Basic Configuration

In [1]:

```
# import packages
import json
from pathlib import Path
from scripts import dsit_workshop as dsit

import pandas as pd

# set path variables
ROOT = Path("/home/azureuser/").resolve()
DATA_RAW = ROOT / "data-raw"
DATA_PROCESSED = ROOT / "data-processed"
```

ChatGPT Warm-Up

1. How does ChatGPT work?

It basically uses input text to generate (predict) the next word in the sequence of words fed to the model. The model is trained on a large corpora of text data and predicts the next word iteartively until it's done completing the input text (either because it hit a stop sequence or the predicted probability of the next word is below a certain threshold).

In [2]:

```
question_a = "Who won the most FIFA Women's Soccer World Cup?"
answer_a = dsit.ask_chatGPT_openai(question_a)
print(answer_a)
```

```
The United States women's national soccer team has won the most FIFA Women's W
orld Cup titles, with four championships in 1991, 1999, 2015, and 2019.
```

Note that the answer above was built based on GPT's memory. It had seen similar text before and inferred the relationship of words to generate the answer and provide it back to us (the users).

We can however trick the model by asking a question with less context so that it doesn't get the answer right.

In [3]:

```
question_b = "Who played the final in 2019?"
answer_b = dsit.ask_chatGPT_openai(question_b)
print(answer_b)
```

```
The final of what event are you referring to?
```

2. Chat with language model

To show how the model uses the input sequences (context), notice that we can make it answer `question_b` correctly by prepending `question_a` to the instruction. If we do this iteratively, we construct a chat bot by allowing the model to use previous questions and answers to answer new questions.

In [4]:

```
history = f"""
    User: {question_a}

    System: {answer_a}

    User: "Who played the final in 2019?"
"""

print(history)
```

```
    User: Who won the most FIFA Women's Soccer World Cup?

    System: The United States women's national soccer team has won the most FI
FA Women's World Cup titles, with four championships in 1991, 1999, 2015, and
2019.

    User: "Who played the final in 2019?"
```

Let's now see how the model can correctly answer the question we asked earlier.

In [5]:

```
print(dsit.ask_chatGPT_openai(history))
```

```
System: The final of the 2019 FIFA Women's World Cup was played between the Un
ited States and the Netherlands. The United States won the match 2-0 to secure
their fourth World Cup championship.
```

How can we make it more useful?

There's, however, a list of limitations when using large language models the way we've used so far:

1. It is only a chatbot (and it might get things wrong).
2. It returns long-form, long text.
3. It might leak information to the internet.
4. It only returns one answer at a time.
5. It is not integrated with anything else.

Let's address them one by one.

Limitation 01: It is only a chatbot

Let's use it for a different task, i.e., data extraction. Below we have a paragraph about... well... me. Let's see if GPT can help us extract my information, structure it, and return it in a format we can use.

In [6]:

```
document = """
    My name is Andre Assumpcao and I am a data scientist at the National
    Center for State Courts. I moved to the United States in 2015 to
    pursue my Ph.D. at the University of North Carolina at Chapel Hill.

    Please extract the name, occupation, employer, and education level
    of the person in the paragraph above.
```

```
"""

data = dsit.ask_chatGPT_openai(document)
print(data)
```

```
Name: Andre Assumpcao
Occupation: Data scientist
Employer: National Center for State Courts
Education Level: Ph.D.
```

Limitation 02: It returns long-form, written text

However, this is not good enough. We still need to parse out the information from the long-form text above. We would need to split it by each of the keys (name, occupation, employer, education level) and attach the values of each key. GPT, however, is smart enough to do this for us... **as long as we ask for it.** Let's see how that would work.

