Harnessing the Power of LLMs to Construct Structured Datasets from Unstructured Sources

David Zarruk

Disclaimer: All views expressed are my own.

Empirical research stages

1. Dataset construction:
   - Access to raw files:
     - Available online
     - Administrative or bureaucratic process
     - Go to archives and find historical documents
   - Data processing:
     - Readily available excel files
     - Press releases, news articles, speeches
     - Historical archives (pdfs, physical paper, etc)
     - Manually-filled survey forms
2. Data analysis
3. Paper writing

Unavoidable

Easy processing

Can take a significant amount of time (50%? 90%?)

Difficult/slow to process

Some examples of costly data processing

- Papers that study impact of speeches/press releases:
  - The Importance of Fed Chair Speeches as a Monetary Policy Tool – Swanson (2023)
  - Shocking language: Understanding the macroeconomic effects of central bank communication – Hansen and McMahon (2016)

- Datasets based on historical archives:
  - The Colonial Origins of Comparative Development: An Empirical Investigation – Acemoglu, Johnson and Robinson (2001)
  - History, Institutions, and Economic Performance: The Legacy of Colonial Land Tenure Systems in India – Banerjee and Lakshmi (2005)

LLMs to the rescue

- LLMs excel at reading and understanding text and images
- Can easily transcribe images into text and extract relevant information
- Can analyze a text, find relevant parts, and analyze sentiments

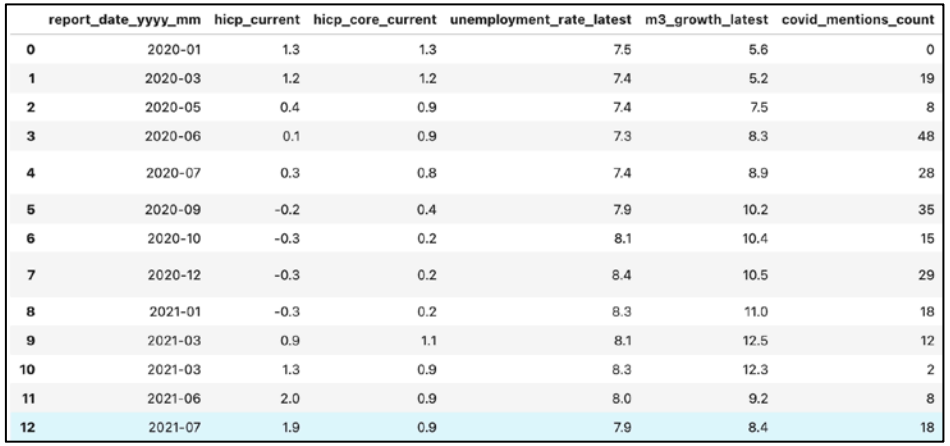
2 examples

1. Construct a dataset with European Central Bank bulletins
    - Hard data (inflation, unemployment, etc)
    - Qualitative features:
        - Risks
        - Labor market conditions: improving, deterioriating, etc.
        - Financial conditions: tightening, easing, etc

EUROPEAN CENTRAL BANK
EUROSYSTEM

Economic Bulletin

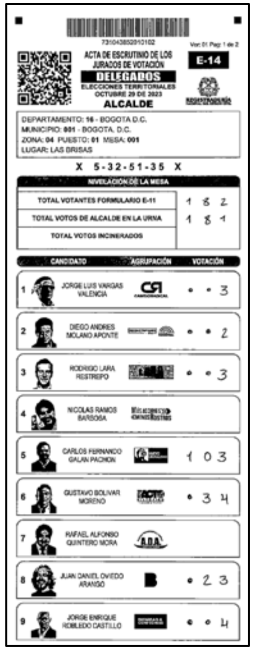
Issue 2 / 2020



	report_date_yyyy_mm	hicp_current	hicp_core_current	unemployment_rate_latest	m3_growth_latest	covid_mentions_count
0	2020-01	1.3	1.3	7.5	5.6	0
1	2020-03	1.2	1.2	7.4	5.2	19
2	2020-05	0.4	0.9	7.4	7.5	8
3	2020-06	0.1	0.9	7.3	8.3	48
4	2020-07	0.3	0.8	7.4	8.9	28
5	2020-09	-0.2	0.4	7.9	10.2	35
6	2020-10	-0.3	0.2	8.1	10.4	15
7	2020-12	-0.3	0.2	8.4	10.5	29
8	2021-01	-0.3	0.2	8.3	11.0	18
9	2021-03	0.9	1.1	8.1	12.5	12
10	2021-03	1.3	0.9	8.3	12.3	2
11	2021-06	2.0	0.9	8.0	9.2	8
12	2021-07	1.9	0.9	7.9	8.4	18

