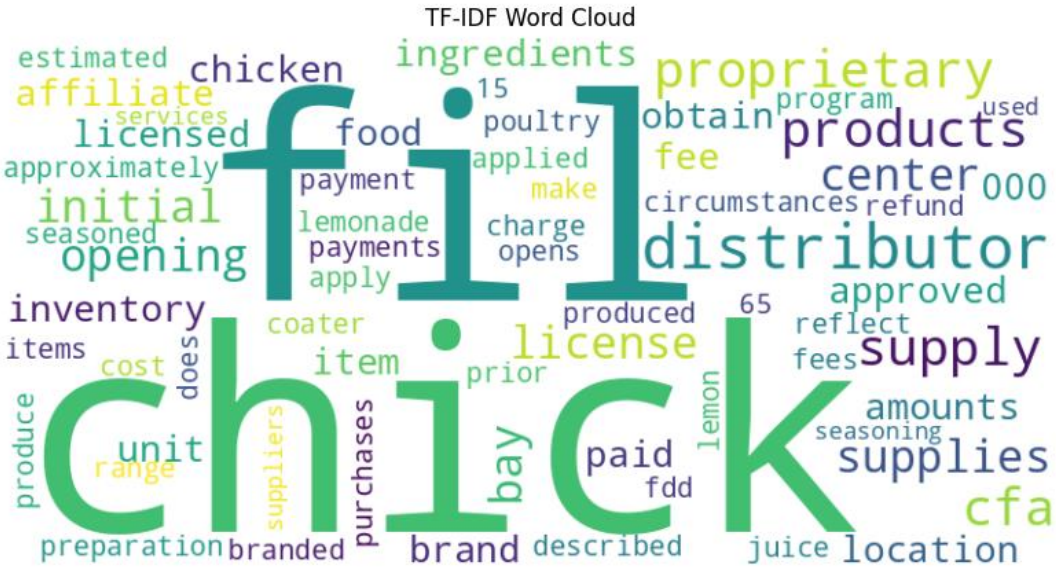

Figure 2. Term Frequency-Inverse Document Frequency (TF-IDF) word cloud of Chick-Fil-A Item 5 dataset. Word relevancy to dataset is determined by text size.

Combination of Methods

Based on the criteria set beforehand, the TF-IDF method identified 2 different topics but grouped in 5 clusters as expected. The topics identified, “Franchise Agreements & Restaurant Businesses” and “Retail & Service Franchises,” meet the criterion of its top terms. While all text files had “franchise” and “development” as their top terms, other files did not have similar/exact top terms as the other. As so, the topic “Franchise Agreements & Restaurant Businesses” was further divided into alternate versions since the top terms across the text files overlapped. This is further supported by Figure 1. The PCA components indicate that the clusters are not distinctly separated and thus, literally overlapping due to the lack of major differences within the clusters. For instance, the clusters of the last 2 topics of “Franchise Agreements & Restaurant Businesses” significantly overlap, likely indicating that a large similarity is present with little variance or difference.

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

Table 1. Topic modeling and clustering of franchises using TF-IDF.