In [7]:

```
document += " Please return the response as a JSON object - nothing else neede
print(document)
```

```
    My name is Andre Assumpcao and I am a data scientist at the National
    Center for State Courts. I moved to the United States in 2015 to
    pursue my Ph.D. at the University of North Carolina at Chapel Hill.

    Please extract the name, occupation, employer, and education level
    of the person in the paragraph above.
 Please return the response as a JSON object - nothing else needed.
```

Now we ask GPT to process the entire prompt above!

In [8]:

```
data = dsit.ask_chatGPT_openai(document)
print(data)
```

```
{
  "name": "Andre Assumpcao",
  "occupation": "data scientist",
  "employer": "National Center for State Courts",
  "education_level": "Ph.D."
}
```

The return is now a JSON object! It'll require some reformatting but much easier than anything we would have to do using the long-form, written out text.

In [9]:

```
data = data.strip().replace("\n", "").replace("\t", "")
df = pd.DataFrame.from_records(json.loads(data), index=[1])
df
```

Out[9]:

	education_level	employer	name	occupation
1	Ph.D.	National Center for State Courts	Andre Assumpcao	data scientist

In [10]:

```
df.to_csv("data-processed/aassumpcao.csv", index=False, quoting=2)
```

Limitation 03: It might leak information to the internet

Another limitation of using ChatGPT is that you would be under the terms of service of OpenAI, and it's possible that these won't meet your security requirements for processing sensitive data. Fortunately, you can use GPT on a dedicated, exclusive server on the Microsoft Azure cloud. You and you alone have access to Azure's REST API endpoints and you can make calls that never leave your network environment.

In [11]:

```
data = dsit.ask_chatGPT_azure(document)
print(data)
```

```
{
  "name": "Andre Assumpcao",
  "occupation": "data scientist",
  "employer": "National Center for State Courts",
  "education_level": "Ph.D."
}
```

Limitation 04: It only returns one answer at a time

If we are working with data, we probably want to process a large number of documents or observations and have GPT structure all documents. When we use the API, we can either make a single call with multiple document and one prompt or multiple calls each with a different prompt. Let's see how that would work below.

In [14]:

```
document = """
    My name is Andre Assumpcao and I am a data scientist at the National
    Center for State Courts. I moved to the United States in 2015 to
    pursue my Ph.D. at the University of North Carolina at Chapel Hill.

    Diane Robinson, Ph.D., is a principal court research
    associate at the National Center for State Courts.

    Please extract the name, occupation, employer, and education level
    of the people in the paragraph above. Return only a JSON object with
    the results. Each person should be one element of a list in each key.
"""

data = dsit.ask_chatGPT_azure(document)
data = data.strip().replace("\n", "").replace("\t", "")
data = json.loads(data)
data
```

Out[14]:

```
{'name': ['Andre Assumpcao', 'Diane Robinson'],
 'occupation': ['data scientist', 'principal court research associate'],
 'employer': ['National Center for State Courts',
  'National Center for State Courts'],
 'education_level': ['Ph.D.', 'Ph.D.']}
```

In [15]:

```
df = pd.DataFrame.from_dict(data, orient='columns')
df
```

Out[15]:

	name	occupation	employer	education_level
0	Andre Assumpcao	data scientist	National Center for State Courts	Ph.D.
1	Diane Robinson	principal court research associate	National Center for State Courts	Ph.D.

In [16]:

```
df.to_csv("data-processed/ncscpeople.csv", index=False, quoting=2)
```

Limitation 05: It is not integrated with anything else

Once we know how to work with the GPT API, it is actually quite easy to build a system that allows us to process court documents and store their information in a structured, tabular format. How about we do this?

Steps:

1. Read in the court documents and OCR them when needed. We will use the Azure Form Recognizer API service for this.
2. Then pass the text from each of these documents through GPT to extract the relevant information.
3. Save the info as a csv file.

Step 01: Read In Files

We have saved three notice of removal cases from CA in the `data-raw` folder. You can feel free to open them up if you would like.

We have prepared a function to extract the text from the PDF files, which is what we do below.

