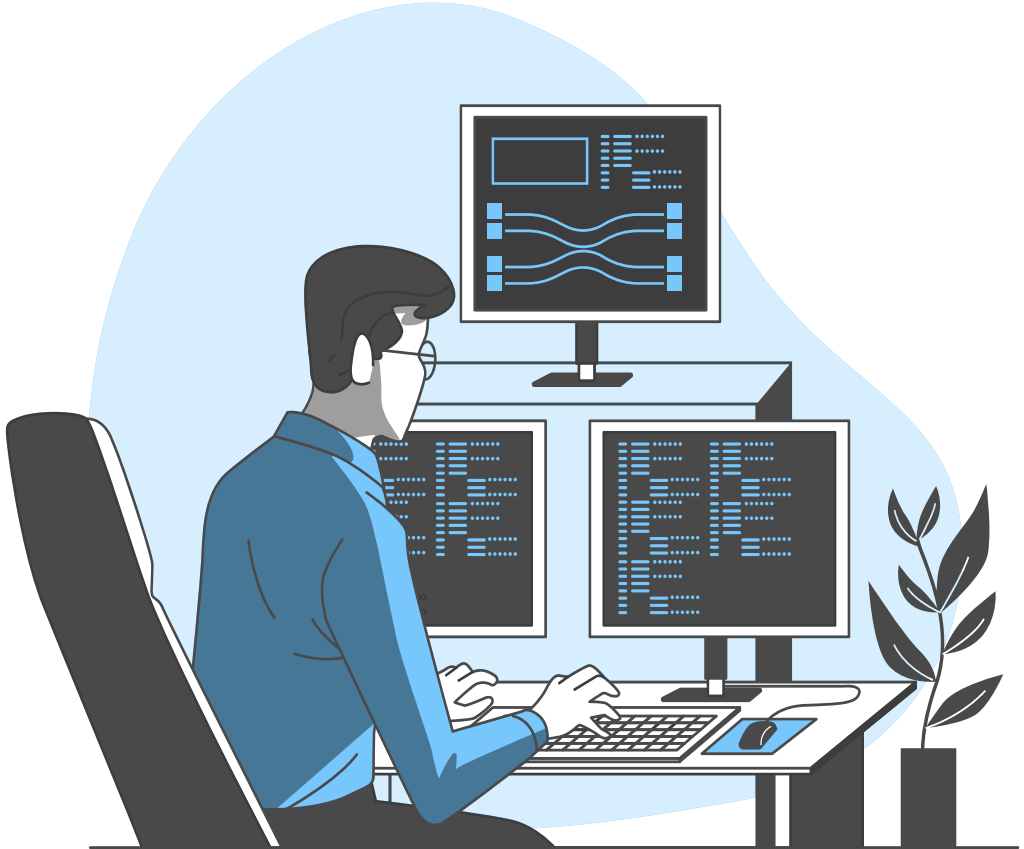

Week 1: Data Wrangling

Alfonso Vieyra, Denise Wei,
Raha Ghazi, Tom Gao

MeetingBank Dataset

- The MeetingBank Dataset contains 1,366 meetings transcripts from the city councils of 6 major U.S. cities.
- 6,892 segment-level summarization instances
- Relatively new data from 2022.
- Summary has varying lengths, ranging from 75 to 1.1k characters.

meeting_id string · *lengths* 27 37	source string · *lengths* 596 386k	type string · *lengths* 4 42	reference string · *lengths* 75 1.1k	city string · *classes* 6 values
LongBeachCC_08092022_22-0922	Speaker 4: Thank you. And can we do the functions for…	Agenda Item	Recommendation to increase appropriations in the General…	LongBeachCC
LongBeachCC_08092022_22-0917	Speaker 4: We're going to hear all of the presentations at…	Public Hearing	Recommendation to conduct a Budget Hearing to receive and…	LongBeachCC
LongBeachCC_08092022_22-0946	Speaker 4: Thank you very much. We will. We're going to…	Ordinance	Recommendation to declare ordinance amending the Long…	LongBeachCC
LongBeachCC_08092022_22-0932	Speaker 4: Great. Thank you. And, you know, we do have I…	Resolution	Recommendation to adopt resolution approving the…	LongBeachCC

DialogSum Dataset

- 13,000+ rows of dialogue, observed summary, and topic of the dialogue
- Recent data (from 2017 and 2019)
- Topics of dialogue vary greatly
- Dialogue length ranges from 200 to a little under 4,500 characters

dialogue string · *lengths* 190⇔689 54.2%	summary string · *lengths* 31⇔132 52.1%	topic string · *lengths* 7⇔12 30.5%
#Person1#: Hi, Mr. Smith. I'm Doctor Hawkins. Why are you here today?…	Mr. Smith's getting a check-up, and Doctor Hawkins advises him to have one…	get a check-up
#Person1#: Hello Mrs. Parker, how have you been? #Person2#: Hello Dr. Peters.…	Mrs Parker takes Ricky for his vaccines. Dr. Peters checks the record and then…	vaccines
#Person1#: Excuse me, did you see a set of keys? #Person2#: What kind of keys?…	#Person1#'s looking for a set of keys and asks for #Person2#'s help to find them.	find keys
#Person1#: Why didn't you tell me you had a girlfriend? #Person2#: Sorry, I though…	#Person1#'s angry because #Person2# didn't tell #Person1# that #Person2# had…	have a girlfriend

IMDB Dataset

- Data is from Kaggle and contains movie reviews
- Samples labeled with 0 (negative) and 1 (positive)
- Dataset is from 2022
- Review length ranges from 32 to 13704
- 40000 Total Reviews

	text	label
0	I grew up (b. 1965) watching and loving the Th...	0
1	When I put this movie in my DVD player, and sa...	0
2	Why do people who do not know what a particula...	0
3	Even though I have great interest in Biblical ...	0
4	Im a die hard Dads Army fan and nothing will e...	1

Yelp Dataset

	stars	text	date
0	3.0	If you decide to eat here, just be aware it is...	2018-07-07 22:09:11
1	5.0	I've taken a lot of spin classes over the year...	2012-01-03 15:28:18
2	3.0	Family diner. Had the buffet. Eclectic assortm...	2014-02-05 20:30:30
3	5.0	Wow! Yummy, different, delicious. Our favo...	2015-01-04 00:01:03
4	4.0	Cute interior and owner (?) gave us tour of up...	2017-01-14 20:54:15

- Data is from yelp.com and consists of reviews on various retail and ecommerce chains
- Approximately 700,000 reviews in total
- Text length ranges from 1 - 5000 characters
- Stars rating column is from 1 - 5
- Date of reviews ranges from 2005 - 2022

Amazon Dataset

- Data is from Kaggle and features reviews from customers on Amazon.com
- Dataset is from 2022
- 4915 Total Reviews
- Text length ranges from 3 - 8638 characters
- Ratings column is from 1 - 5

	reviewText	overall
0	No issues.	4.0
1	Purchased this for my device, it worked as adv...	5.0
2	it works as expected. I should have sprung for...	4.0
3	This think has worked out great.Had a diff. br...	5.0
4	Bought it with Retail Packaging, arrived legit...	5.0

Bonus Datasets

- Financial Phrases Sentiment Analysis: Focused on very industry specific language
- Customer Support on Twitter Summary: Usually automated responses from the customer service provider
- Meta Casual Conversations: No observed summary/sentiment
- TalkBank Phone Conversations: No observed summary/sentiment
- DailyDialog: Observed emotion but no observed sentiment