Rohini Vishwanathan
Natural Language Processing
IST 332

Lab Assignment #2: Create basic NLP features

Google colab- https://colab.research.google.com/drive/1yxtUkpMzSwomSKOmKtvRNYogNB5W24LR?usp=sharing

Q1. Read the data into a DataFrame

```
print("Shape:", df.shape)
df.head(5)
```

Shape: (63230, 14)

	Stars	ReviewDate	Review	BusinessName	City	State	Zip	Price	_review_raw	_stems_list	_stems_top	Cleaned_Review_Stems	_lemmas_
0	5.0	2008-09-29 09:33:04	its the best pizza hut i'v found in florida al...	Pizza Hut	Orlando	FL	32819	3.50	its the best pizza hut i'v found in florida al...	[best, pizza, hut, found, florida, alway, got,...	[best, pizza, found, alway, got, great, deal, ...	best pizza found alway got great deal offer st...	
1	3.0	2013-12-06 23:22:26	This place is an interesting combo. The chef, ...	La Mongerie Bakery & Bistro	Atlanta	GA	30308	3.00	This place is an interesting combo. The chef, ...	[place, interest, combo, chef, david, incred, ...	[place, chef, incred, cook, food, perfect, tas...	place chef incred cook food perfect tast break...	
2	5.0	2008-11-16 09:44:04	Pizza Hut is great! You get huge pizzas for yo...	Pizza Hut	Orlando	FL	32819	3.50	Pizza Hut is great! You get huge pizzas for yo...	[pizza, hut, great, get, huge, pizza, money, s...	[pizza, great, get, huge, pizza, money, servic...	pizza great get huge pizza money servic quick ...	
3	5.0	2016-04-11 17:00:34	Always a great place to taste some tea with fr...	Teavana	Atlanta	GA	30326	3.00	Always a great place to taste some tea with fr...	[alway, great, place, tast, tea, friend, staff...	[alway, great, place, tast, tea, friend, staff...	alway great place tast tea friend staff friend	
4	4.0	2015-05-26 06:47:56	Delish! Stopped in on a random Sunday and so g...	Native Foods Cafe	Happy Valley	OR	97086	4.50	Delish! Stopped in on a random Sunday and so g...	[delish, stop, random, sunday, glad, normal, e...	[stop, sunday, glad, eat, found, make, good, c...	stop sunday glad eat found make good choic eat...	

2. Create cleaned review (stems) and keep top 500 stems

```
# confirm top-K
TOP_K, len(top_stems)
```

(500, 500)

3. Snippets of first 5 cleaned reviews (stems)

	Cleaned_Review_Stems
0	best pizza found alway got great deal offer st...
1	place chef incred cook food perfect tast break...
2	pizza great get huge pizza money servic quick ...
3	alway great place tast tea friend staff friend
4	stop sunday glad eat found make good choic eat...

4. Save the stems data frame as CSV

```
stems_out = df.copy()
stems_out.to_csv("/content/cleaned_restaurant_review_stems.csv", index=False)
print("Saved -> /content/cleaned_restaurant_review_stems.csv")
stems_out.head(3)
```

Saved -> /content/cleaned_restaurant_review_stems.csv

	Stars	ReviewDate	Review	BusinessName	City	State	Zip	Price	_review_raw	_stems_list	_stems_top	Cleaned_Review_Stems
0	5.0	2008-09-29 09:33:04	its the best pizza hut i'v found in florida al...	Pizza Hut	Orlando	FL	32819	3.50	its the best pizza hut i'v found in florida al...	[best, pizza, hut, found, florida, alway, got,...	[best, pizza, found, alway, got, great, deal, ...	best pizza found alway got great deal offer st...
1	3.0	2013-12-06 23:22:26	This place is an interesting combo. The chef, ...	La Mongerie Bakery & Bistro	Atlanta	GA	30308	3.00	This place is an interesting combo. The chef, ...	[place, interest, combo, chef, david, incred, ...	[place, chef, incred, cook, food, perfect, tas...	place chef incred cook food perfect tast break...
2	5.0	2008-11-16 09:44:04	Pizza Hut is great! You get huge pizzas for yo...	Pizza Hut	Orlando	FL	32819	3.50	Pizza Hut is great! You get huge pizzas for yo...	[pizza, hut, great, get, huge, pizza, money, s...	[pizza, great, get, huge, pizza, money, servic...	pizza great get huge pizza money servic quick ...

5. Aggregate to unique restaurants (new DataFrame)

```
df_rest[["BusinessName","City","State","All_Reviews_Text"]].head(5)
```

	BusinessName	City	State	All_Reviews_Text
0	1/2 Roll	Vancouver	BC	The ambience is kind of funky & interesting. F...
1	21 Nickels	Watertown	MA	I like 21 Nickels a ton and they would attract...
2	24 Hour Pizza Delivery	Boston	MA	This place needs to be shut down. I could not ...
3	3 Monkeys Pub & Grill	Vancouver	WA	Well, it is re-opened. Stopped in today just t...
4	5th St Ribs-N-Blues	Atlanta	GA	First visit here for my weekly southern BBQ fi...

▶ Start coding or generate with AI.

6. Annotate each row with NER tag categories

```
ner_cols = [c for c in df_rest.columns if c.startswith("NER_")]
df_rest[["BusinessName","City","State"] + ner_cols].head(5)
```

	BusinessName	City	State	NER_PERSON	NER_NORP	NER_FAC	NER_ORG	NER_GPE	NER_LOC	NER_PRODUCT	...	NER_WORK_OF_ART	NER_LAW	NER_LANGU
0	1/2 Roll	Vancouver	BC	31	11	6	27	8	1	2	...	0	0	
1	21 Nickels	Watertown	MA	30	8	2	48	26	1	2	...	2	0	
2	24 Hour Pizza Delivery	Boston	MA	10	2	1	15	4	0	0	...	0	0	
3	3 Monkeys Pub & Grill	Vancouver	WA	36	1	3	23	12	2	4	...	1	0	
4	5th St Ribs-N-Blues	Atlanta	GA	14	1	5	46	7	2	1	...	2	1	

5 rows × 21 columns

7. Create cleaned review (lemmas) as a string

	Cleaned_Review_Lemmas
0	its the best pizza hut found in florida always...
1	this place is an interesting combo the chef da...
2	pizza hut is great you get huge pizzas for you...
3	always great place to taste some tea with frie...
4	delish stopped in on random sunday and so glad...

8. Frequency based Bag-Of-words features

```
#frequency based Bag of Words features
full_bow.shape[1]
```

500

```
df_bow.head(5)
```

	BusinessName	City	State	Stars	ReviewDate	BOW_abl	BOW_absolut	BOW_across	BOW_actual	BOW_ad	...	BOW_worst	BOW_worth	BOW_would	BO
0	Pizza Hut	Orlando	FL	5.0	2008-09-29 09:33:04	0	0	0	0	0	...	0	0	0	
1	La Mongerie Bakery & Bistro	Atlanta	GA	3.0	2013-12-06 23:22:26	0	0	0	0	0	...	0	0	0	
2	Pizza Hut	Orlando	FL	5.0	2008-11-16 09:44:04	0	0	0	0	0	...	0	0	0	
3	Teavana	Atlanta	GA	5.0	2016-04-11 17:00:34	0	0	0	0	0	...	0	0	0	
4	Native Foods Cafe	Happy Valley	OR	4.0	2015-05-26 06:47:56	0	0	0	0	0	...	0	0	0	

5 rows × 505 columns