In [17]:

```
# list files in data-raw
files = [file for file in DATA_RAW.iterdir()]
files = sorted(files)

# ocr case files
case_files = dsit.ocr_document(files[1])
```

Let's print out the content of the first file to check whether the OCR function worked.

In [18]:

```
print(case_files)
```

```
Case 4:21-cv-08580
Document 1
Filed 11/03/21
Page 1 of 10
1
MATTHEW J. BLASCHKE (State Bar No. 281938)
mblaschke@kslaw.com
KING & SPALDING LLP
50 California Street Suite 3300
San Francisco, CA 94111
Telephone:
(415) 318-1212
Facsimile:
(415) 318-1300
2
3
4
5 DAVID L. BALSER (pro hac vice forthcoming) dbalser@kslaw.com
6 8 9 S. STEWART HASKINS II (pro hac vice forthcoming) shaskins@kslaw.com 7 KI
NG & SPALDING LLP 1180 Peachtree Street, N.E. Atlanta, GA 30309 Telephone.: (4
04) 572-4600 Facsimile: (404) 572-5100
10 Counsel for Capital One Defendants
11
UNITED STATES DISTRICT COURT
12
FOR THE NORTHERN DISTRICT OF CALIFORNIA OAKLAND DIVISION
13
14 15 RONEY MIRANDA and ALAIN MICHAEL, on behalf of themselves and all others
similarly situated,
Case No.
CAPITAL ONE DEFENDANTS’
NOTICE OF REMOVAL
16
17
Plaintiffs,
18
v.
19
CAPITAL ONE FINANCIAL CORPORATION, CAPITAL ONE BANK (USA), N.A., CAPITAL ONE,
N.A., AMAZON.COM, Inc., and AMAZON WEB SERVICES, Inc.
20
21
22
Defendants.
23
24
25
26
27
28
NOTICE OF REMOVAL
Case 4:21-cv-08580
Document 1 Filed 11/03/21
Page 2 of 10
1 Defendants Capital One Financial Corporation, Capital One Bank (USA), N.A.,
and Capital
2 3 5 6 7 One, N.A. (collectively, “Capital One”) hereby remove this purported
class action from the Superior Court of the State of California for the County
of Alameda to the United States District Court for the Northern District of Ca
lifornia, Oakland Division. This Notice of Removal is filed pursuant to 28 U.
```

```
S.C. §§ 1332, 1441, 1446, and 1453, with the consent of Defendants Amazon.com,
Inc. (“Amazon”) and Amazon Web Services, Inc. (“AWS”), and on the basis of the
following facts, which show that this case may properly be removed to this Cou
rt:
4
8
BACKGROUND AND PLAINTIFFS’ COMPLAINT
9
1. Plaintiffs Roney Miranda and Alain Michael filed this case on September 17,
2021, in
10 11 12 13 the Superior Court of California for the County of Alameda, Case N
umber RG21113096, styled as Roney Miranda and Alain Michael, on behalf of them
selves and all others similarly situated v. Capital One Financial Corporation,
Capital One Bank (USA), N.A., Capital One, N.A., Amazon.com, Inc., and Amazon
Web Services, Inc. (the “State Court Action”).
14
2. The Complaint filed in the State Court Action is attached as Exhibit A (the
“Complaint” or “Compl.”). Capital One was served with the Complaint on October
22, 2021, making this removal timely.
15
16
17
3. The Complaint asserts the following claims: (1) negligence, (2) negligence
per se, (3) unjust enrichment, (4) declaratory judgment and injunctive relief,
(5) breach of confidence, (6) breach of contract, (7) breach of implied contra
ct, (8) violation of California’s Unfair Competition Law (Cal. Bus. & Prof. Co
de §§ 17200, et seq.), and (9) violation of California’s Consumer Legal Remedi
es Act (Cal. Civ. Code §§ 1750, et seq.). See Compl. ¶¶ 152-244
18
19
20
21
22
4. The foregoing claims all arise from the cyber-attack Capital One announced
on July 29, 2019 (the “Cyber Incident”), which affected nearly 100 million U.
S. consumers. See id. ¶¶ 1-2.
23
24
5. As set forth in the Complaint, the Cyber Incident was perpetrated by a hack
er named Paige Thompson. Id. ¶¶ 3, 55. Plaintiffs allege that Thompson exfiltr
ated personal information pertaining to consumers who applied for Capital One
credit card products between 2005 and 2019. Id. ¶¶ 3-5, 79.
25
26
27
28
2
NOTICE OF REMOVAL
```

