Managing Data Platforms for Smart Cities Using Large Language Models

Supplement

Marcin Krystek, Mikołaj Basiński, Mikołaj Morzy, Cezary Mazurek

This document serves as a supplement to the paper "*Managing Data Platforms for Smart Cities Using Large Language Models*" presented at the 32nd International Conference on Information Systems Development (ISD 2024). It contains detailed results of experiments described in the paper.

Experiment 1: identify and label concepts

The detailed results for individual tables are presented in Table 1 (*Graves*), Table 2 (*MPK Stops*), Table 3 (*Bike Stations*), and Table 4 (*Address Points*). Figure 1 presents an example of a dataset query with context.

Experiment 2: Generate human-like descriptions

The detailed descriptions of tables used in Experiment 2 are listed in Table 5. Figure 2 presents an example of the *Bike Stations* dataset description query.

Examples of questions and answers regarding the *Bike Stations* dataset

Question: *Considering type and nature of each identified features in Bike Stations dataset, for each feature suggest two different tools or programming libraries which are most suitable for processing this kind of data. Suggest also an analytical technique which is the most relevant for particular feature.*

Answer:

```
1. For processing the coordinates feature in the city bike station dataset, suitable tools or programming libraries could include:
   - GeoPandas: A Python library for working with geospatial data that provides easy-to-use tools for handling geometric data
        structures.
   - Leaflet: A JavaScript library for interactive maps that can be used to visualize and interact with the geographic coordinates
        .
   Relevant analytical technique: Spatial analysis, which involves analyzing patterns and relationships within geographic data.
2. For processing the bike racks feature in the city bike station dataset, suitable tools or programming libraries could include:
   - Pandas: A Python library for data manipulation and analysis that can handle numerical data efficiently.
   - Matplotlib: A Python library for creating static, animated, and interactive visualizations in Python.
   Relevant analytical technique: Descriptive statistics, which can be used to summarize and analyze the distribution of bike
        racks at different stations.
3. For processing the updated feature in the city bike station dataset, suitable tools or programming libraries could include:
   - datetime: A Python module for manipulating dates and times that can be used to handle timestamps.
   - Moment.js: A JavaScript library for parsing, validating, manipulating, and formatting dates and times.
   Relevant analytical technique: Time series analysis, which involves analyzing data collected over time to identify patterns and
        trends.
```

Question: *Elaborate Bike Stations dataset domain context description. Explain relationship between Bike Stations dataset and urban lifestyle or city inhabitants habits.*

Answer:

```
The city bike station dataset provides information about bike stations in a city, including details such as the coordinates of the
     station, the number of bike racks available, the number of bikes parked at the station, the number of free racks available,
     the name of the station, and the last time the station status was updated. This dataset is relevant to urban lifestyle and
     city inhabitants' habits as it reflects the availability and usage of bike-sharing services in the city.
Bike stations play a crucial role in promoting sustainable transportation options, reducing traffic congestion, and improving air
     quality in urban areas. City inhabitants who use bike-sharing services can benefit from a convenient and environmentally
     friendly mode of transportation for short trips around the city. The dataset can also provide insights into the popularity
     of cycling as a means of commuting or recreation in the city, as well as the overall interest in promoting a healthier and
     more active lifestyle among residents.
```

Table 1: Names and descriptions of the *Graves* dataset features generated by the LLM

