Codalab R2VQ - Competence-based Multimodal Question Answering

Authors: Arihant Bhat (abhat8@uncc.edu), Jyothi Ravipudi (jravipud@uncc.edu), Vishal Reddy Pachika (vpachika@uncc.edu), Kishore Jayasai Vutukury (kvutukur@uncc.edu), Srilakshmi Neha Yalamanchili (syalama5@uncc.edu)

Current date: *October 24th, 2021*

1. ABSTRACT:

The R2VQ (Recipe to Video Questions) challenge is divided into question-answer pairs that test how effectively a system understands the semantics of recipes drawn from a library of cooking recipes and movies. Each question is part of a "question family" that reflects a certain level of thinking ability. The linked R2VQ dataset is intended for assessing computers' competence-based comprehension of a multimodal recipe collection. We will utilize the R2VQ dataset given by codalab for our study, which consists of a collection of recipes taken from https://recipes. fandom.com/wiki/Recipes_Wiki and foodista.com, and they are labeled according to three distinct annotation layers: (i) Cooking Role Labeling (CRL), (ii) Semantic Role Labeling (SRL), and (iii) aligned image frames taken from cooking videos in the YouCook2 dataset. The R2VQ corpus will contain 1,000 recipes, each having an expected average of ten ingredients, eight words, and 35 tokens per step. For training, 800 recipes will be utilized, with 100 recipes for validation and testing.

2. INTRODUCTION:

This project is primarily concerned with how well a system understands the semantics of recipes obtained from a collection of cooking recipes and videos when a question is posed and the system responds with an answer.

This approach consists of three basic phases, which are as follows:

1. Data Transformation: Parsing the conllu annotation schema dataset and converting it to JSON format.
2. The dataset is divided into training and testing phases.
3. Training the model: We will train the model using a multimodal training dataset, which will respond to unknown questions using either visual and textual information combined or textual information only.

3. PROBLEM STATEMENT:

Our goal is to build a system that delivers responses to open-ended questions based on the textual and visual information encoded in the dataset. All systems will be assessed entirely on the responses to the questions. For answers to cardinality, yes/no, and unanswerable questions, accuracy will be used to compute the results. For responses to natural language questions, exact match, word-level F1 will be used to compute the results.

4. DATA DESCRIPTION :

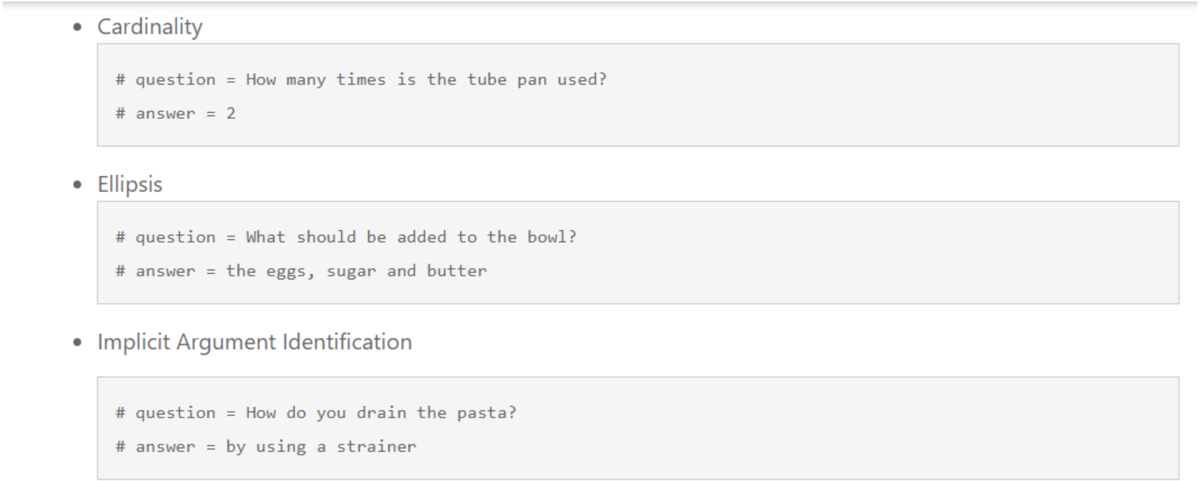
Input Data :

The training data provided for textual extraction is a csv file which is in CoNLL-U format and has the details of recipes sourced from the above two mentioned websites. These ConLlu files consist of Id and text(a line from recipe) along with a set of question and answers to train the data. Each word in the text is represented in seperate lines and these sentences are separated by a next line. Each column represents one annotation(string value about a word) and every word in the corpus has the same number of columns. For Visual extraction images are sourced from YouCook2 dataset.

