

Generating Analytic Specifications for Data Visualization from Natural Language Queries using Large Language Models

Subham Sah

ssah1@uncc.edu

Rishab Mitra

rmitra34@gatech.edu

Arpit Narechania

anarechania3@gatech.edu

Alex Endert

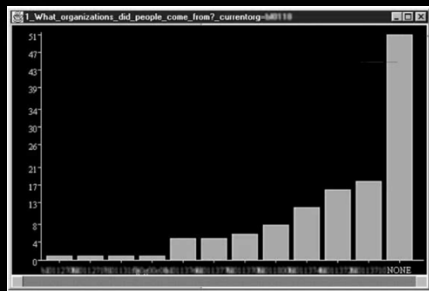
endert@gatech.edu

John Stasko

stasko@cc.gatech.edu

Wenwen Dou

wdou1@uncc.edu



InfoStill 2001



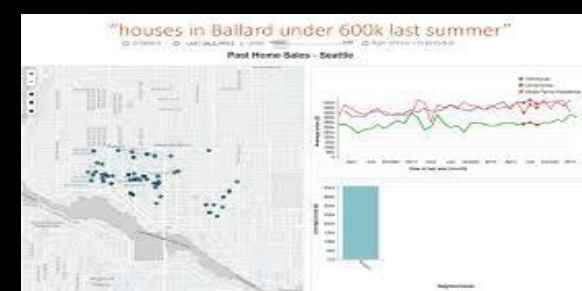
Articulate 2011



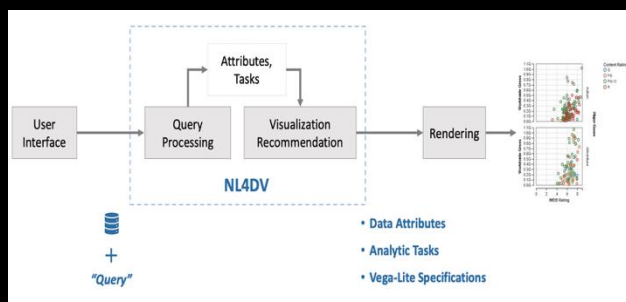
DataTone 2015



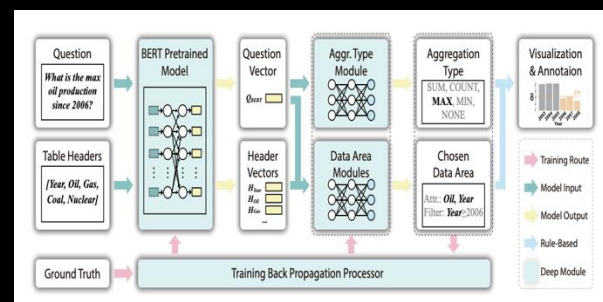
Orko 2018



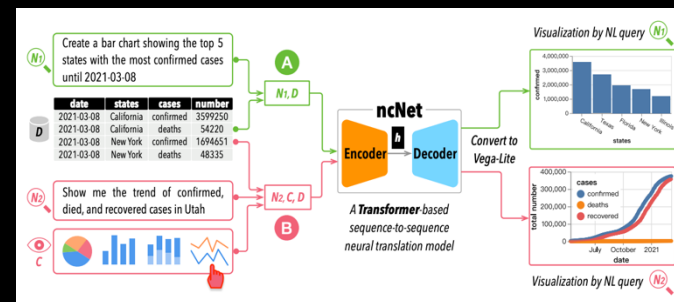
Evizeon 2018



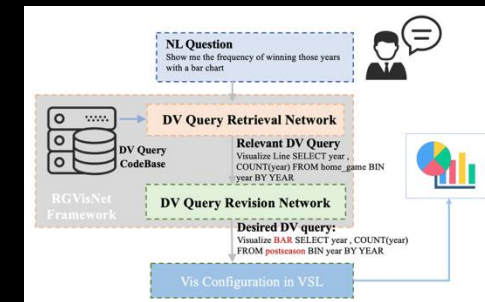
NL4DV 2020



Advisor 2021



ncNet 2021



RGVisNet 2022



IMDB
Movies

Title	Worldwide Gross	Production Budget	Release Year	Content Rating	Running Time	Genre	Creative Type	Rotten Tomatoes Rating	IMDB Rating
Titanic	1.84G	200M	1997	PG-13	194	Thriller	Historical Fiction	82	7.4
Pirates of the Caribbean: Dead Man's Chest	1.07G	225M	2006	PG-13	151	Adventure	Fantasy	54	7.3
The Dark Knight	1.02G	185M	2008	PG-13	152	Action	Super Hero	93	8.9
Pirates of the Caribbean: At World's End	961M	300M	2007	PG-13	167	Adventure	Fantasy	45	7