attribute name	**description**	**type**	**code**	**LLM name**	**LLM description**	**LLM type**	**uniq values**
cm_id	graveyard identifier	number	0kqa	graveyard_id	The id of the graveyard where the grave is located	int	1
cm_nr	graveyard number	number	qawU	grave_id	Unknown attribute due to lack of data.	int	50
g_date_birth	person's date of birth	date	wRc2	birth_date	The date when the person was born	date	50
g_date_burial	person's burial date	date	sAK1	burial_date	The date when the person was buried	date	50
g_date_death	person's date of death	date	c1gE	death_date	The date when the person died	date	50
g_family	type of the grave	string	dXUz	-	Unknown attribute due to lack of data.	string	0
g_field	graveyard field identifier	string	O8m2	-	Unknown attribute due to lack of data.	string	0
g_name	person's name	string	Q+R2	first_name	The first name of the person buried in the grave	string	48
g_place	place number	string	pfeb	grave_location	The location of the grave in the graveyard	string	50
g_quarter	graveyard quarter	string	ambd	grave_type	The type of the grave (P for private, L for public)	string	21
g_row	row identifier inside quarter	string	PBeA	grave_number	The number of the grave in the graveyard	string	48
g_size	size of the grave	string	6Itn	-	Unknown attribute due to lack of data.	string	0
g_surname_name	surname and name	string	dstp	-	Unknown attribute due to lack of data.	string	0
g_surname	person's surname	string	dyjH	last_name_lowercase	The last name of the buried person in lowercase	string	50
paid	is the fee paid	number	Pai5	-	Unknown attribute due to lack of data.	int	1
print_name	capital name	string	nqiu	full_name	The full name of the person buried in the grave	string	48
print_surname_name	capital surname and name	string	vWTd	-	Unknown attribute due to lack of data.	string	0
print_surname	capital surname	string	dK4E	last_name	The last name of the person buried in the grave	string	50

Table 2: Names and descriptions of the *MPK Stops* dataset features generated by the LLM

title	**description**	**code**	**LLM name**	**LLM description**	**LLM type**	**uniq values**
stop type	distinguish between tram and bus stop	DIkr	direction_count	The number of directions at the stop	int	2
lines	list of public transport lines using this stop	EkmW	line_numbers	List of lines that stop at the stop	List[int]	43
name	common name	cLND	stop_name	Name of the stop	string	49
string stop type	distinguish between tram and bus stop	PVtE	vehicle_type	Type of the vehicle that stops at the stop (T for tram, A for bus)	string	2
description	stop type description	IOLd	stop_type	Type of the stop (tram or bus)	string	3

Table 3: Names and descriptions of the *Bike Stations* dataset features generated by the LLM

attribute name	code	LLM name	LLM description	LLM type	uniq values
bike_racks	yeNA	total_docks	The total number of docks at the bike station. It is a string that can be converted to an integer.	int	11
bikes	lXWC	broken_bikes	The number of broken bikes at the bike station. It is a string that can be converted to an integer.	int	1
free_racks	Xc78	available_bikes	The number of bikes currently available at the bike station. It is a string that can be converted to an integer.	int	14
label	HIkK	station_name	The name of the bike station	string	50
updated	Ebor	last_update	The date and time of the last update of the bike station's data. It is a string in the format "YYYY-MM-DD HH:MM	date	1

Table 4: Names and descriptions of the *Address Points* dataset features generated by the LLM