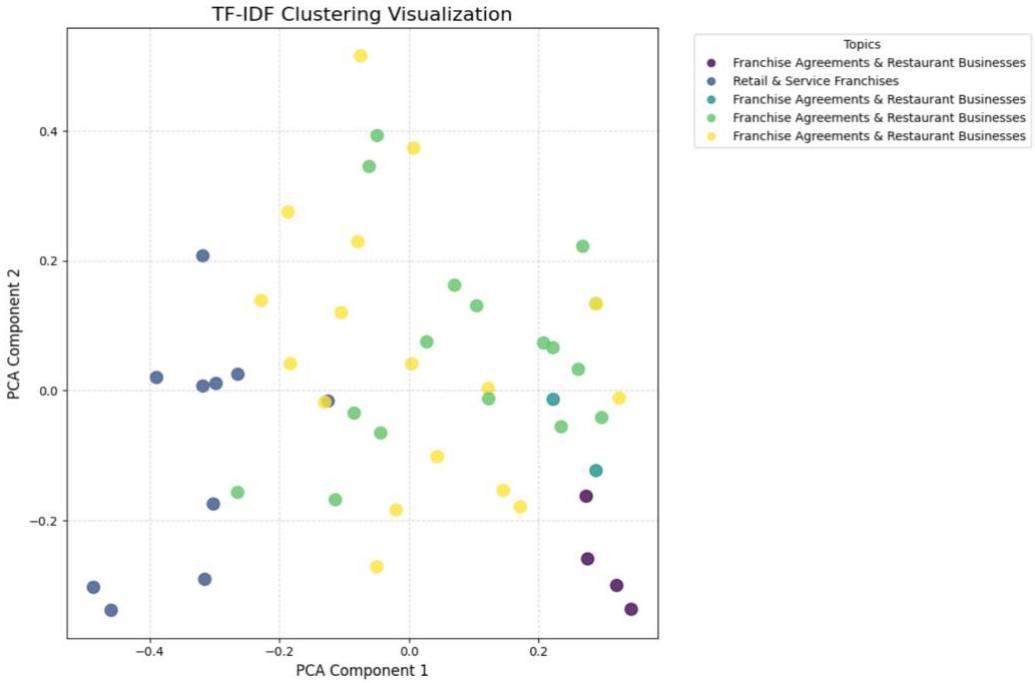

Figure 5. TF-IDF clustering of franchises. Clustering defined with K-Means and topics modeling based on top terms.

The Sentence-BERT method identified exactly 5 topics and clusters as previously defined. Examples of franchises were used as the criterion for topic modeling. For instance, examples of “Food & Service Franchises” included McDonald’s and Kentucky Fried Chicken to exclusively include food-related franchises. Or, examples of “Retail &

Service Franchises" included "Midas" and Ace Hardware" to exclusively include retail and automotive franchises. Unusual topic grouping that was observed was "Five Guys" or "Panera Bread" as "Convenience Stores & Fitness Franchises," which can suggest that some terms, such as "quick" or "convenience," were found in the text files. Therefore, these franchises were selected due to their quick and convenient service even if they offer dining and are strictly food-related only. In reference to Figure 5, the clusters are significantly scattered across the graph, indicating that while each franchise was assigned to one topic, they overlap due to similar characteristics. For instance, Chick-Fil-A can be be considered as both a "Food & Service Franchise" and "Food & Beverage Franchise" because it sells food and beverages and provides top-notch service.

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

Table 2. Topic modeling and clustering of franchises using Sentence-BERT.