"Show the relationship between budget and IMDB rating for action and adventure movies that grossed over 100M"

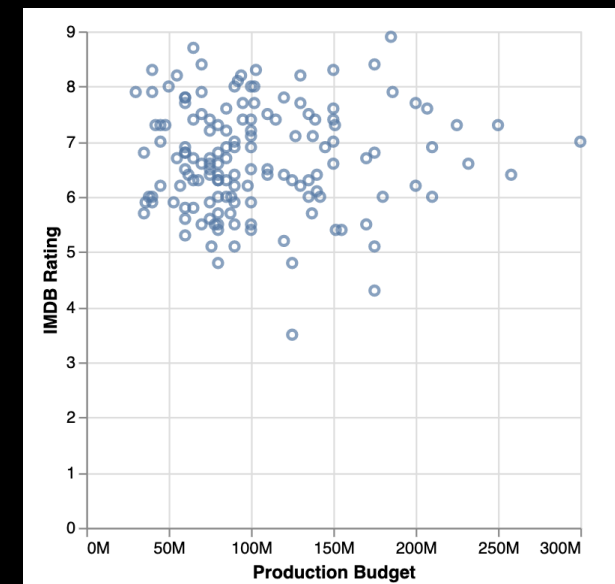
NL4DV

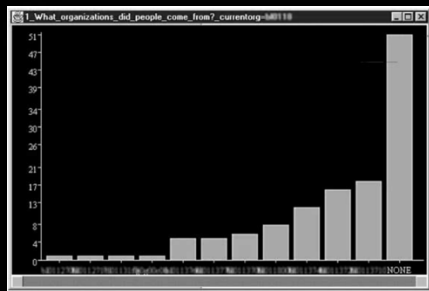
Attributes

```
{
  "Production Budget": {...},
  "IMDB Rating": {...},
  "Genre": {...},
  "Worldwide Gross": {...},
}
```

Tasks

```
{
  "Correlation": ["Production Budget",
                  "IMDB Rating"],
  "Filter": ["Genre" IN ["Action", "Adventure"],
             "Worldwide Gross" > 100000000]
}
```





