Question Answering on SQUAD Dataset

```
!pip install html2text --quiet
!pip install simpletransformers --quiet
```

```
     |████████████████████████████████| 204kB 8.1MB/s
     |████████████████████████████████| 2.9MB 15.4MB/s
     |████████████████████████████████| 7.4MB 45.1MB/s
     |████████████████████████████████| 317kB 59.8MB/s
     |████████████████████████████████| 1.5MB 59.4MB/s
     |████████████████████████████████| 1.8MB 59.3MB/s
     |████████████████████████████████| 51kB 8.6MB/s
     |████████████████████████████████| 71kB 9.8MB/s
     |████████████████████████████████| 1.1MB 56.4MB/s
     |████████████████████████████████| 4.5MB 55.2MB/s
     |████████████████████████████████| 163kB 66.5MB/s
     |████████████████████████████████| 112kB 52.9MB/s
     |████████████████████████████████| 81kB 12.5MB/s
     |████████████████████████████████| 890kB 50.7MB/s
     |████████████████████████████████| 133kB 56.0MB/s
     |████████████████████████████████| 102kB 14.9MB/s
     |████████████████████████████████| 122kB 62.0MB/s
     |████████████████████████████████| 71kB 11.6MB/s
  Building wheel for seqeval (setup.py) ... done
  Building wheel for blinker (setup.py) ... done
  Building wheel for sacremoses (setup.py) ... done
  Building wheel for subprocess32 (setup.py) ... done
ERROR: google-colab 1.0.0 has requirement ipykernel~=4.10, but you'll have ipykernel 5.4.2 which is incompatible.
```

```
!pip install -U ipykernel
!pip install modin[dask]
```

```
import numpy as np # Math
import requests # Getting text from websites
import html2text # Converting wiki pages to plain text
from googlesearch import search # To performing Google searches
```

```
import re
from simpletransformers.question_answering import QuestionAnsweringModel
from IPython.display import display
from IPython.html import widgets # Graphical display
from bs4 import BeautifulSoup
from markdown import markdown
```

```
/usr/local/lib/python3.6/dist-packages/IPython/html.py:14: ShimWarning: The `IPython.html` package has been deprecated
  "`IPython.html.widgets` has moved to `ipywidgets`.", ShimWarning)
```

```
model = QuestionAnsweringModel('distilbert', 'distilbert-base-uncased-distilled-squad')
```

Downloading: 100% 451/451 [00:00<00:00, 13.0kB/s]

Downloading: 100% 265M/265M [00:03<00:00, 81.1MB/s]

Downloading: 100% 232k/232k [00:00<00:00, 850kB/s]

```
# Testing for a single question

question_data = {
  'qas':
  [{'question': 'What color is the sky?',
    'id': 0,
    'answers': [{'text': ' ', 'answer_start': 0}],
    'is_impossible': False}],
  'context': 'the sky is blue'
  }

prediction = model.predict([question_data])
print(prediction)
```

```
convert squad examples to features: 100%|██████████| 1/1 [00:00<00:00, 196.46it/s]
```

```
# Dividing the context into small context of size 512

def predict_answer(model, question, contexts, seq_len=512, debug=False):
  split_context=[]
  if not isinstance(contexts, list):
    contexts=[]

  for context in contexts:
    for i in range(0, len(context), seq_len):
      split_context.append(context[i:i+seq_len])
  split_context= contexts

  f_data=[]
  for i,c in enumerate(split_context):
    f_data.append(
        {
        'qas':
    [{'question': question,
      'id': i,
      'answers':[{'text': ' ', 'answer_start':0}],
      'is_impossible':False}], # for unanswerable questions
    'context': c
        })
    prediction = model.predict(f_data)
    ans= prediction[0][0]['answer'][0]
    prob= prediction[1][0]['probability'][0]
    print("Answer: ",ans,", Probability: ",prob)
```

```
predict_answer(model, 'what colour is sky?', ['the sky is blue in colour'])
```

```
convert squad examples to features: 100%|██████████| 1/1 [00:00<00:00, 253.37it/s]
add example index and unique id: 100%|██████████| 1/1 [00:00<00:00, 12633.45it/s]
Running Prediction: 100%                                                    1/1 [00:00<00:00, 20.88it/s]
Answer:  blue , Probability:  0.9413280059545852
```

```
predict_answer(model, 'which is the largest animal?', ['Although elephants are quite big but the blue whale is the largest a
```

```
convert squad examples to features: 100%|██████████| 1/1 [00:00<00:00, 7269.16it/s]
add example index and unique id: 100%|██████████| 1/1 [00:00<00:00, 6775.94it/s]
```