Output Data :

Each template is associated with a functional program. It contains a set of functions that allows you to query and filter the annotated recipe to get the answer to that template-based question.

Sample Question-Answer Pairs :



- Cardinality

```
# question = How many times is the tube pan used?
# answer = 2
```

- Ellipsis

```
# question = What should be added to the bowl?
# answer = the eggs, sugar and butter
```

- Implicit Argument Identification

```
# question = How do you drain the pasta?
# answer = by using a strainer
```

4.1 Subsection motivating your approach

Our approach is divided into phases: Preprocessing of data, Analyzing the dataset and determining the best features suitable for classification, Training the model and Testing the model on a validation set.

4.2 Subsection describing the details of your methods

Creating DataSet :

Data set is built for the recipe websites sourced from foodista.com and recipes.fandom.com/wiki/Recipes_Wiki, and labeled accordingly to make the system understand the semantics of recipes.

Preprocessing of data :

As part of pre-processing, Worked on creating the dataset from the given websites using the BeautifulSoup library in Python.

Analyzing the dataset :

We have a single text file each for train, validation and test split. Each file has the annotation of CoNLL-U format - A textual format specific to the Universal Dependencies corpora where each line represents a word in a sentence, and columns represent features or labels of that word. Parsed the data using the conllu module. Once parsed, a file is presented as a list of sentences, each containing tokens.

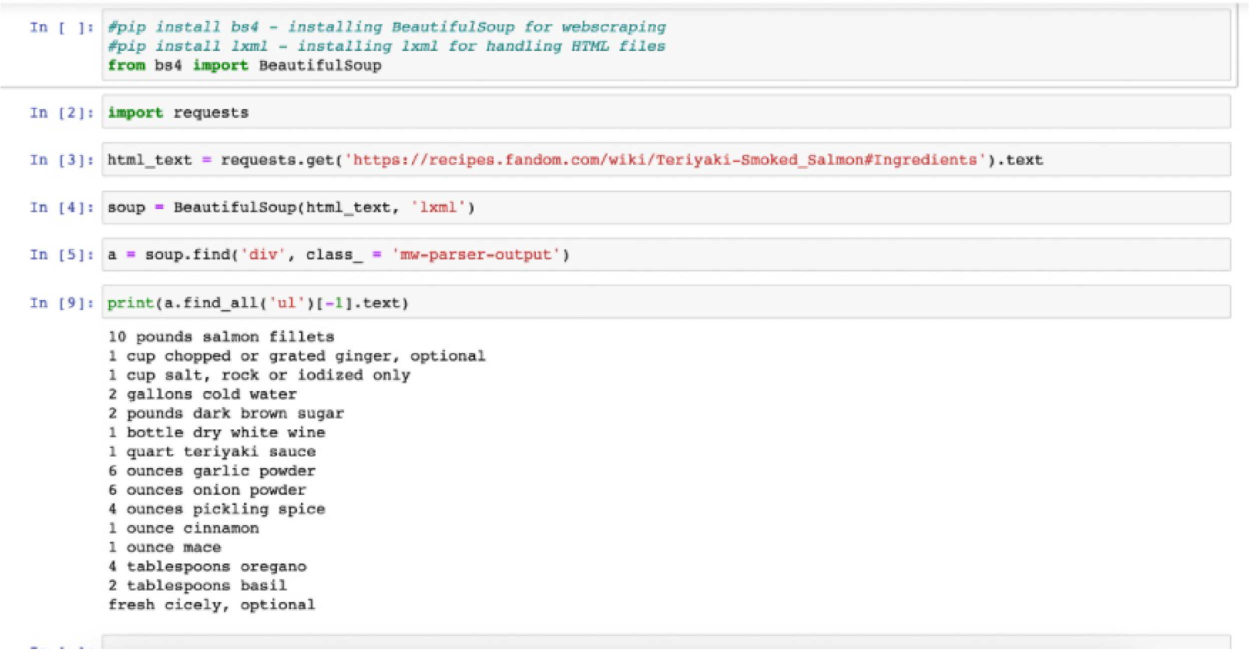
Training and validation :

For textual extraction we are using BERT model to train the dataset and for visual extraction we will be using Convolutional Neural networks for training and validation purpose