Let's build a simple dataset containing metadata about these cases. We would want to extract the following information:

1. The name of the court in which the case was filed.
2. The name (or names) of those who filed the case.
3. The name (or names) of those against whom the case was filed.
4. The type of case filed.
5. The U.S. Code basis for the federal court to take up the case.
6. The contact details of plaintiffs' lawyers.

7. The contact details of defendants' lawyers.

Step 02: Send the text to GPT

How do we extract these data points? We write out a prompt with instructions to do so and return a structured JSON object that we can then use to build a tabular dataset.

In [19]:

```
prompt = """
    Please extract the following information from the document above. Return
    a JSON object with keys and values as listed below. If any of information

    1. "court": the name of the court in which the case was filed.
    2. "plaintiffs": the name (or names) of those who filed the case. Multiple
    3. "defendants": the name (or names) of those against whom the case was fi
    4. "case_type": the type of case filed.
    5. "us_code": the U.S. Code basis for the court to take up the case.
    6. "plaintiff_lawyer_info": the contact details of the plaintiffs' lawyer.
    7. "defendant_lawyer_info": the contact details of the defendants' lawyer.
"""
```

How would the full prompt look? This is what we'll send over to GPT.

In [20]:

```
message = f"Begin document:{case_files}\nEnd of document.\n\n{prompt}"
print(message)
```

```
Begin document:Case 4:21-cv-08580
Document 1
Filed 11/03/21
Page 1 of 10
1
MATTHEW J. BLASCHKE (State Bar No. 281938)
mblaschke@kslaw.com
KING & SPALDING LLP
50 California Street Suite 3300
San Francisco, CA 94111
Telephone:
(415) 318-1212
Facsimile:
(415) 318-1300
2
3
4
5 DAVID L. BALSER (pro hac vice forthcoming) dbalser@kslaw.com
6 8 9 S. STEWART HASKINS II (pro hac vice forthcoming) shaskins@kslaw.com 7 KI
NG & SPALDING LLP 1180 Peachtree Street, N.E. Atlanta, GA 30309 Telephone.: (4
04) 572-4600 Facsimile: (404) 572-5100
10 Counsel for Capital One Defendants
11
UNITED STATES DISTRICT COURT
12
FOR THE NORTHERN DISTRICT OF CALIFORNIA OAKLAND DIVISION
13
14 15 RONEY MIRANDA and ALAIN MICHAEL, on behalf of themselves and all others
similarly situated,
Case No.
CAPITAL ONE DEFENDANTS’
NOTICE OF REMOVAL
16
17
Plaintiffs,
18
v.
19
CAPITAL ONE FINANCIAL CORPORATION, CAPITAL ONE BANK (USA), N.A., CAPITAL ONE,
N.A., AMAZON.COM, Inc., and AMAZON WEB SERVICES, Inc.
20
21
22
Defendants.
23
24
25
26
27
28
NOTICE OF REMOVAL
Case 4:21-cv-08580
Document 1 Filed 11/03/21
Page 2 of 10
1 Defendants Capital One Financial Corporation, Capital One Bank (USA), N.A.,
and Capital
2 3 5 6 7 One, N.A. (collectively, “Capital One”) hereby remove this purported
class action from the Superior Court of the State of California for the County
of Alameda to the United States District Court for the Northern District of Ca
lifornia, Oakland Division. This Notice of Removal is filed pursuant to 28 U.
```

```
S.C. §§ 1332, 1441, 1446, and 1453, with the consent of Defendants Amazon.com,
Inc. (“Amazon”) and Amazon Web Services, Inc. (“AWS”), and on the basis of the
following facts, which show that this case may properly be removed to this Cou
rt:
4
8
BACKGROUND AND PLAINTIFFS’ COMPLAINT
9
1. Plaintiffs Roney Miranda and Alain Michael filed this case on September 17,
2021, in
10 11 12 13 the Superior Court of California for the County of Alameda, Case N