2 examples

2. Construct a dataset based on hand-written forms
   - Elections in Colombia



731043852010102

Ver: 01 Pag: 1 de 2

ACTA DE ESCRUTINIO DE LOS JURADOS DE VOTACIÓN

DELEGADOS

E-14

ELECCIONES TERRITORIALES
OCTUBRE 29 DE 2023

ALCALDE

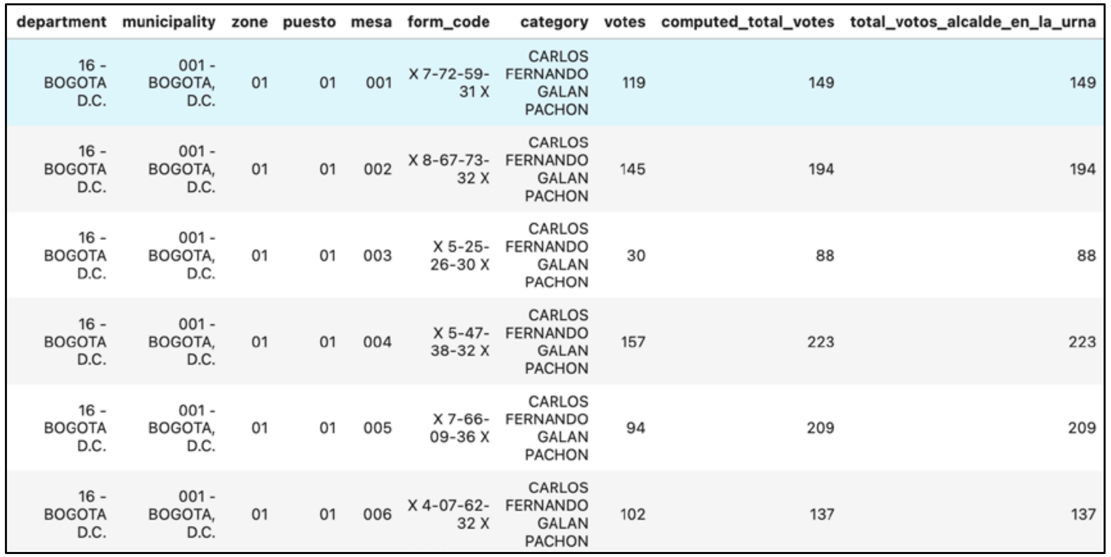
DEPARTAMENTO: 16 - BOGOTA D.C.
MUNICIPIO: 001 - BOGOTA, D.C.
ZONA: 04 PUESTO: 01 MESA: 001
LUGAR: LAS BRISAS

X 5-32-51-35 X

NIVELACIÓN DE LA MESA

TOTAL VOTANTES FORMULARIO E-11	1 8 2
TOTAL VOTOS DE ALCALDE EN LA URNA	1 8 1
TOTAL VOTOS INCINERADOS	

	CANDIDATO	AGRUPACIÓN	VOTACIÓN
1	JORGE LUIS VARGAS VALENCIA		• • 3
2	DIEGO ANDRES MOLANO APONTE		• • 2
3	RODRIGO LARA RESTREPO		• • 3
4	NICOLAS RAMOS BARBOSA		
5	CARLOS FERNANDO GALAN PACHON		1 0 3
6	GUSTAVO BOLIVAR MORENO		• 3 4
7	RAFAEL ALFONSO QUINTERO MORA		
8	JUAN DANIEL OVIEDO ARANGO		• 2 3
9	JORGE ENRIQUE ROBLEDO CASTILLO		• • 4



department	municipality	zone	puesto	mesa	form_code	category	votes	computed_total_votes	total_votos_alcalde_en_la_urna
16 - BOGOTA D.C.	001 - BOGOTA, D.C.	01	01	001	X 7-72-59-31 X	CARLOS FERNANDO GALAN PACHON	119	149	149
16 - BOGOTA D.C.	001 - BOGOTA, D.C.	01	01	002	X 8-67-73-32 X	CARLOS FERNANDO GALAN PACHON	145	194	194
16 - BOGOTA D.C.	001 - BOGOTA, D.C.	01	01	003	X 5-25-26-30 X	CARLOS FERNANDO GALAN PACHON	30	88	88
16 - BOGOTA D.C.	001 - BOGOTA, D.C.	01	01	004	X 5-47-38-32 X	CARLOS FERNANDO GALAN PACHON	157	223	223
16 - BOGOTA D.C.	001 - BOGOTA, D.C.	01	01	005	X 7-66-09-36 X	CARLOS FERNANDO GALAN PACHON	94	209	209
16 - BOGOTA D.C.	001 - BOGOTA, D.C.	01	01	006	X 4-07-62-32 X	CARLOS FERNANDO GALAN PACHON	102	137	137

Tips for good prompt engineering

1. Enclose sections in the prompt using <> keys
2. Define a set of rules that define boundaries for the LLM:
   - “Do not hallucinate”
   - “Return a JSON – nothing else”
3. Give examples to the LLM!
4. Ask the LLM to return citations or the thought process that support the output
5. ITERATE: writing a good prompt is an iterative process – usually the LLM never gets it right on the first attempt