InfoStill 2001



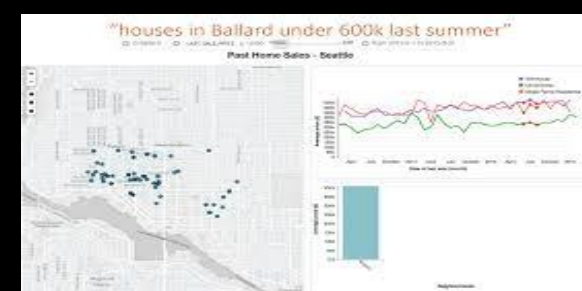
Articulate 2011



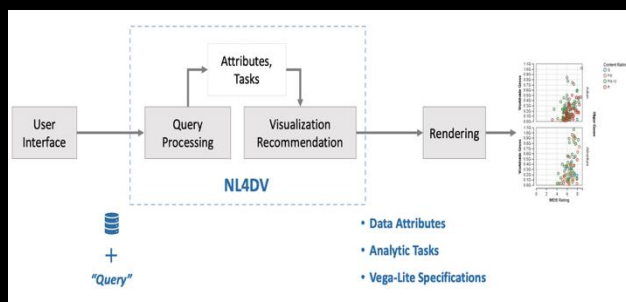
DataTone 2015



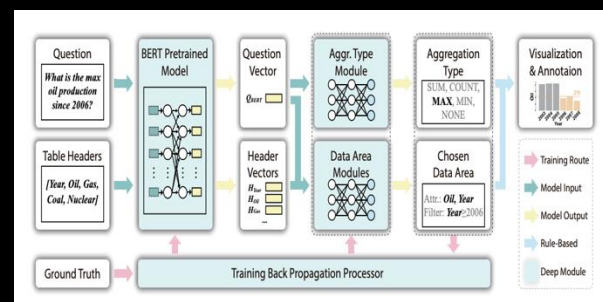
Orko 2018



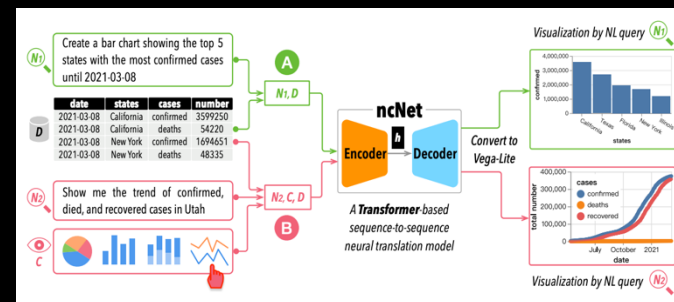
Evizeon 2018



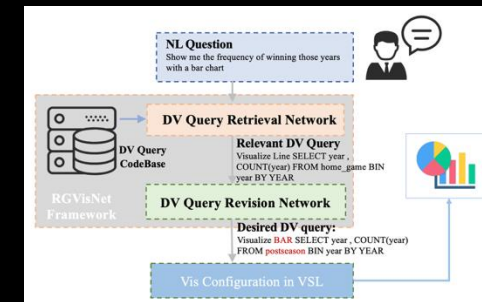
NL4DV 2020



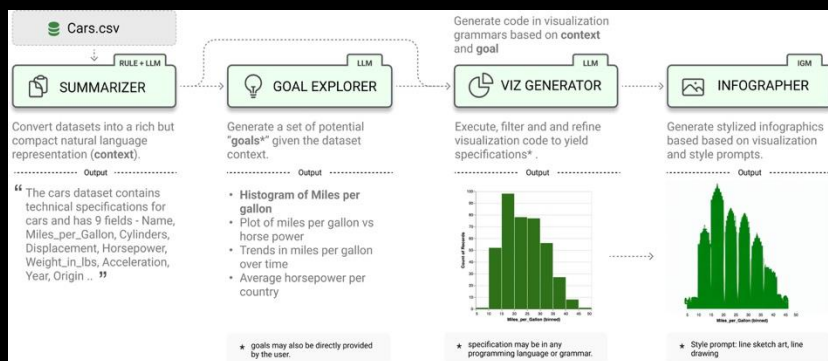
Advisor 2021



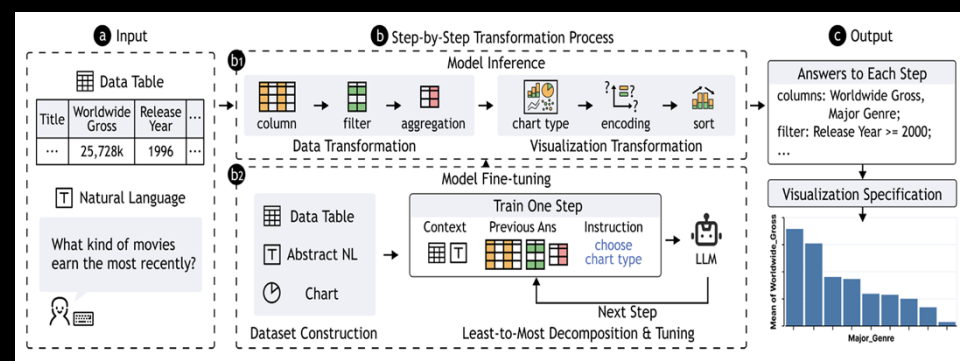
ncNet 2021



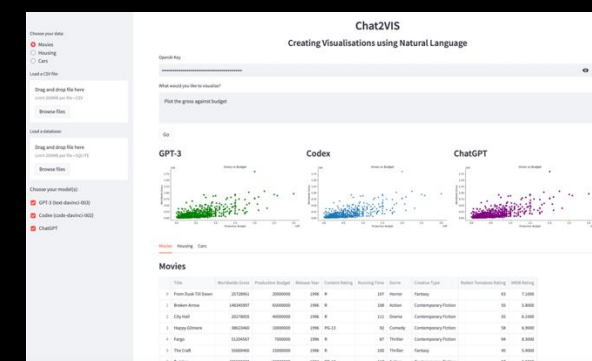
RGVisNet 2022



LIDA 2023



chartGPT 2023



chat2vis 2023

LIDA 2023 (Viz Generator Module)