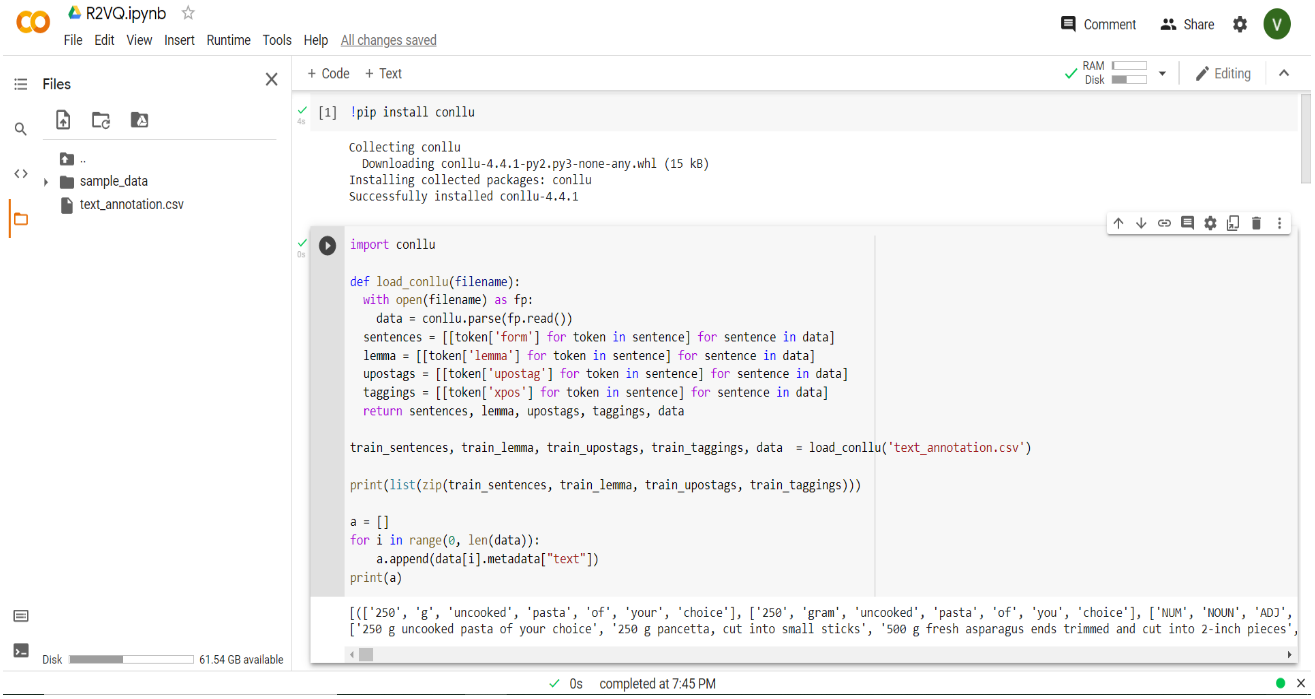
4.3 Subsection details of your experiments

Worked on creating the dataset from the given websites using the BeautifulSoup library in Python Tested CDQA (Closed Domain Question Answering) by taking recipe details in pdf format for textual extraction. Since training data is in Conllu format in codalab, parsed data using conllu module and extracted tokens. Since, Conllu format is difficult to train the data, we thought of converting the conllu format to json format and then train the data. For conversion, we are trying to use pyjsonnlp module converter.

Screenshot of experiments performed:



```
In [ ]: #pip install bs4 - installing BeautifulSoup for webscraping
        #pip install lxml - installing lxml for handling HTML files
        from bs4 import BeautifulSoup

In [2]: import requests

In [3]: html_text = requests.get('https://recipes.fandom.com/wiki/Teriyaki-Smoked_Salmon#Ingredients').text

In [4]: soup = BeautifulSoup(html_text, 'lxml')

In [5]: a = soup.find('div', class_ = 'mw-parser-output')

In [9]: print(a.find_all('ul')[-1].text)

        10 pounds salmon fillets
        1 cup chopped or grated ginger, optional
        1 cup salt, rock or iodized only
        2 gallons cold water
        2 pounds dark brown sugar
        1 bottle dry white wine
        1 quart teriyaki sauce
        6 ounces garlic powder
        6 ounces onion powder
        4 ounces pickling spice
        1 ounce cinnamon
        1 ounce mace
        4 tablespoons oregano
        2 tablespoons basil
        fresh cicely, optional
```