attribute name pl	code	LLM name	LLM description	LLM type	uniq values
adres	Lihs	secaddress_lowercase	represents the address in lowercase	string	50
gimnazjum	rswz	-	unknown attribute, lack of data to identify the concept.	string	5
goap	Hi1E	sector	represents the sector of the city where the address point is located	string	3
goap_odpady	ZE3B	agglomeration	represents the agglomeration where the city is located.	string	2
kod	IdFd	postal_code	represents the postal code of the address point	string	8
komisariat	dkJA	police_station	represents the nearest police station to the address point	string	2
krajobrazowa	epFx	urbanization_status	represents the urbanization status of the address point.	string	2
nazwa_parafii	q4tJ	parish	represents the parish the address point belongs to	string	4
nr	xKXa	house_number	represents the house number of the address point	string	36
obwody_ro	JGa4	-	unknown attribute, lack of data to identify the concept	int	3
obwody	tCdY	-	unknown attribute, lack of data to identify the concept.	int	5
okregi_ro	MiAs	district	represents the district of the city where the address point is located	string	5
okregi	UUBF	-	unknown attribute, lack of data to identify the concept	string	2
parafie	Wobd	-	unknown attribute, lack of data to identify the concept	int	4
podstawowa	nDES	-	unknown attribute, lack of data to identify the concept	int	4
rejon	KiOW	-	unknown attribute, lack of data to identify the concept	-	0
samorzady_lokalne	V9C8	neighborhood	represents the neighborhood of the city where the address point is located	string	4
typ	nt1q	street_prefix	represents the prefix of the street name (e.g., "ul." for "ulica" which means "street" in Polish).	string	1
wg_imienia	r4Lg	-	unknown attribute, lack of data to identify the concept.	string	38
wg_imienia_wydruk	RjaC	street_name	string, represents the name of the street of the address point	string	38
wg_nazwiska	lWQ9	street_name_lowercase	represents the name of the street of the address point in lowercase	string	38
wg_nazwiska_wydruk	lsMf	street_name_capitalized	represents the name of the street of the address point with the first letter capitalized	string	38

```
{
  "features": [
    {
      "geometry": {
        "coordinates": [
          16.7779975682727,
          52.4801856747057
        ],
        "type": "Point"
      },
      "id": 88835,
      "properties": {
        "Hi1E": "Sektor III",
        "IdFd": "60-480",
        "JGa4": 1,
        "KiOW": null,
        "Lihs": "turystyczna 1",
        "MiAs": "I-Kiekrz",
        "RjaC": "Turystyczna",
        "UUBF": "VI",
        "V9C8": "Kiekrz",
        "WObd": 1202,
        "ZE3B": "Aglomeracja Poznań",
        "dkJA": "Komisariat Policji Poznań-Jeżyce",
        "epFx": "Zurbanizowany",
        "lWQ9": "turystyczna",
        "lsMf": "Turystyczna",
        "nDES": 28,
        "nt1q": "ul.",
        "q4tJ": "pw. Świętego Michała Archanioła i
                Wniebowzięcia Najświętszej Maryi Panny
                w Poznaniu - Kiekrzu",
        "r4Lg": "turystyczna",
        "rswz": 0,
        "tCdY": 188,
        "xKXa": "1"
      },
      "type": "Feature"
    },
    [..]
  ]
}
```


Figure 1: Example of a dataset query with context

```
By analyzing the data from city bike stations, urban planners and policymakers can make informed decisions about expanding bike-
     sharing networks, improving infrastructure for cyclists, and promoting initiatives to encourage more people to use bicycles
     as a mode of transportation. Ultimately, the city bike station dataset contributes to creating a more sustainable and
     livable urban environment that supports healthy and active lifestyles for city inhabitants.
```

Question: *Considering Bike Stations dataset domain context and data origin describe two possible analytical scenarios which can be carried out by professional data scientist. Each scenario should include problem statement and goal of the analysis.*

Answer:

```
Scenario 1: Problem Statement: A city wants to optimize the distribution of bikes among its bike stations to ensure that there are
      enough bikes available for users at all times.
Goal of the Analysis: The data scientist can analyze the historical data of bike racks, bikes, and free racks at each station
     along with the timestamps of when the station status was updated. By identifying patterns in bike usage and station
     availability, the goal is to create a predictive model that can recommend the optimal distribution of bikes among stations
     to minimize the chances of stations running out of bikes or racks.
Scenario 2: Problem Statement: A city wants to improve the efficiency of its bike sharing system by identifying stations that are
     consistently underutilized or overutilized.
Goal of the Analysis: The data scientist can analyze the data on bike racks, bikes, and free racks at each station to identify
     stations that have a consistently low or high number of bikes available. By clustering stations based on their usage
     patterns, the goal is to recommend adjustments to the bike distribution system to ensure a more balanced utilization of all
     stations and improve the overall efficiency of the bike sharing system.
```