umber RG21113096, styled as Roney Miranda and Alain Michael, on behalf of them
selves and all others similarly situated v. Capital One Financial Corporation,
Capital One Bank (USA), N.A., Capital One, N.A., Amazon.com, Inc., and Amazon
Web Services, Inc. (the “State Court Action”).
14
2. The Complaint filed in the State Court Action is attached as Exhibit A (the
“Complaint” or “Compl.”). Capital One was served with the Complaint on October
22, 2021, making this removal timely.
15
16
17
3. The Complaint asserts the following claims: (1) negligence, (2) negligence
per se, (3) unjust enrichment, (4) declaratory judgment and injunctive relief,
(5) breach of confidence, (6) breach of contract, (7) breach of implied contra
ct, (8) violation of California’s Unfair Competition Law (Cal. Bus. & Prof. Co
de §§ 17200, et seq.), and (9) violation of California’s Consumer Legal Remedi
es Act (Cal. Civ. Code §§ 1750, et seq.). See Compl. ¶¶ 152-244
18
19
20
21
22
4. The foregoing claims all arise from the cyber-attack Capital One announced
on July 29, 2019 (the “Cyber Incident”), which affected nearly 100 million U.
S. consumers. See id. ¶¶ 1-2.
23
24
5. As set forth in the Complaint, the Cyber Incident was perpetrated by a hack
er named Paige Thompson. Id. ¶¶ 3, 55. Plaintiffs allege that Thompson exfiltr
ated personal information pertaining to consumers who applied for Capital One
credit card products between 2005 and 2019. Id. ¶¶ 3-5, 79.
25
26
27
28
2
NOTICE OF REMOVAL
End of document.


    Please extract the following information from the document above. Return
    a JSON object with keys and values as listed below. If any of information
is not found, return a `null` value.

    1. "court": the name of the court in which the case was filed.
    2. "plaintiffs": the name (or names) of those who filed the case. Multiple
entries should be a single string.
    3. "defendants": the name (or names) of those against whom the case was fi
led. Multiple entries should be a single string.
```

```
    4. "case_type": the type of case filed.
    5. "us_code": the U.S. Code basis for the court to take up the case.
    6. "plaintiff_lawyer_info": the contact details of the plaintiffs' lawyer.
Multiple entries should be a single string.
    7. "defendant_lawyer_info": the contact details of the defendants' lawyer.
Multiple entries should be a single string.
```

What are the structured data points?

In [21]:

```
response = dsit.ask_chatGPT_azure(message)
json.loads(response)
```

Out[21]:

```
{'court': 'United States District Court for the Northern District of Californi
a, Oakland Division',
 'plaintiffs': 'Roney Miranda and Alain Michael, on behalf of themselves and a
ll others similarly situated',
 'defendants': 'Capital One Financial Corporation, Capital One Bank (USA), N.
A., Capital One, N.A., Amazon.com, Inc., and Amazon Web Services, Inc.',
 'case_type': 'purported class action',
 'us_code': '28 U.S.C. §§ 1332, 1441, 1446, and 1453',
 'plaintiff_lawyer_info': 'MATTHEW J. BLASCHKE (State Bar No. 281938)\nmblasch
ke@kslaw.com\nKING & SPALDING LLP\n50 California Street Suite 3300\nSan Franci
sco, CA 94111\nTelephone:\n(415) 318-1212\nFacsimile:\n(415) 318-1300',
 'defendant_lawyer_info': 'DAVID L. BALSER (pro hac vice forthcoming) dbalser@
kslaw.com\nS. STEWART HASKINS II (pro hac vice forthcoming) shaskins@kslaw.com
\nKING & SPALDING LLP\n1180 Peachtree Street, N.E.\nAtlanta, GA 30309\nTelepho
ne.: (404) 572-4600\nFacsimile: (404) 572-5100'}
```