Running Prediction: 100% 1/1 [00:00<00:00, 21.32it/s]

```
Answer:  blue whale , Probability:  0.7292108232300563
```

```
# Pre-Processing  text
links = list(search('what colour is the sky?', stop=2))

html_conv= html2text.HTML2Text()
html_conv.ignore_links= True
html_conv.escape_all= True

text=[]
for l in links:
  req= requests.get(l)
  text.append(html_conv.handle(req.text))
```

```
text
```

```
['PNG \x1a \x00\x00\x00 IHDR\x00\x00\x00\x00\x00\x00\x00\x08\x03\x00\x00\x00bH\x01wPLTE\n\x0
 '\n\n# Chicago Sky\n\nFrom Wikipedia, the free encyclopedia\n\nJump to navigation Jump to search\n\nChicago Sky  \n---
```

```
# To remove extra html tags
def markdown_to_text(markdown_string):
    """ Converts a markdown string to plaintext """
    html = markdown(markdown_string)
    html = re.sub(r'<pre>(.*?)</pre>', ' ', html)
    html = re.sub(r'<code>(.*?)</code >', ' ', html)

    # extract text
    soup = BeautifulSoup(html, "html.parser")
    text = ''.join(soup.findAll(text=True))
    return text
```

```
def format_text(text):
    text = markdown_to_text(text)
    text = text.replace('\n', ' ')
    return text
```

```
links = list(search('what color is the sky?', stop=2)) # stop represents number of links to consider
print(links)

html_conv = html2text.HTML2Text()
html_conv.ignore_links = True
html_conv.escape_all = True

text = []
for link in links:
    req = requests.get(link)
    text.append(html_conv.handle(req.text))
    text[-1] = format_text(text[-1])

print(text)
```

```
['https://www.universetoday.com/74020/what-color-is-the-sky/', 'https://science.discoveryplace.org/blog/ever-wonder-why
['503 Service Temporarily Unavailable  nginx', "Skip to main Content ADVANCED RESERVATION REQUIRED MAKE A RESERVATION A
```

```
def query_pages(query, n=5):
    return list(search(query, stop=n))

query_pages('life of pi story')
```

```
['https://en.wikipedia.org/wiki/Life_of_Pi',
 'https://en.wikipedia.org/wiki/Life_of_Pi#Plot',
 'https://en.wikipedia.org/wiki/Life_of_Pi#Inspiration',
 'https://en.wikipedia.org/wiki/Life_of_Pi#Characters',
 'https://en.wikipedia.org/wiki/Life_of_Pi#Reception']
```

```
def query_to_text(query, n=5):
    html_conv = html2text.HTML2Text()
```

```
    html_conv.ignore_links = True
    html_conv.escape_all = True

    text = []
    for link in query_pages(query, n):
        req = requests.get(link)
        text.append(html_conv.handle(req.text))
        text[-1] = format_text(text[-1])
    return text
```

```
question = 'where is Taj Mahal?'
context = query_to_text(question, n=1)
pred = predict_answer(model, question, context)
print(pred)
```

```
convert squad examples to features: 100%|██████████| 1/1 [00:00<00:00, 43.57it/s]
add example index and unique id: 100%|██████████| 1/1 [00:00<00:00, 5645.09it/s]
Running Prediction: 100%                                    1/1 [00:00<00:00, 18.67it/s]
Answer:   , Probability:  0.7531785793195008
None
```

```
def q_to_a(model, question, n=2):
    context = query_to_text(question, n=n)
    pred = predict_answer(model, question, context)
    return pred
```

```
q_to_a(model, 'what color is the sky')
```

```
convert squad examples to features: 100%|██████████| 1/1 [00:00<00:00, 217.56it/s]
```

GUI Output

```
Answer:  503 Service Temporarily Unavailable nginx , Probability:  0.5240630634217339
```

```
text = widgets.Text(description='Question:', width=300)
display(text)

button = widgets.Button(description='Get an Answer')
display(button)

def on_button_click(b):
    answer = q_to_a(model, text.value, n=2)
    print('Answer:', answer)

button.on_click(on_button_click)
```

Question: covid 19 symptoms

Get an Answer

```
convert squad examples to features: 100%|██████████| 1/1 [00:00<00:00, 155.32it/s]
add example index and unique id: 100%|██████████| 1/1 [00:00<00:00, 9845.78it/s]
```

Running Prediction: 100% 1/1 [00:00<00:00, 24.50it/s]

```
Answer:  18.24d5c217.1609946740.20229d , Probability:  0.18799871132166562
convert squad examples to features: 100%|██████████| 2/2 [00:01<00:00,  1.56it/s]
add example index and unique id: 100%|██████████| 2/2 [00:00<00:00, 13774.40it/s]
```

