Project Name: CSML1010 NLP Course Project - Part 1 - Proposal): Problem, Dataset, and Exploratory Data Analysis

Authors (Group3): Paul Doucet, Jerry Khidaroo

2. Data Clean-up and NLP Notebook

This notebook will review the Data Cleaning tasks performed as part of our project proposal:

- Categorize Groups
- Connect to Database
- Cleaning the Dataset for NLP
- NLP
- Store to Database

Categorize Groups

In [17]:

```python
import pandas as pd
```

In [18]:

```python
# Import CSV
df = pd.read_csv("./data/DF_selfDialogs.csv")
```

In [19]:

```python
print (df.groupby('Instruction_id').size())
```

```
Instruction_id
auto-repair-appt-1    1161
coffee-ordering-1      735
coffee-ordering-2      641
movie-finder            54
movie-ticket-1          37
movie-tickets-1        642
movie-tickets-2        377
movie-tickets-3        195
pizza-ordering-1       257
pizza-ordering-2      1211
restaurant-table-1     704
restaurant-table-2     494
restaurant-table-3     102
uber-lyft-1            646
uber-lyft-2            452
dtype: int64
```

We need to fix the 37 movie-ticket-1 instruction_ids

In [4]:
```
df = df.replace(['movie-ticket-1'], 'movie-tickets-1')
```

In [5]:
```
print (df.groupby('Instruction_id').size())
```

```
Instruction_id
auto-repair-appt-1    1161
coffee-ordering-1      735
coffee-ordering-2      641
movie-finder            54
movie-tickets-1        679
movie-tickets-2        377
movie-tickets-3        195
pizza-ordering-1       257
pizza-ordering-2      1211
restaurant-table-1     704
restaurant-table-2     494
restaurant-table-3     102
uber-lyft-1            646
uber-lyft-2            452
dtype: int64
```

Add the Service Type as a column (i.e. auto, coffee, movie, etc.)

In [6]:
```
df['service_type'] = df['Instruction_id'].str.split('-',expand=True)[0]
print (df.groupby('service_type').size())
```

```
service_type
auto          1161
coffee        1376
movie         1305
pizza         1468
restaurant    1300
uber          1098
dtype: int64
```

In [7]:

```
df
```

Out[7]:

	id	Conversation	Instruction_id	service_type
0	dlg-00055f4e-4a46-48bf-8d99-4e477663eb23	Hi, I'm looking to book a table for Korean fod...	restaurant-table-2	restaurant
1	dlg-0009352b-de51-474b-9f13-a2b0b2481546	Hi I would like to see if the Movie What Men W...	movie-tickets-1	movie
2	dlg-00123c7b-15a0-4f21-9002-a2509149ee2d	I want to watch avengers endgame where do you ...	movie-tickets-3	movie
3	dlg-0013673c-31c6-4565-8fac-810e173a5c53	I want to order a pizza from Bertuccis in Chel...	pizza-ordering-2	pizza
4	dlg-001d8bb1-6f25-4ecd-986a-b7eeb5fa4e19	Hi I'd like to order two large pizzas. Sure, w...	pizza-ordering-2	pizza
...	...	...	...	...
7703	dlg-ffc0c5fb-573f-40e0-b739-0e55d84100e8	I feel like eating at a nice restaurant tonigh...	restaurant-table-1	restaurant
7704	dlg-ffc87550-389a-432e-927e-9a9438fc4f1f	Hi Sally, I need a Grande iced Americano with ...	coffee-ordering-2	coffee
7705	dlg-ffcd1d53-c080-4acf-897d-48236513bc58	Good afternoon. I would like to order a pizza ...	pizza-ordering-2	pizza
7706	dlg-ffd9db94-36e3-4534-b99d-89f7560db17c	Hey. I'm thinking of seeing What Men Want toni...	movie-tickets-1	movie
7707	dlg-fffa6565-32bb-4592-8d30-fff66df29633	Hello. Can you help me purchase a couple of mo...	movie-tickets-3	movie

7708 rows × 4 columns

Connect to Database

In [8]:

```
import sqlite3
con = sqlite3.connect('selfdialogs.db')
```

Cleaning the Dataset for NLP

Cleaning Function

In [9]:

```
import re
def clean(s):
    s = s.replace(r'<lb>', "\n")
    s = s.replace(r'<tab>', "\i")
    s = re.sub(r'<br */*>', "\n", s)
    s = s.replace("&lt;", "<").replace("&gt;", ">").replace("&amp;", "&")
    s = s.replace("&amp;", "&")
    # markdown urls
    s = re.sub(r'\(https*://[^\)]*\)', "", s)
    # normal urls
    s = re.sub(r'https*://[^\s]*', "", s)
    s = re.sub(r'_+', ' ', s)
    s = re.sub(r'"+', '"', s)
    return str(s)
```