Step 3: Save the data as a csv file

We can finally build a tabular dataset...

In [22]:

```
df = pd.json_normalize(json.loads(response))
df
```

Out[22]:

	court	plaintiffs	defendants	case_type	us_code	plaintiff_lawyer_info	defendant_lawyer
0	United States District Court for the Northern ...	Roney Miranda and Alain Michael, on behalf of ...	Capital One Financial Corporation, Capital One...	purported class action	28 U.S.C. §§ 1332, 1441, 1446, and 1453	MATTHEW J. BLASCHKE (State Bar No. 281938)\nmb...	DAVID L. BALSEF hac vice forthco

...and save it as a csv file.

In [23]:

```
df.to_csv("data-processed/onerecord_processed.csv", index=False, quoting=2)
```

Boom, we have a data extraction pipeline! Let's look at how the entire pipeline and execute it

In [25]:

```
def data_extraction_pipeline(filepaths):

    # read in and ocr case document
    case_files = [dsit.ocr_document(filepath) for filepath in filepaths]
    # create prompt for data extraction
    prompt = """
        Please extract the following information from the document above. Retu
        a JSON object with keys and values as listed below. If any of informat

        1. "court": the name of the court in which the case was filed.
        2. "plaintiffs": the name (or names) of those who filed the case. Mult
        3. "defendants": the name (or names) of those against whom the case wa
        4. "case_type": the type of case filed.
        5. "us_code": the U.S. Code basis for the court to take up the case.
        6. "plaintiff_lawyer_info": the contact details of the plaintiffs' law
        7. "defendant_lawyer_info": the contact details of the defendants' law
    """

    # join document and prompt
    messages = [f"Begin document:{case_file}\nEnd of document.\n\n{prompt}" fo

    # send information over to GPT
    data = [dsit.ask_chatGPT_azure(message) for message in messages]

    # store the info into a tabular format
    df = [pd.json_normalize(json.loads(dataset)) for dataset in data]
    df = pd.concat(df)

    # save to disk
    df.to_csv("data-processed/multiplerecords_processed.csv", index=False, quo

    # return to user
    return df


# execute pipeline
data_extraction_pipeline(files[1:])
```

Out[25]:

	court	plaintiffs	defendants	case_type	us_code	plaintiff_lawyer_info	defendant_
0	United States District Court for the Northern ...	RONEY MIRANDA and ALAIN MICHAEL, on behalf of ...	CAPITAL ONE FINANCIAL CORPORATION, CAPITAL ONE...	purported class action	28 U.S.C. §§ 1332, 1441, 1446, and 1453	MATTHEW J. BLASCHKE (State Bar No. 281938)\nmb...	DAVID L. hac vice
0	U.S. DISTRICT COURT CENTRAL DIST. OF CALIFORNIA	EDEN RUIZ ARI	ARCHON INFORMATION SYSTEMS, LLC, BRYAN P. BARR...	NOTICE OF REMOVAL	28 USC §§ 1332 and 1441	Joshua S. Force (State Bar No. 176383) SHER GA...	Attorney Informat
0	United States District Court for the Central D...	JP MORGAN CHASE BANK, N.A.	ROXZANA BOTAS HTTA ROXANA BOTASH; and DOES 1 to 6	Unlawful Detainer	28 U.S.C. § 1331, 1441(a), 1443(1)	Not available	In Pro P Defendar