

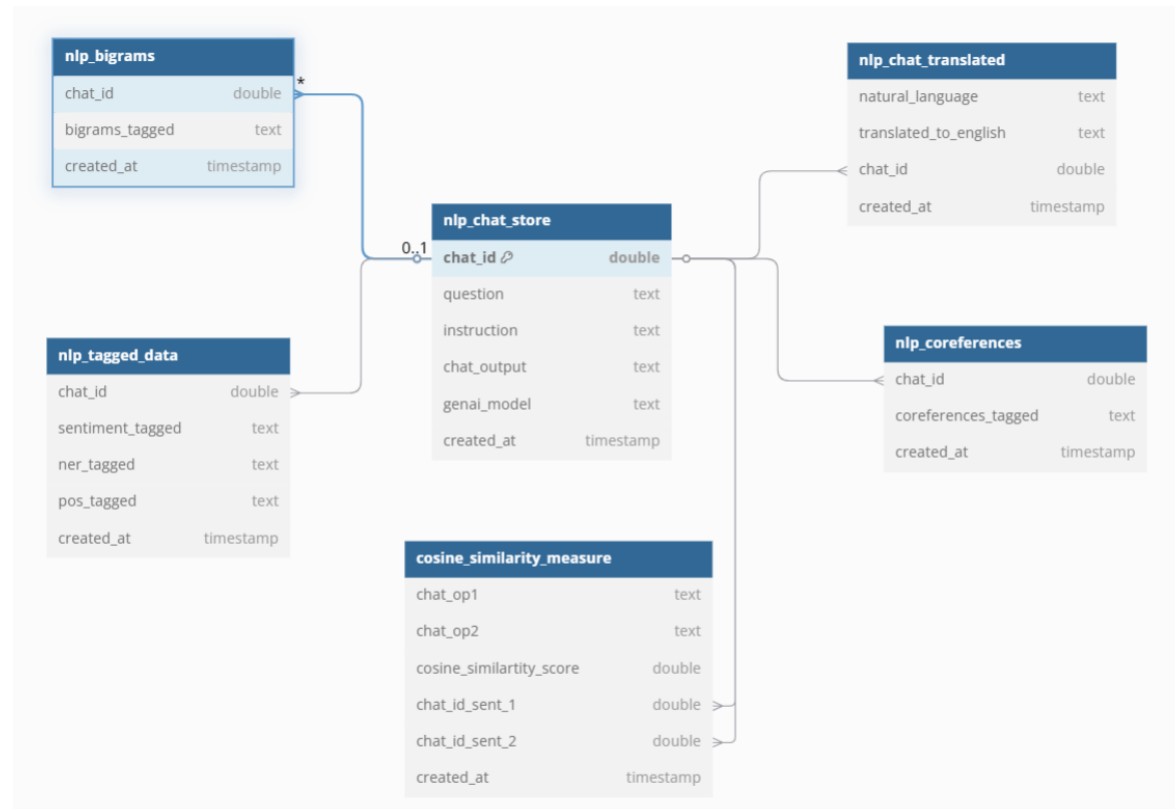
This document implements few NLP tasks using GenAI plugins step using Pentaho Data Integration (PDI) which calls OpenAI based LLM. The tests are done via “gpt-4o-mini” model.

Below is a snapshot of various Natural Language Processing (NLP) tasks are categorized as below. Few of these NLP tasks we will address using PDI and LLM (OpenAI).