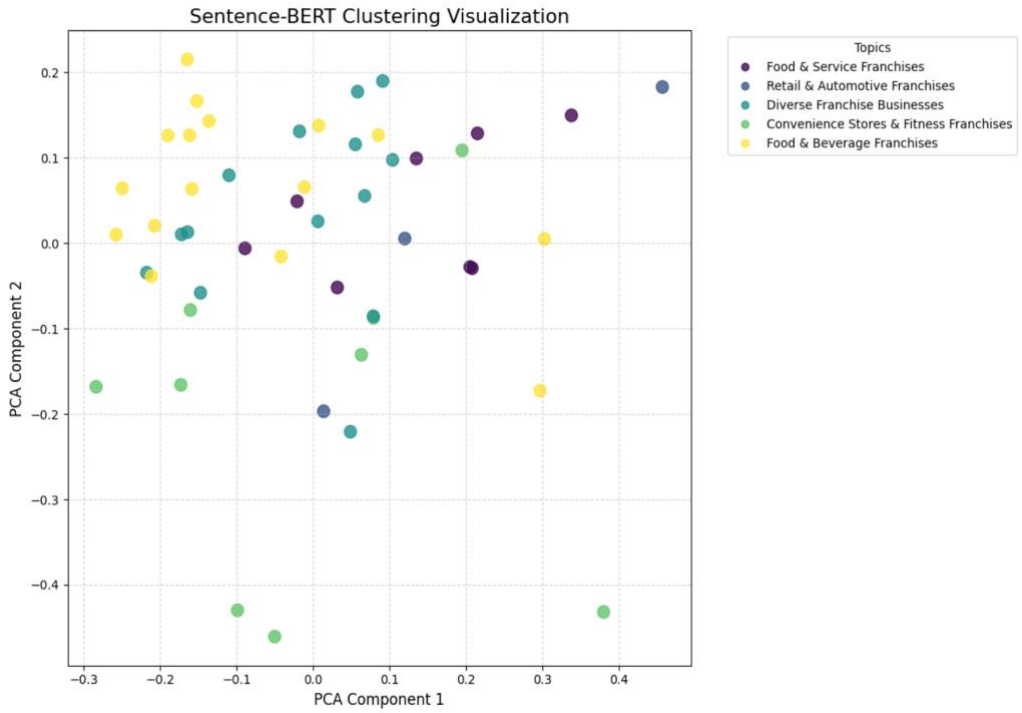

Figure 5. Sentence-BERT clustering of franchises. Clustering defined with K-Means and topics modeling based on top terms.

Question-Answering

For the Q&A task, RoBERTa was used as the pre-trained model. The model is provided with 4 questions to extract the initial fee from the text:

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)?'

The table below shows the true vs. predicted values for each question.

filename	True Initial Fee	Prediction 1	Prediction 2	Prediction 3	Prediction 4
1 McDonald's.pdf-all_items_item_5.txt	$45,000	$45,000	$45,000	$45,000	$22,500
3 KFC.pdf-all_items_item_5.txt	$45,000	$45,000	$3,000	$45,000	$45,000
4 Burger King.pdf-all_items_item_5.txt	$50,000	$25,000	$25,000	$25,000	$50,000
5 Ace Hardware.pdf-all_items_item_5.txt	$5,000	$5,000	$5,000	$5,000	$5,000
6 Chick-fil-A.pdf-all_items_item_5.txt	$0	$0	$0	$0	$0
7 Subway.pdf-all_items_item_5.txt	$15,000	$15,000	$15,000	$15,000	$15,000
8 Domino's.pdf-all_items_item_5.txt	$10,000	$0	$88,950	$88,950	$88,950
9 Circle K.pdf-all_items_item_5.txt	$25,000	$25,000	$25,000	$25,000	$60,000