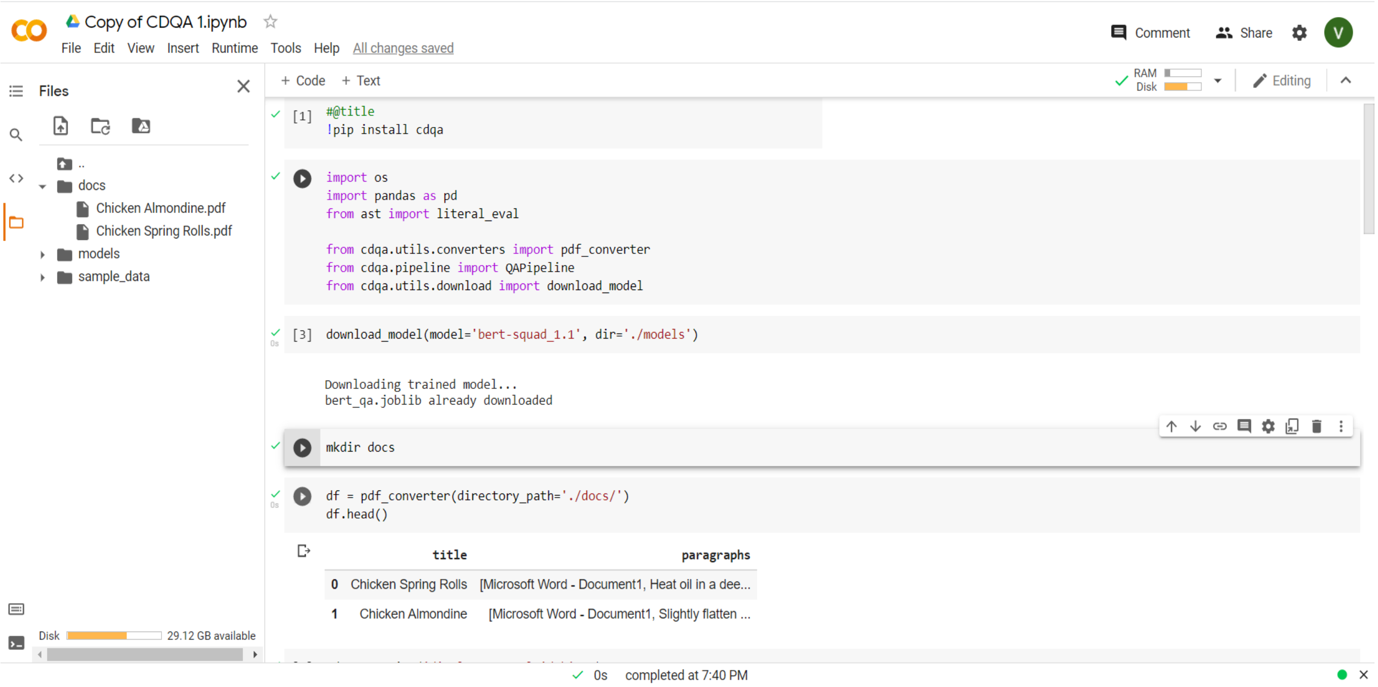
```
[1] !pip install conllu
```

```
Collecting conllu
  Downloading conllu-4.4.1-py2.py3-none-any.whl (15 kB)
Installing collected packages: conllu
Successfully installed conllu-4.4.1
```

```
import conllu

def load_conllu(filename):
  with open(filename) as fp:
    data = conllu.parse(fp.read())
  sentences = [[token['form'] for token in sentence] for sentence in data]
  lemma = [[token['lemma'] for token in sentence] for sentence in data]
  upostags = [[token['upostag'] for token in sentence] for sentence in data]
  taggings = [[token['xpos'] for token in sentence] for sentence in data]
  return sentences, lemma, upostags, taggings, data

train_sentences, train_lemma, train_upostags, train_taggings, data  = load_conllu('text_annotation.csv')

print(list(zip(train_sentences, train_lemma, train_upostags, train_taggings)))

a = []
for i in range(0, len(data)):
    a.append(data[i].metadata["text"])
print(a)
```

```
[(['250', 'g', 'uncooked', 'pasta', 'of', 'your', 'choice'], ['250', 'gram', 'uncooked', 'pasta', 'of', 'you', 'choice'], ['NUM', 'NOUN', 'ADJ',
['250 g uncooked pasta of your choice', '250 g pancetta, cut into small sticks', '500 g fresh asparagus ends trimmed and cut into 2-inch pieces',
```

✓ 0s completed at 7:45 PM



```
[1] #@title
    !pip install cdqa
```

```
import os
import pandas as pd
from ast import literal_eval

from cdqa.utils.converters import pdf_converter
from cdqa.pipeline import QAPipeline
from cdqa.utils.download import download_model
```

```
[3] download_model(model='bert-squad_1.1', dir='./models')
```