You are a helpful assistant highly skilled in recommending a DIVERSE set of visualization code. Your input is an example visualization code, a summary of a dataset and an example visualization goal that the user has already seen. Given this input, your task is to recommend additional visualizations that a user may be interested. Your recommendation may consider different types of valid data aggregations, chart types, clearer ways of displaying information and uses different variables from the data summary. THE CODE YOU GENERATE MUST BE CORRECT (follow the language syntax and syntax of the visualization grammar) AND FOLLOW VISUALIZATION BEST PRACTICES.

chat2vis 2023

Use a dataframe called df from <dataset>.csv with columns <insert columns here>.

Label the x and y axes appropriately.

Add a title. Set the fig suptitle as empty.

Using Python version 3.9.12, create a script using the dataframe df to graph the following:

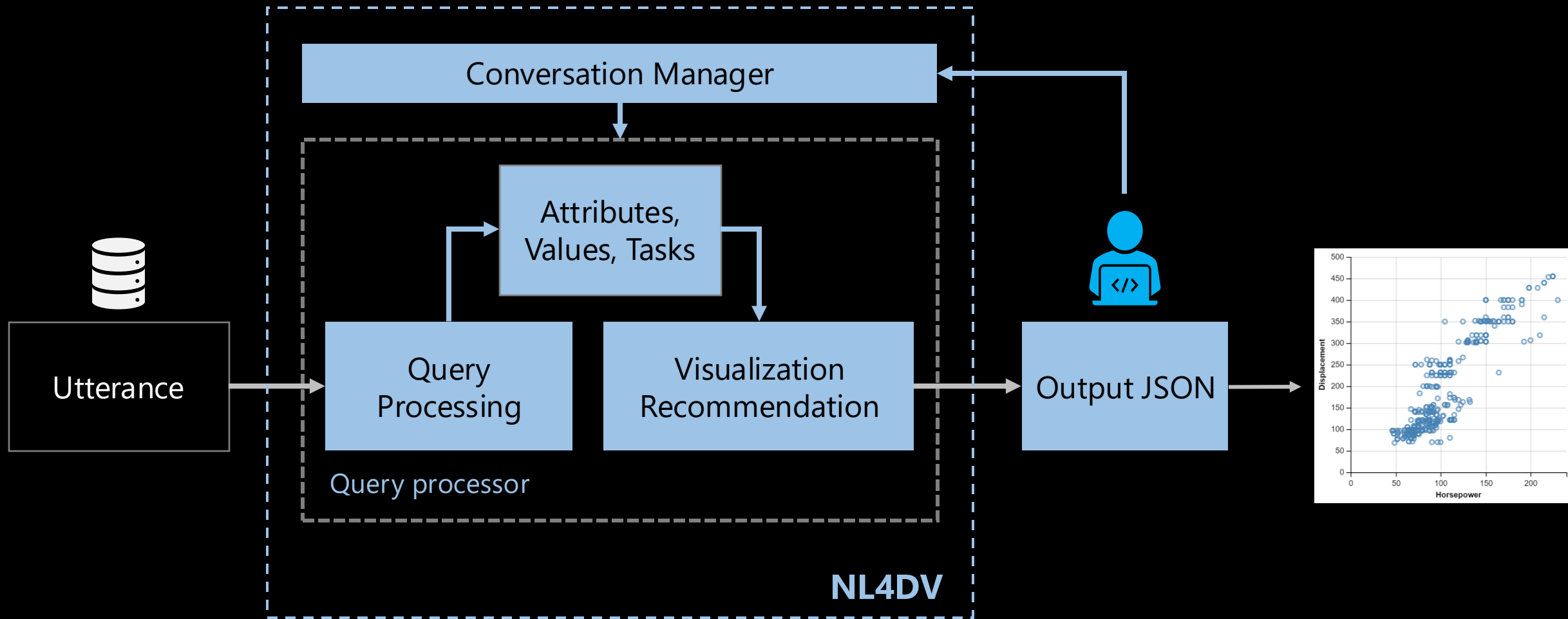
<INSERT NL query here>

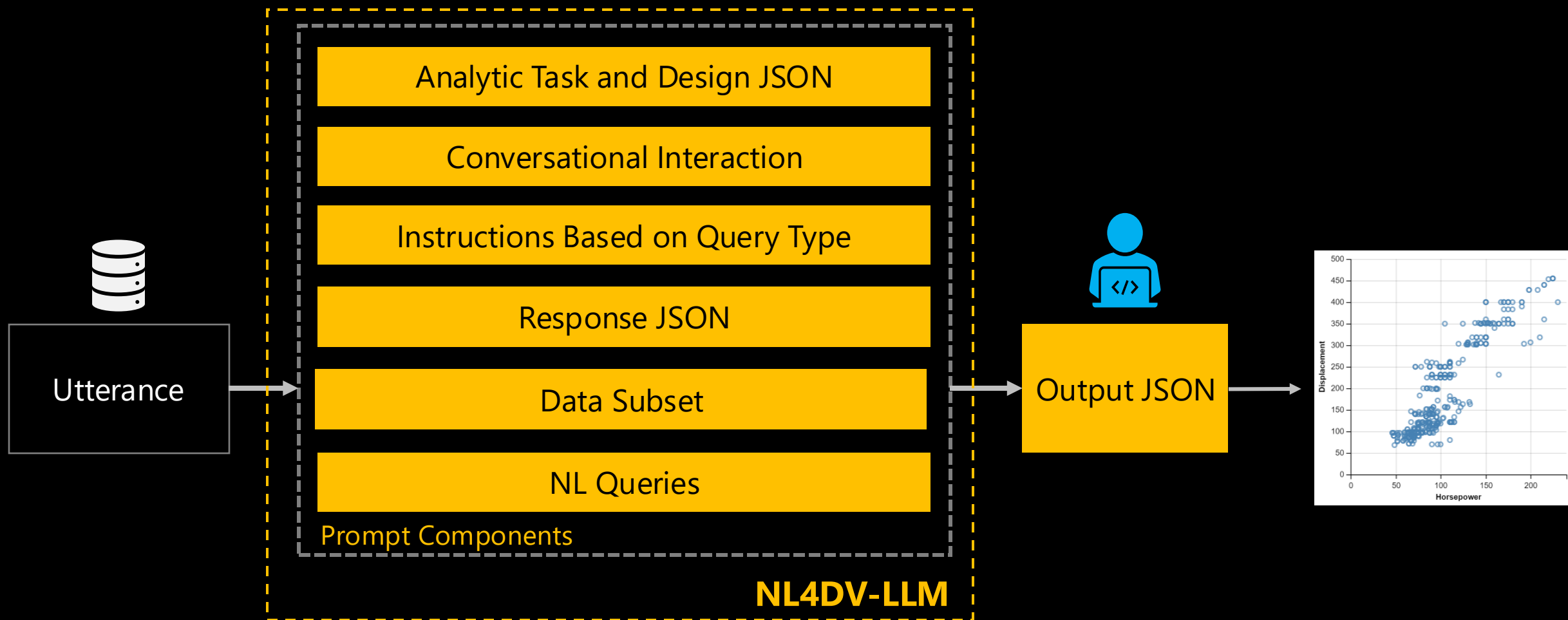
Use import pandas as pd, and import matplotlib.pyplot as plt in your code.