In [10]:

```
df["selfdialog_clean"] = ''
```

Iterate and Clean

In [11]:

```
for i, row in df.iterrows():
    df.at[i, "selfdialog_clean"] = clean(row.Conversation)
```

In [12]:

```
df.head()
```

Out[12]:

	id	Conversation	Instruction_id	service_type	selfdialog_clean
0	dlg-00055f4e-4a46-48bf-8d99-4e477663eb23	Hi, I'm looking to book a table for Korean fod...	restaurant-table-2	restaurant	Hi, I'm looking to book a table for Korean fod...
1	dlg-0009352b-de51-474b-9f13-a2b0b2481546	Hi I would like to see if the Movie What Men W...	movie-tickets-1	movie	Hi I would like to see if the Movie What Men W...
2	dlg-00123c7b-15a0-4f21-9002-a2509149ee2d	I want to watch avengers endgame where do you ...	movie-tickets-3	movie	I want to watch avengers endgame where do you ...
3	dlg-0013673c-31c6-4565-8fac-810e173a5c53	I want to order a pizza from Bertuccis in Chel...	pizza-ordering-2	pizza	I want to order a pizza from Bertuccis in Chel...
4	dlg-001d8bb1-6f25-4ecd-986a-b7eeb5fa4e19	Hi I'd like to order two large pizzas. Sure, w...	pizza-ordering-2	pizza	Hi I'd like to order two large pizzas. Sure, w...

NLP

In [13]:

```
import spacy
nlp = spacy.load('en')
```

Iterate and Perform NLP

In [14]:

```
for i, row in df.iterrows():
    if i % 1000 == 0:
        print(i)
    if(row["selfdialog_clean"] and len(str(row["selfdialog_clean"])) < 1000000):
        doc = nlp(str(row["selfdialog_clean"]))
        adjectives = []
        nouns = []
        verbs = []
        lemmas = []

        for token in doc:
            lemmas.append(token.lemma_)
            if token.pos_ == "ADJ":
                adjectives.append(token.lemma_)
            if token.pos_ == "NOUN" or token.pos_ == "PROPN":
                nouns.append(token.lemma_)
            if token.pos_ == "VERB":
                verbs.append(token.lemma_)

        df.at[i, "selfdialog_lemma"] = " ".join(lemmas)
        df.at[i, "selfdialog_nouns"] = " ".join(nouns)
        df.at[i, "selfdialog_adjectives"] = " ".join(adjectives)
        df.at[i, "selfdialog_verbs"] = " ".join(verbs)
        df.at[i, "selfdialog_nav"] = " ".join(nouns+adjectives+verbs)
        df.at[i, "no_tokens"] = len(lemmas)
```

```
0
1000
2000
3000
4000
5000
6000
7000
```

In [15]:
```
df.head()
```

Out[15]:

	id	Conversation	Instruction_id	service_type	selfdialog_clean	selfdialog_lemma	sel
0	dlg-00055f4e-4a46-48bf-8d99-4e477663eb23	Hi, I'm looking to book a table for Korean fod...	restaurant-table-2	restaurant	Hi, I'm looking to book a table for Korean fod...	hi , -PRON- be look to book a table for korean...	so
1	dlg-0009352b-de51-474b-9f13-a2b0b2481546	Hi I would like to see if the Movie What Men W...	movie-tickets-1	movie	Hi I would like to see if the Movie What Men W...	hi -PRON- would like to see if the movie what ...	n tic
2	dlg-00123c7b-15a0-4f21-9002-a2509149ee2d	I want to watch avengers endgame where do you ...	movie-tickets-3	movie	I want to watch avengers endgame where do you ...	-PRON- want to watch avenger endgame where do ...	av tin
3	dlg-0013673c-31c6-4565-8fac-810e173a5c53	I want to order a pizza from Bertuccis in Chel...	pizza-ordering-2	pizza	I want to order a pizza from Bertuccis in Chel...	-PRON- want to order a pizza from bertuccis in...	wh
4	dlg-001d8bb1-6f25-4ecd-986a-b7eeb5fa4e19	Hi I'd like to order two large pizzas. Sure, w...	pizza-ordering-2	pizza	Hi I'd like to order two large pizzas. Sure, w...	hi -PRON- would like to order two large pizza ...	I

Store to Database

In [16]:
```
df.to_sql('posts_nlp', con, if_exists='replace')
```

In [ ]:
```
```