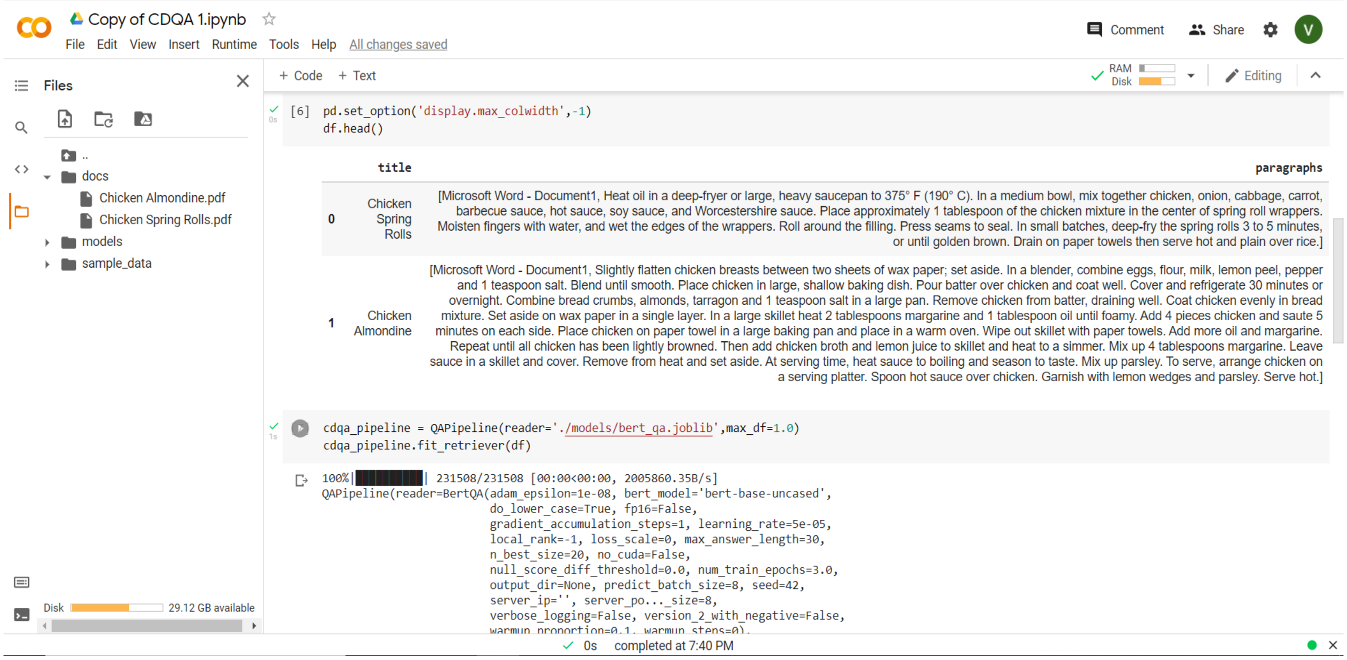
```
Downloading trained model...
bert_qa.joblib already downloaded
```

```
mkdir docs
```

```
df = pdf_converter(directory_path='./docs/')
df.head()
```

	title	paragraphs
0	Chicken Spring Rolls	[Microsoft Word - Document1, Heat oil in a dee...
1	Chicken Almondine	[Microsoft Word - Document1, Slightly flatten ...

✓ 0s completed at 7:40 PM



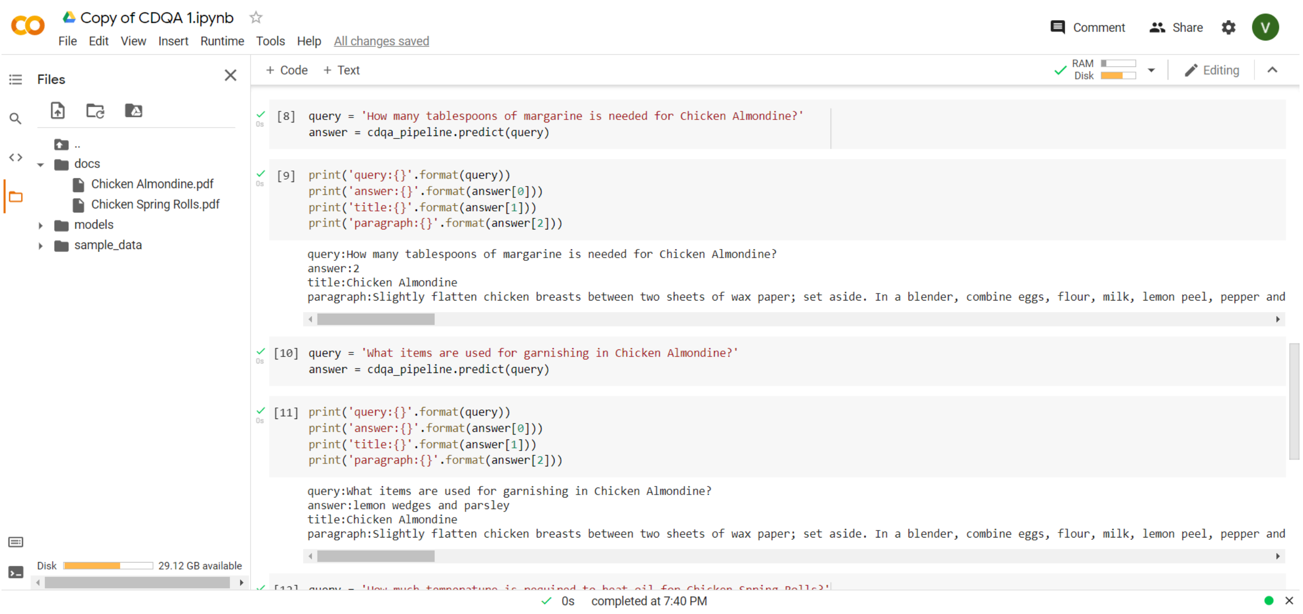
```
[6] pd.set_option('display.max_colwidth',-1)
    df.head()
```