10 Taco Bell.pdf-all_items_item_5.txt	$22,500	$22,500	$22,500	$22,500	$25,000
11 Wendy's.pdf-all_items_item_5.txt	$50,000	$50,000	$50,000	$50,000	$50,000
12 Pizza Hut.pdf-all_items_item_5.txt	$25,000	$25,000	$25,000	$25,000	$25,000
13 REMAX.pdf-all_items_item_5.txt	$35,000(high density)	$0	$1,000	$25,000	$1,000
14 Dunkin'.pdf-all_items_item_5.txt	$90,000(Type1)	$4,000	$45,000	$22,500	$10,000
15 Keller Williams.pdf-all_items_item_5.txt	$35,000	$35,000	$35,000	$35,000	$35,000
16 Tim Hortons.pdf-all_items_item_5.txt	$50,000(STANDARD)	$25,000	$25,000	$25,000	$50,000
17 Panera Bread.pdf-all_items_item_5.txt	$35,000	$0	$0	$35,000	$0
18 Popeyes Louisiana Kitchen.pdf-all_items_item_5.txt	$50,000	$50,000	$50,000	$50,000	$50,000
19 Dairy Queen.pdf-all_items_item_5.txt	$45,000	$45,000	$45,000	$45,000	$45,000
20 Sonic Drive-In.pdf-all_items_item_5.txt	$45,000(traditional)	$30,000	$30,000	$30,000	$10,000
21 Little Caesars.pdf-all_items_item_5.txt	$20,000	$20,000	$20,000	$20,000	$20,000
22 Papa John's.pdf-all_items_item_5.txt	$5,000-$25,000	$5,000	$5,000	$5,000	$5,000
23 Arby's.PDF-all_items_item_5.txt	$37,500(standard)	$12,500	$12,500	$37,500	$37,500
24 Applebees.pdf-all_items_item_5.txt	$35,000	$35,000	$35,000	$35,000	$15,000
25 Express Employment Professionals.pdf-all_items_item_5.txt	$40,000	$40,000	$40,000	$40,000	$25,000
26 Chili's.pdf-all_items_item_5.txt	$60,000	$0	$40,000	$60,000	$60,000
27 Jack In The Box.pdf-all_items_item_5.txt	$50,000	$50,000	$50,000	$50,000	$50,000
28 Buffalo Wild Wings.pdf-all_items_item_5.txt	$30,000	$5,000	$5,000	$15,000	$0
29 Planet Fitness.pdf-all_items_item_5.txt	$20,000	$20,000	$20,000	$20,000	$20,000
30 Berkshire Hathaway.pdf-all_items_item_5.txt	$25,000	$25,000	$25,000	$25,000	$25,000
31 The UPS Store.pdf-all_items_item_5.txt	$29,950	$7,500	$7,500	$29,950	$29,950
32 Servpro.pdf-all_items_item_5.txt	$90,000	$90,000	$90,000	$90,000	$90,000
33 IHOP.pdf-all_items_item_5.txt	$50,000(single)	$50,000	$0	$0	$0
35 Five Guys.pdf-all_items_item_5.txt	$25,000	$25,000	$25,000	$25,000	$25,000
36 Denny's.pdf-all_items_item_5.txt	$30,000	$30,000	$30,000	$30,000	$30,000
37 Culver's.pdf-all_items_item_5.txt	$55,000	$55,000	$55,000	$55,000	$45,000
38 Wingstop.pdf-all_items_item_5.txt	$20,000	$5,000	$20,000	$20,000	$20,000
39 Jersey Mikes.pdf-all_items_item_5.txt	$18,500	$8,500	$18,500	$25,500	$5,000
40 Paris Baguette.pdf-all_items_item_5.txt	$50,000	$0	$0	$0	$0
41 Hardee's.pdf-all_items_item_5.txt	$25,000	$25,000	$25,000	$25,000	$25,000
42 Jimmy John's.pdf-all_items_item_5.txt	$35,000	$0	$1,500	$25,000	$35,000
43 Vavoline Instant Oil Change.pdf-all_items_item_5.txt	$30,000(first center)	$30,000	$30,000	$30,000	$5,000
44 Home Instead.pdf-all_items_item_5.txt	$54,000	$54,000	$54,000	$54,000	$54,000
45 Zaxbys.pdf-all_items_item_5.txt	$35,000	$35,000	$35,000	$35,000	$7,000
46 Carl's Jr..pdf-all_items_item_5.txt	$25,000	$25,000	$25,000	$25,000	$25,000
47 Baskin Robbins.pdf-all_items_item_5.txt	$25,000	$12,500	$12,500	$12,500	$6,000
49 Snap-On Tools.pdf-all_items_item_5.txt	$8,000 --$16,000	$8,000	$2,800	$8,000	$8,000
50 Anytime Fitness.pdf-all_items_item_5.txt	$42,500	$42,500	$42,500	$42,500	$42,500
51 Jiffy Lube.pdf-all_items_item_5.txt	$35,000 (new center)	$17,500	$35,000	$35,000	$10,000
52 Midas.pdf-all_items_item_5.txt	$35,000 (new center)	$35,000	$35,000	$35,000	$35,000
54 Homevesters.pdf-all_items_item_5.txt	$85,000 (Full)	$85,000	$85,000	$85,000	$85,000

Table 3. Question-Answering predicted & actual initial fees using RoBERTa.

Evaluation metrics were calculated for each question, shown in the table below. EM represents the ratio of responses that had an 'Exact Match' with the true value. F1 represents the harmonic mean of the precision and recall. Manual scoring is the score obtained after giving 1 point to responses that match the true value, and half a point to those that match one of the alternative answers. Overall, question 3 had the best performance on the Q&A task.

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

Table 2. Evaluation Metrics for Each Question