Week 2 Report

Nate

Public NLP Datasets Exploration

In this section, we analyze three foundational datasets that have shaped the field of NLP.

Required Table

Dataset	Content Description	Data Format(s)
Project Gutenberg	Over 70,000 public domain books (literary classics).	Plain Text (.txt), HTML
Yelp Reviews	Millions of user reviews, business info, and star ratings.	JSON, CSV
20 Newsgroups	~20,000 newsgroup posts across 20 different interest groups.	Plain Text

Dataset Analysis

- Project Gutenberg: This is a cornerstone for historical language analysis. Because it consists of complete books, it is ideal for training models to understand long-range context and formal grammar.

- Yelp Reviews: This dataset provides labled data, making it perfect for supervised learning where a model learns to associate specific words with positive or negative sentiment.

- 20 Newsgroups: For bag-of-word and clustering tasks. Since the data is pre-sorted into categories (like *sci.med* or *rec.autos*), it serves as a benchmark for how well a model can distinguish between subjects.

  This is all to say that though the content might be focused on some picturler subject or context, different sorts of contect which are more general can be derived & learned from, like that of using classic texts to model the why humans use their grammar within their langauges.

  I also recall here that a bag-of-words method is actually derived from a lingustical context where it's thought that in both phenolog & morhoplogy either two different elements can be of the same 'kind' or two similer elements of the same 'meaning' can appear in different contexts.

Structured vs. Unstructured Data in NLP

Modern NLP involves converting human language into a format computers can calculate.

Required Table

Data Type	Definition	NLP Examp
Structured Data	Data that resides in fixed fields within a record or file (e.g., a table).	A SQL data Spreadsheet
Unstructured Data	Information that does not have a pre-defined data model or organization.	A raw trans A collection

Discussion

Human language is naturally fluid, idiomatic, and context-dependent. Forcing language into a rigid table usually strips away the nuances (sarcasm, tone, and flow) that make NLP valuable. Unstructured data requires heavy preprocessing. Before a model can use it, tokenization must be performed (splitting text into pieces), normalization (lowercasing, removing punctuation), and embedding (converting text into numerical vectors).

Futhermore it can be seen that the desire for some model outcome depends not wholly of the type of data but rather the way it's designed or engineered.

Cloud Storage and LLM Challenges

Cloud Storage Platforms

1. AWS S3: An object storage service used for Data Lakes. It allows NLP practitioners to store massive raw text files and access them via high-speed distributed computing.

2. Google Cloud Storage: Integrated deeply with TPU (Tensor Processing Unit) clusters, making it a preferred choice for training very large transformer models.

3. Azure Blob Storage: Used extensively in enterprise NLP for storing unstructured data like logs and customer service transcripts in a secure, compliant environment.

   Such cloud storage solutions optitions will usually included features like formatting, processing, and visioning in order to do a mulitude of things related to them; think creating a table or catelogying or creating branches to try something with a NLP model idea.

Data Scale

Training a modern LLM requires trillions of tokens. The sheer scale creates a "bottleneck" where the time it takes to move data from storage to the GPU/TPU can become longer than the actual computation time. This requires complex distributed systems to ensure the "data pipeline" never runs dry.

Reflection

Overcoming Traditional Limitations

Modern cloud storage utilizes "distributed architectures." Unlike a traditional hard drive, cloud storage spreads data across thousands of nodes. This allows for Parallelism where multiple GPUs can take in data at different parts of the dataset simultaneously. This eliminates the storage-access bottleneck that previously made training LLMs impossible.

Looking deeper into parallelism, often vGPU or GPU Over Frabic is used to give nodes partial access to other nodes, thereby creating a orgnaized logic of hardward into a distribued network of compute.

Excitement and Concerns

My exciment is towards datasets remaining public and future datasets being so, but it's contrary would be my fear: a great amount of privite data can result in solos which can stop modeling like sentiment analysis.