Can we get the best of both worlds?







nl4dv/nl4dv/promptgenie/pro

x

+

https://github.com/nl4dv/nl4dv/blob/master/nl4dv/promptgenie/promptgenie.py

Relaunch to update

Files

master

Go to file

examples

nl4dv

attributegenie

autogenie

conversationgenie

datagenie

promptgenie

querygenie

taskgenie

tests

updategenie

utils

visgenie

vlgenie

__init__.py

.gitignore

LICENSE

MANIFEST.in

README.md

Code

Blame

225 lines (216 loc) · 21 KB

Raw

```
1 class PromptGenie:
2
3     def __init__(self, nl4dv_instance):
4         # nl4dv instance
5         self.nl4dv_instance = nl4dv_instance
6         self.prompt = ""
7         Consider the below JSON array of objects describing low-level analytic tasks (fundamental operations that users perform when interacting with c
8         [
9         {
10            "Name": "Correlation",
11            "Description": "Given a set of data cases and two attributes, determine useful relationships between the values of those attributes.",
12            "Pro Forma Abstract": "What is the correlation between attributes X and Y over a given set S of data cases?",
13            "Examples": ["Is there a correlation between carbohydrates and fat?", "Do different genders have a preferred payment method?", "What is the rel
14            "taskMap Encoding": "correlation",
15            "Attribute Data Types and Visual Encodings":
16                [{
17                    "X axis": "Quantitative",
18                    "Y axis": "Quantitative",
19                    "Other Encoding": "<choose from : [\"Nominal\", \"Ordinal\", \"Quantitative\", \"Temporal\"]>",
20                    "Sort": null,
21                    "Filter": null,
22                }],
23            "Attributes and Visual Encodings Description": "The \"X axis\" key indicates that the first attribute is used for the horizontal or x-axis of the
24            "Recommended Visualization": ["Scatterplot"]
25        },
26        {
27            "Name": "Derived Value",
28            "Description": "Given a set of data cases, compute an aggregate numeric representation of those data cases.",
29            "Pro Forma Abstract": "What is the value of aggregation function F over a given set S of data cases?",
30            "Examples": ["What is the average calorie content of Post cereals?", "What is the gross income of all stores combined?", "How many manufactur
31            "taskMap Encoding": "derived_value",
32            "Attribute Data Types and Visual Encodings":
33                [{
34                    "X axis": "Nominal" or "Ordinal",
35                    "Y axis": "Quantitative",
36                    "Other Encoding": "<choose from : [\"Nominal\", \"Ordinal\", \"Quantitative\", \"Temporal\"]>",
37                    "Sort": null,
38                    "Filter": null,
39                }],
```



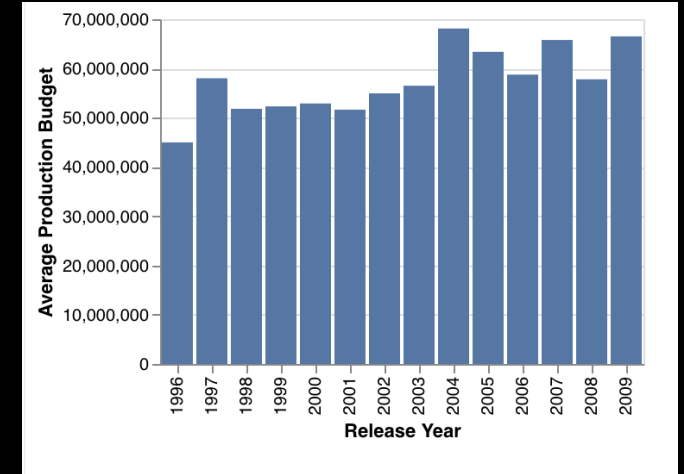
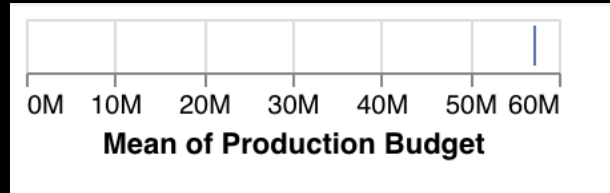
IMDb
Movies