	title	paragraphs
0	Chicken Spring Rolls	[Microsoft Word - Document1, Heat oil in a deep-fryer or large, heavy saucepan to 375° F (190° C). In a medium bowl, mix together chicken, onion, cabbage, carrot, barbecue sauce, hot sauce, soy sauce, and Worcestershire sauce. Place approximately 1 tablespoon of the chicken mixture in the center of spring roll wrappers. Moisten fingers with water, and wet the edges of the wrappers. Roll around the filling. Press seams to seal. In small batches, deep-fry the spring rolls 3 to 5 minutes, or until golden brown. Drain on paper towels then serve hot and plain over rice.]
1	Chicken Almondine	[Microsoft Word - Document1, Slightly flatten chicken breasts between two sheets of wax paper; set aside. In a blender, combine eggs, flour, milk, lemon peel, pepper and 1 teaspoon salt. Blend until smooth. Place chicken in large, shallow baking dish. Pour batter over chicken and coat well. Cover and refrigerate 30 minutes or overnight. Combine bread crumbs, almonds, tarragon and 1 teaspoon salt in a large pan. Remove chicken from batter, draining well. Coat chicken evenly in bread mixture. Set aside on wax paper in a single layer. In a large skillet heat 2 tablespoons margarine and 1 tablespoon oil until foamy. Add 4 pieces chicken and saute 5 minutes on each side. Place chicken on paper towel in a large baking pan and place in a warm oven. Wipe out skillet with paper towels. Add more oil and margarine. Repeat until all chicken has been lightly browned. Then add chicken broth and lemon juice to skillet and heat to a simmer. Mix up 4 tablespoons margarine. Leave sauce in a skillet and cover. Remove from heat and set aside. At serving time, heat sauce to boiling and season to taste. Mix up parsley. To serve, arrange chicken on a serving platter. Spoon hot sauce over chicken. Garnish with lemon wedges and parsley. Serve hot.]

```
cdqa_pipeline = QAPipeline(reader='./models/bert_qa.joblib',max_df=1.0)
cdqa_pipeline.fit_retriever(df)
```

```
100%|██████████| 231508/231508 [00:00<00:00, 2005860.35B/s]
QAPipeline(reader=BertQA(adam_epsilon=1e-08, bert_model='bert-base-uncased',
                         do_lower_case=True, fp16=False,
                         gradient_accumulation_steps=1, learning_rate=5e-05,
                         local_rank=-1, loss_scale=0, max_answer_length=30,
                         n_best_size=20, no_cuda=False,
                         null_score_diff_threshold=0.0, num_train_epochs=3.0,
                         output_dir=None, predict_batch_size=8, seed=42,
                         server_ip='', server_po..._size=8,
                         verbose_logging=False, version_2_with_negative=False,
                         warmup_proportion=0.1, warmup_steps=0),
```

✓ 0s completed at 7:40 PM



```
[8] query = 'How many tablespoons of margarine is needed for Chicken Almondine?'
    answer = cdqa_pipeline.predict(query)
```

```
[9] print('query:{}'.format(query))
    print('answer:{}'.format(answer[0]))
    print('title:{}'.format(answer[1]))
    print('paragraph:{}'.format(answer[2]))
```

```
query:How many tablespoons of margarine is needed for Chicken Almondine?
answer:2
title:Chicken Almondine
paragraph:Slightly flatten chicken breasts between two sheets of wax paper; set aside. In a blender, combine eggs, flour, milk, lemon peel, pepper and
```

```
[10] query = 'What items are used for garnishing in Chicken Almondine?'
     answer = cdqa_pipeline.predict(query)
```

```
[11] print('query:{}'.format(query))
     print('answer:{}'.format(answer[0]))
     print('title:{}'.format(answer[1]))
     print('paragraph:{}'.format(answer[2]))
```

```
query:What items are used for garnishing in Chicken Almondine?
answer:lemon wedges and parsley
title:Chicken Almondine
paragraph:Slightly flatten chicken breasts between two sheets of wax paper; set aside. In a blender, combine eggs, flour, milk, lemon peel, pepper and
```

