Week 2: Data Management

Alfonso Vieyra, Denise Wei,
Raha Ghazi, Tom Gao

Dialogue Sum Train Set

4 Columns: ID, Dialogue, Summary, Topic:

- **test_id** corresponds to the ***unique*** test of each row - Will change this to a single integer ID
- **Dialogue** column contains the interaction between entities
- **Summary** column contains **multiple** summaries for a given dialogue
- **Topic** is a categorical variable that identifies the type of dialogue it corresponds to.

...

Dialogue Sum Insights

Topics column and duplicate dialogues/summaries:

- Row 1626 and 9118 have the same dialogue **but**...they have different summaries of the dialogue. Only two cases of duplicate dialogues have been found in training dataset. Test set contains more.
- Likewise, we have found a total of 24 summary duplicates
- There are 7,434 different topics - not sure if this will be a useful attribute to include
- Top 5 most popular are: shopping: 174, job interview: 161, daily casual talk: 125, phone call: 89, order food: 79

Example: Dialogue Duplicates

- We can see that despite the duplicate dialogue, there contains a **different** topic **categorization** and **summary**.

	id	dialogue	summary	topic
1626	train_1626	#Person1#: Any plans tonight? \n#Person2#: Not...	#Person1# invites #Person2# to have a drink be...	social communication
9118	train_9118	#Person1#: Any plans tonight? \n#Person2#: Not...	#Person1# invites #Person2# to take a hang-out...	after inteview

	id	dialogue	summary	topic
7618	train_7618	#Person1#: I'm searching for an old music box....	#Person1# is searching for an old music box ma...	old music box
11492	train_11492	#Person1#: I'm searching for an old music box....	#Person1# wants to buy an old music box with d...	music box

Example: Duplicate Summary

- Same topic, written differently.
- Dialogue for each summary are identical BUT were not detected due to small character differences (i.e. spaces)

	id	dialogue	summary	topic	dialogue_tokens	summary_tokens	sentiment
1085	train_1085	#Person1#: I'd like to take this opportunity t...	#Person1# thanks #Person2# for #Person2#'s help.	gratitude	61.0	12.0	1
9605	train_9605	#Person1#: I'd like to take this opportunity t...	#Person1# thanks #Person2# for #Person2#'s help.	thanks	62.0	12.0	0

```
train_df.iloc[1085, 1]
```

```
'#Person1#: I'd like to take this opportunity to thank you for everything you did for me.\n#Person2#: It's my pleasure. I enjoyed working with you.\n#Person1#: I wouldn't be able to make it without your help.\n#Person2#: Then keep up the good work.'
```

```
train_df.iloc[9605, 1]
```

```
'#Person1#: I'd like to take this opportunity to thank you for everything you did for me. \n#Person2#: It's my pleasure. I enjoyed working with you. \n#Person1#: I wouldn't be able to make it without your help. \n#Person2#: Then keep up the good work. '
```

Data Management

- Group data based on Dialogue.
- May need to use ML techniques to process dialogue for similarity - make uniqueness reliable and reduce redundancy
- Currently, thoughts are on creating 4 separate entity tables:

LLM's {
llmid;
Type;
}

Dialogue {
Did;
Dialogue;
dtokens;
}

Summary {
Sid;
did (references dialogue);
stokens;
}

Scores {
Llmid (references LLMs);
Rogue;
Bleu;
Meteor;
}

Final Thoughts

- Dialogues and summaries seem to have strong similarities that are sometimes the exact same minus a character or so
- Inclusion of sentiment column with the appropriate labels
- Topics columns seem too varied
- Inclusion of token counts might be useful for dialogue and summary tables

Thank you