Running Prediction: 100% 8/8 [00:00<00:00, 34.40it/s]

```
Answer:  18.24d5c217.1609946740.20229d , Probability:  0.18775240077788657
Answer: None
convert squad examples to features: 100%|██████████| 1/1 [00:00<00:00, 13.49it/s]
add example index and unique id: 100%|██████████| 1/1 [00:00<00:00, 4760.84it/s]
```

Running Prediction: 100% 2/2 [00:00<00:00, 29.82it/s]

```
Answer:  ###RX#####R##2R[#J##砾###м##s##)ю##Q]#####k6##kE}#O#x#r#(#W#?#u##G##NL#######&#m##y#####3#(##M#
convert squad examples to features: 100%|██████████| 2/2 [00:00<00:00,  7.63it/s]
add example index and unique id: 100%|██████████| 2/2 [00:00<00:00, 19553.86it/s]
```

Running Prediction: 100% 3/3 [00:00<00:00, 34.77it/s]

```
Answer:  ###RX#####R##2R[#J##砾###м##s##)ю##Q]#####k6##kE}#O#x#r#(#W#?#u##G##NL#######&#m##y#####3#(##M#
Answer: None
convert squad examples to features: 100%|██████████| 1/1 [00:00<00:00,  4.24it/s]
add example index and unique id: 100%|██████████| 1/1 [00:00<00:00, 9238.56it/s]
```

Running Prediction: 100% 2/2 [00:00<00:00, 28.60it/s]

```
Answer:  a boy , Probability:  0.22889368385201297
convert squad examples to features: 100%|██████████| 2/2 [00:00<00:00,  4.32it/s]
add example index and unique id: 100%|██████████| 2/2 [00:00<00:00, 17512.75it/s]
```

Running Prediction: 100% 4/4 [00:00<00:00, 36.24it/s]

```
Answer:  a boy , Probability:  0.22889368385201297
Answer: None
convert squad examples to features: 100%|██████████| 1/1 [00:00<00:00,  4.28it/s]
add example index and unique id: 100%|██████████| 1/1 [00:00<00:00, 9986.44it/s]
```

Running Prediction: 100% 2/2 [00:00<00:00, 27.68it/s]

```
Answer:  "https://en.wikipedia.org/w/index.php?title=Life_of_Pi&oldid=995976305" , Probability:  0.34457417209953944
convert squad examples to features: 100%|██████████| 2/2 [00:00<00:00,  4.33it/s]
add example index and unique id: 100%|██████████| 2/2 [00:00<00:00, 17296.10it/s]
```

```
Running Prediction: 100%                                                    4/4 [00:00<00:00, 34.37it/s]

Answer:  "https://en.wikipedia.org/w/index.php?title=Life_of_Pi&oldid=995976305" , Probability:  0.34457417209953944
Answer: None
convert squad examples to features: 100%|██████████| 1/1 [00:00<00:00, 12.15it/s]
add example index and unique id: 100%|██████████| 1/1 [00:00<00:00, 11491.24it/s]

Running Prediction: 100%                                                    1/1 [00:00<00:00, 18.76it/s]

Answer:  magma within the Earth and lava , Probability:  0.1861369956488139
convert squad examples to features: 100%|██████████| 2/2 [00:00<00:00, 19.83it/s]
add example index and unique id: 100%|██████████| 2/2 [00:00<00:00, 19108.45it/s]

Running Prediction: 100%                                                    1/1 [00:00<00:00, 21.76it/s]
```

```
convert squad examples to features: 100%|██████████| 1/1 [00:00<00:00,  5.95it/s]
add example index and unique id: 100%|██████████| 1/1 [00:00<00:00, 9776.93it/s]

Running Prediction: 100%                                                    2/2 [00:00<00:00, 30.50it/s]

Answer:  difference , Probability:  0.22545115362312046
convert squad examples to features: 100%|██████████| 2/2 [00:00<00:00,  3.02it/s]
add example index and unique id: 100%|██████████| 2/2 [00:00<00:00, 16448.25it/s]

Running Prediction: 100%                                                    6/6 [00:00<00:00, 33.47it/s]

Answer:  difference , Probability:  0.22561775835923645
Answer: None
convert squad examples to features: 100%|██████████| 1/1 [00:00<00:00, 11.44it/s]
add example index and unique id: 100%|██████████| 1/1 [00:00<00:00, 5907.47it/s]

Running Prediction: 100%                                                    1/1 [00:00<00:00, 20.21it/s]

Answer:  mild symptoms to severe illness , Probability:  0.30260054380744483
convert squad examples to features: 100%|██████████| 2/2 [00:00<00:00,  9.35it/s]
add example index and unique id: 100%|██████████| 2/2 [00:00<00:00, 15917.66it/s]

Running Prediction: 100%                                                    2/2 [00:00<00:00, 31.36it/s]

Answer:  mild symptoms to severe illness , Probability:  0.302774932449822
Answer: None
```