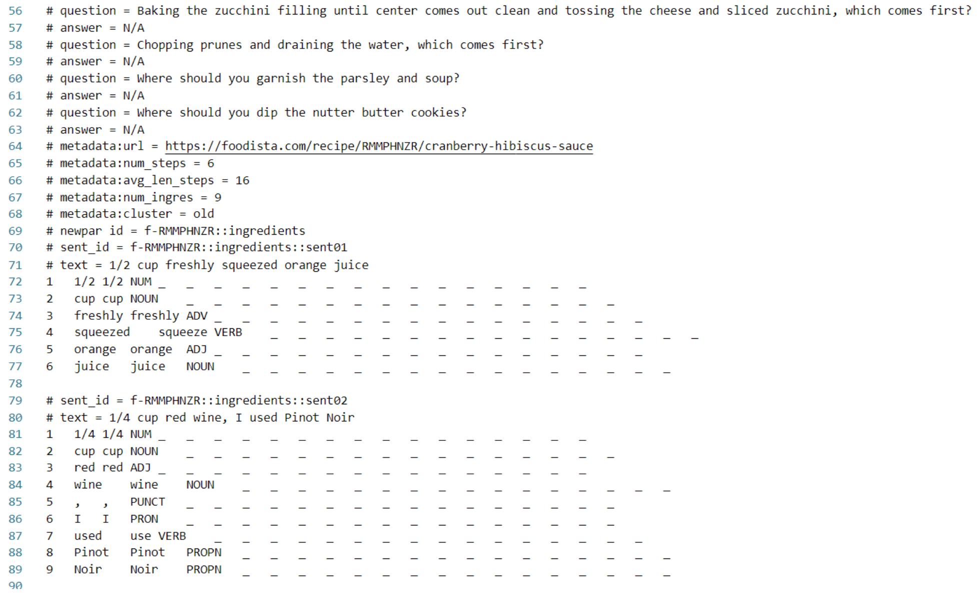
```
[12] query = 'How much temperature is required to heat oil for Chicken Spring Rolls?'
```

✓ 0s completed at 7:40 PM

5. DISCUSSION AND RELATED WORK

Initially, test data is prepared from the given two websites. Using the BeautifulSoup library in Python, the data is scraped from those websites, using this test data to understand the implementation process. The trained dataset provided in the CodaLab site is in CoNLL-U format where every word is represented in one line and every column represents an annotation. Later, we are parsing the CoNLL-U(Conference on

Computational Natural Language Learning) annotation schema dataset and transforming it to JSON (Javascript Object Notation) format. After that, we are dividing the dataset into training and testing datasets, and we train our model using a multimodal training dataset. The same process is performed on textual datasets and visual datasets.



```
# question = Baking the zucchini filling until center comes out clean and tossing the cheese and sliced zucchini, which comes first?
# answer = N/A
# question = Chopping prunes and draining the water, which comes first?
# answer = N/A
# question = Where should you garnish the parsley and soup?
# answer = N/A
# question = Where should you dip the nutter butter cookies?
# answer = N/A
# metadata:url = https://foodista.com/recipe/RMMPHNZR/cranberry-hibiscus-sauce
# metadata:num_steps = 6
# metadata:avg_len_steps = 16
# metadata:num_ingres = 9
# metadata:cluster = old
# newpar id = f-RMMPHNZR::ingredients
# sent_id = f-RMMPHNZR::ingredients::sent01
# text = 1/2 cup freshly squeezed orange juice
1   1/2 1/2 NUM _   _   _   _   _   _   _   _   _   _   _   _   _   _   _   _
2   cup cup NOUN    _   _   _   _   _   _   _   _   _   _   _   _   _   _   _   _
3   freshly freshly ADV _   _   _   _   _   _   _   _   _   _   _   _   _   _   _   _
4   squeezed    squeeze VERB    _   _   _   _   _   _   _   _   _   _   _   _   _   _   _   _
5   orange  orange  ADJ _   _   _   _   _   _   _   _   _   _   _   _   _   _   _   _
6   juice   juice   NOUN    _   _   _   _   _   _   _   _   _   _   _   _   _   _   _   _

# sent_id = f-RMMPHNZR::ingredients::sent02
# text = 1/4 cup red wine, I used Pinot Noir
1   1/4 1/4 NUM _   _   _   _   _   _   _   _   _   _   _   _   _   _   _   _
2   cup cup NOUN    _   _   _   _   _   _   _   _   _   _   _   _   _   _   _   _
3   red red ADJ _   _   _   _   _   _   _   _   _   _   _   _   _   _   _   _
4   wine    wine    NOUN    _   _   _   _   _   _   _   _   _   _   _   _   _   _   _   _
5   ,   ,   PUNCT   _   _   _   _   _   _   _   _   _   _   _   _   _   _   _   _
6   I   I   PRON    _   _   _   _   _   _   _   _   _   _   _   _   _   _   _   _
7   used    use VERB    _   _   _   _   _   _   _   _   _   _   _   _   _   _   _   _
8   Pinot   Pinot   PROPN   _   _   _   _   _   _   _   _   _   _   _   _   _   _   _   _
9   Noir    Noir    PROPN   _   _   _   _   _   _   _   _   _   _   _   _   _   _   _   _
```