Query

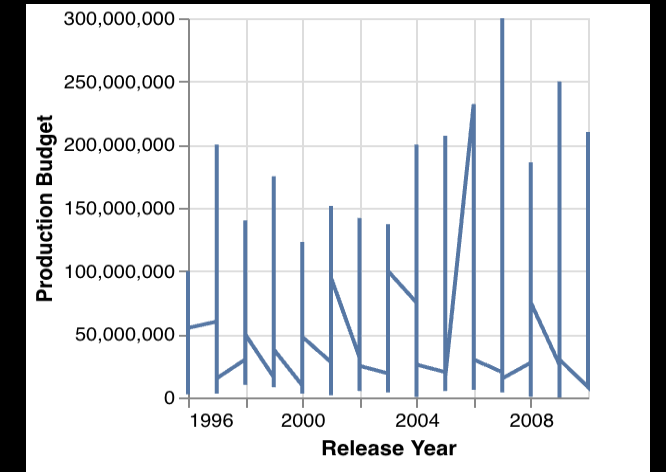
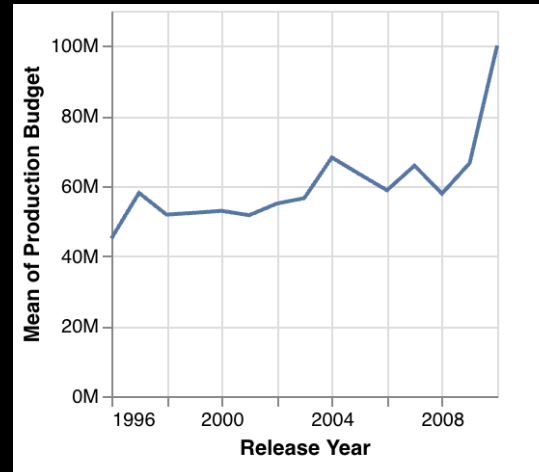
NL4DV

NL4DV-LLM

Average production budget from 1996 to 2009



I want to see trend of production budget by the release year



Evaluation Results on NLVCorpus (~700 queries)

	NL4DV	NL4DV-LLM (on GPT-4)
Accuracy	64.05%	87.02%
Average Response Time	3 seconds	25 seconds
Expressivity	Restricted to supporting queries that only follow certain syntactical structures	Handles a wide range of query forms and structures
Consistency	Always provides reliable, deterministic output for supported query types	Sometimes hallucinates attributes or spuriously fails to provide output