Table 5: Datasets used in Experiment 2

Address Points	address points, their location, categories and affiliation to various functional areas of the city
Bike Stations	city bike stations, their location and current occupancy status
Election Districts	names of all election districts in the city
Env Decisions	case number, status and reference to parcel covered by the environmental decision
Graves	description of the graves located in the city cemetery
Graveyards	names of graveyards and API parameters which can be used to interact with services
Historic Buildings	address, ways of contact, description and classification of historic buildings in the city
MPK Lines	description of each public transport line available in the city
MPK Stops	public transport stops, their location, categories, types of vehicles and possible transfers
MPK Tickets	all locations where public transport tickets are sold including category and opening hours
Parking Machines	location of the machine and accepted payment methods
Polling Places	location, district name and number and accessibility for disabled people
Properties Sale	properties for sale including their location, parcel details, comments, contacts, type of offer and links to other services containing further information
School Districts	names and numbers of school districts, together with recruitment campaign codes
SD Plans	spatial development plans, references to official documents and dates indicating history of the plan development
SD Plans Progress	spatial development plans which are under development, references to official documents and date when work was initially started
Streets	different categories of places: streets, squares and others, including patron names and street length.
Wireless	locations and names of all wireless access points available in the city

Experiment 3: support in decision making

An example query: *As a statistician, I would like to analyze the frequency of occurrence of city residents' names and surnames. I am interested in identifying and tracking general trends over a long period of time and how names and usernames function in public space. Suggest the most relevant datasets from the context which can support this analytical task. List dataset names and justify the choice explaining why particular datasets are suitable for my use case.*

Model response: *Based on your analytical task of analyzing the frequency of occurrence of city residents' names and surnames over a long period of time to track trends and understand how names and surnames function in the public space, the most relevant datasets from the context provided would be the street dataset and the grave dataset. 1. Street Dataset: - Justification: The street dataset contains information about patron surnames and names associated with various streets in the city. By analyzing this dataset, you can track the frequency of occurrence of different surnames and names over time, providing insights into the historical and cultural significance of these names in the public space. This dataset can help you understand naming trends, patterns, and changes in the city's streets, reflecting the evolution of the urban environment and its inhabitants. 2. Grave Dataset: - Justification: The grave dataset can also be valuable for your analysis as it contains information about the names and surnames of deceased individuals buried in the city. By studying this dataset, you can track the frequency of occurrence of names and surnames associated with burial sites over time, providing insights into population demographics, historical naming conventions, and cultural practices related to death and remembrance. This dataset can complement the street dataset by offering a different perspective on the usage and significance of names and surnames in the public space. By combining insights from both the street dataset and the grave dataset, you can gain a comprehensive understanding of how names and surnames have been used, evolved, and functioned in the city's public space over an extended period of time. This analysis can help you identify trends, patterns, and correlations related to city residents' names and surnames, contributing to a deeper understanding of urban lifestyle and habits.*

```
{
  "city_bike_station": {
    "type": "object",
    "title": "city bike station",
    "service_url":
        "https://www.poznan.pl/mim/plan/map_service.html?
        mtype=pub_transport&co=stacje_rowerowe",
    "properties": {
      "geometry": {
        "type": "Point",
        "crs": "EPSG:4326",
        "coordinates": {
          "title": "coordinates",
          "type": "double",
          "description": "coordinates of exact point"
        }
      },
      "bike_racks": {
        "title": "bike racks",
        "type": "number",
        "description": "total number of bike racks available
                        at the station"
      },
      "bikes": {
        "title": "bikes",
        "type": "number",
        "description": "number of bikes parked at the station"
      },
      "free_racks": {
        "title": "free racks",
        "type": "number",
        "description": "number of free racks avaialable
                        at the station"
      },
      "label": {
        "title": "label",
        "type": "string",
        "description": "human readable name of the station"
      },
      "updated": {
        "title": "updated",
        "type": "timestamp",
        "description": "represents date when station
                        status was updated"
      }
    }
  }
}
```

Figure 2: Example of the *Bike Stations* dataset description query.