6. PROBLEM :

We are trying to convert the given training data from conLLu to JSON format using pyjson module but not able to make much progress.

We are now trying to implement alternate methods(like strip newline, split on space char and make components for the dictionary) to see if it works

7. THE OBJECTIVE FOR THE NEXT WEEK IS :

Next week's objective is to complete textual extraction and to make progress on the Visual extraction part.

Summary of our work till now :

	Arihant Bhat	Vishal	Kishore	Jyothi	Neha
WEEK1 (28 September 2021— 5th October 2021)	working on the documentation. Visual Data: Worked on the	Working on Conllu annotation schema of reference data	Worked on creating the datasets from the given websites using	Analyzed problem statement and dataset	Working on textual extraction using BERT and giving my

	implementation and testing process of visual extraction of data using Video detection with ImageAI and YOLOv3 and preparing reports	to extract the context and use it in the BERT model for question answering.	the BeautifulSoup library in Python.		contribution towards testing the convolutional neural network
WEEK2 (6 th October – 13th October 2021)	Report preparation and documentation. Contributed to the implementation and testing of visual data extraction using ImageAI and YOLOv3.	Working on Conllu annotation schema of reference data to extract the context and use it in the BERT model for question answering. (Tried using stanza but unable to progress last week to extract the context, next week I will explore and try to do it in another way).	Working on implementing the Textual process using the scraped data from the websites.	Exploring and testing on Bert Hugging face and transformers for textual extraction.	Testing and implementing the data sets required for the BERT model which will be helpful for implementing the project.
WEEK3 (14th October – 19th October 2021)	Contributed to the implementation and testing of visual data extraction using ImageAI and YOLOv3, as well as report preparation	Progressing on textual extraction from the provided training data.	Analyzing the Visual data provided by codalab.	Conversion of CONLLU format to JSON and train the data using JSON	Testing and developing the data sets necessary for the BERT model, which will be useful for project implementation

	and documentation.				
WEEK4 20th October -26th October)	Analyzing on Visual extraction of data to identify the objects in the image using ImageAI.	Analyzing on Visual extraction of data to identify the objects in the image using ImageAI.	Working on implementing the process on visual datasets.	Performing experiments on ConLLu to JSON format conversion using module pyjsonNLP to train the data	Preparing the documentation part.
WEEK5					

8. CONCLUSION :

Our progress/code is available at :

https://colab.research.google.com/drive/1Ekds0ZN2_kLvqKkw9u-B8Xra77ndg6wt#scrollTo=QvAF3dIYhjx

https://colab.research.google.com/drive/1MUxfQj_-nmkCnQ4o5IamI0cqz8azXCXc

https://colab.research.google.com/drive/1TnvHtiIxVahVmK3_668rCCBFm7QOQ-tC#scrollTo=g2RPxn2Gl9f2

9. REFERENCES :

- https://www.analyticsvidhya.com/blog/2020/02/learn-image-classification-cnn-convolutional-neural-networks-3-datasets/
- https://universaldependencies.org/format.html
- https://medium.com/hucvl-stories/introducing-recipeqa-a-challenge-dataset-for-multimodal-comprehension-of-cooking-recipes-478c46e6a80c
- https://arxiv.org/pdf/2105.05999v1.pdf
- https://github.com/cdqa-suite/cdQA
- https://docs.google.com/presentation/d/1mvuu4QTfOP6CHUfbLdXMisleiH7kIlxpsU1bgyT0oTw/edit#slide=id.g1e8a3a7b60d4c1c3_27