Get started with

`pip install nl4dv`

For documentation and examples, check out

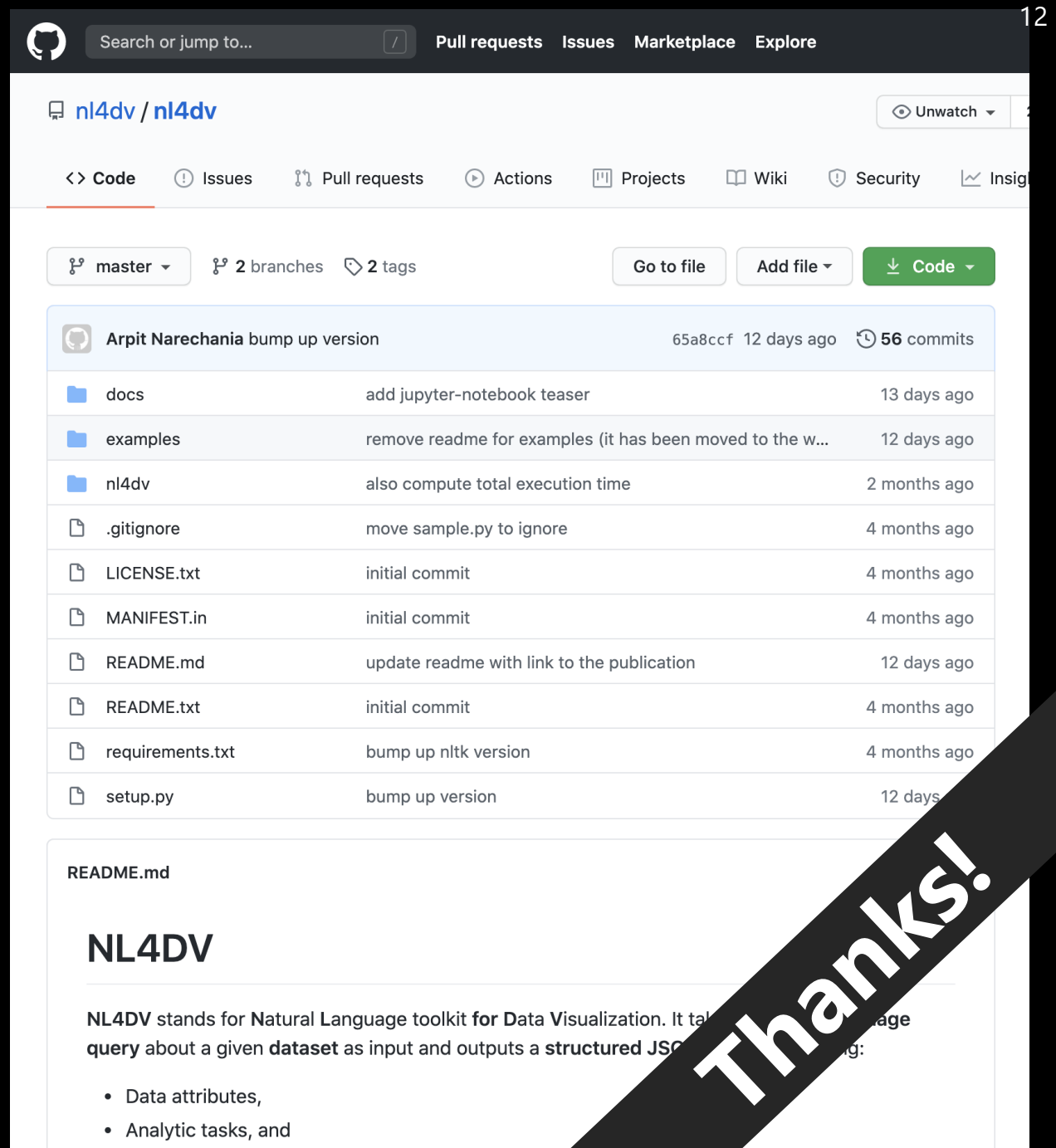
nl4dv.github.io

```
from nl4dv import NL4DV

data_url="https://raw.githubusercontent.com/nl4dv/nl4dv/master/examples/
assets/data/movies-w-year.csv"

nl4dv_instance = NL4DV(data_url=data_url, processing_mode="gpt",
gpt_api_key="[OpenAI KEY HERE]")

output = nl4dv_instance.analyze_query("create a barchart showing average
gross across genres")
```



The screenshot shows the GitHub repository for `nl4dv`. The repository is owned by `nl4dv` and has a `master` branch with 2 branches and 2 tags. The commit history shows a series of updates, including adding a jupyter-notebook teaser, removing the README for examples, and bumping up the version. The README.md file is visible, showing the project's purpose: NL4DV stands for Natural Language toolkit for Data Visualization. It takes a natural language query about a given dataset as input and outputs a structured JSON object. The README also lists features like Data attributes and Analytic tasks.

Search or jump to... Pull requests Issues Marketplace Explore

nl4dv / nl4dv Unwatch

<> Code Issues Pull requests Actions Projects Wiki Security Insights

master 2 branches 2 tags Go to file Add file Code

Arpit Narechania bump up version 65a8ccf 12 days ago 56 commits

docs	add jupyter-notebook teaser	13 days ago
examples	remove readme for examples (it has been moved to the w...	12 days ago
nl4dv	also compute total execution time	2 months ago
.gitignore	move sample.py to ignore	4 months ago
LICENSE.txt	initial commit	4 months ago
MANIFEST.in	initial commit	4 months ago
README.md	update readme with link to the publication	12 days ago
README.txt	initial commit	4 months ago
requirements.txt	bump up nltk version	4 months ago
setup.py	bump up version	12 days ago

README.md

NL4DV

NL4DV stands for Natural Language toolkit for Data Visualization. It takes a natural language query about a given dataset as input and outputs a structured JSON object. For example, given the query "create a barchart showing average gross across genres", it outputs:

- Data attributes,
- Analytic tasks, and

Thanks!

Going Forward

Improving the Usability of NL4DV-LLM

- Experiment and evaluate the prompt with different LLMs
Gpt-4o has shown to be much faster and retain similar accuracies, and o1 could provide enhanced explainability capabilities.
- Evaluate NL4DV-LLM against other LLM-based NL2VIS systems
Also utilize a larger scale dataset like VisEval to gain better understandings of accuracies against query and dataset complexity.