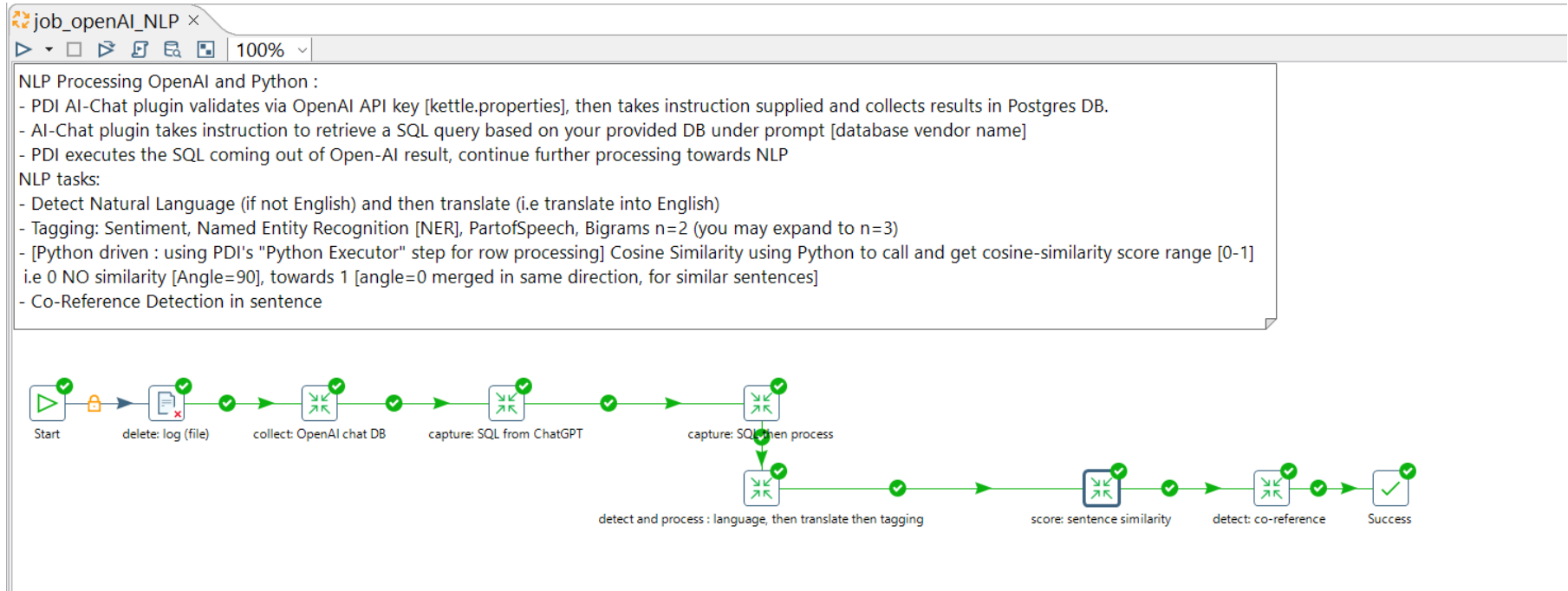


Data Model:

ERD: https://dbdiagram.io/d/ChatGpt_PDI-67c71518263d6cf9a030e12c



Pentaho Data Integration jobs and transformations:



#Generate Data (multiple natural languages) using OpenAI and store in a database for processing NLP

job_openAI_NLP ai_chat_plugin x

PDI using "AI Chat" step to take instruction and storing results in DB

100%

Data grid AI Chat sv: question_instruction_result as: chat_id to: open_ai_chat_store

Data grid

Step name Data grid

Meta Data

#	Question	Instruction
1	Tell 1 interesting fact about Norway and are there polar bears in Norway? Please reply in Englis...	You are a helpfu...
2	Tell 1 interesting fact about Massachusetts. Please reply in Hindi.	You are a helpfu...
3	Tell 1 interesting fact about last Superbowl. Please reply in German.	You are a helpfu...
4	Why polar bears are not found in southern hemisphere?	You are a helpfu...
5	Share a joke in German.	You are a helpfu...

Execution Re

Executing

Help OK Preview Cancel

Spoon - ai_chat_gpt_to_query_TableInput

File Edit View Action Tools Help



View Design

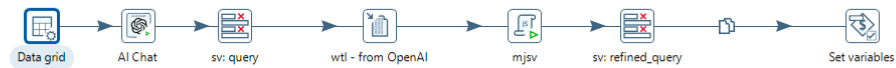
Search

- > Input
- > Output
- > Metadata Discovery
- > Streaming
- > Hierarchical
- > Transform
- > Utility
- > Flow
- > Scripting
- > Pentaho Server
- > Lookup
- > Joins
- > Data Warehouse
- > Validation
- > Statistics
- > Data Mining
- > Big Data
- > Cryptography
- > Job
- > Mapping
- > Bulk loading

job_openAI_NLP ai_chat_gpt_to_qu...

100%

PDI using Open AI API to get results and stored in DB



Data grid

Step name Data grid

Meta Data

#	Question	Instruction
1	select the most recent 5 records based on created_at colu...	Just return SQL in Postgres DB, without any extra characters in output, please.

Execution Re

Execution



Help

OK

Preview

Cancel

#Testing to see SQL works

Spoon - connect_PG

File Edit View Action Tools Help

Connect

View Design

Search

- > Input
- > Output
- > Metadata Discovery
- > Streaming
- > Hierarchical
- > Transform
- > Utility
- > Flow
- > Scripting
- > Pentaho Server
- > Lookup
- > Joins
- > Data Warehouse
- > Validation
- > Statistics
- > Data Mining
- > Big Data
- > Cryptography
- > Job
- > Mapping
- > Bulk loading
- > Inline
- > Experimental
- > Legacy
- > Deprecated
- > AI
- > PMI
- > History

job_openAI_NLP connect_PG x

gvr: REFINED_QUERY wtl: refined_query read ti: chat_gpt tfo: tmp_openai_querying_DB

Table input

Step name: ti: chat_gpt

Connection: openai_nlp

SQL

\$(REFINED_QUERY)

Get SQL select statement...

Execution Result

Execution I

#	Stepna...
1	

Line 1 Column 0

Store column info in step meta data ☐

Enable lazy conversion ☐

Replace variables in script? ☒

Insert data from step

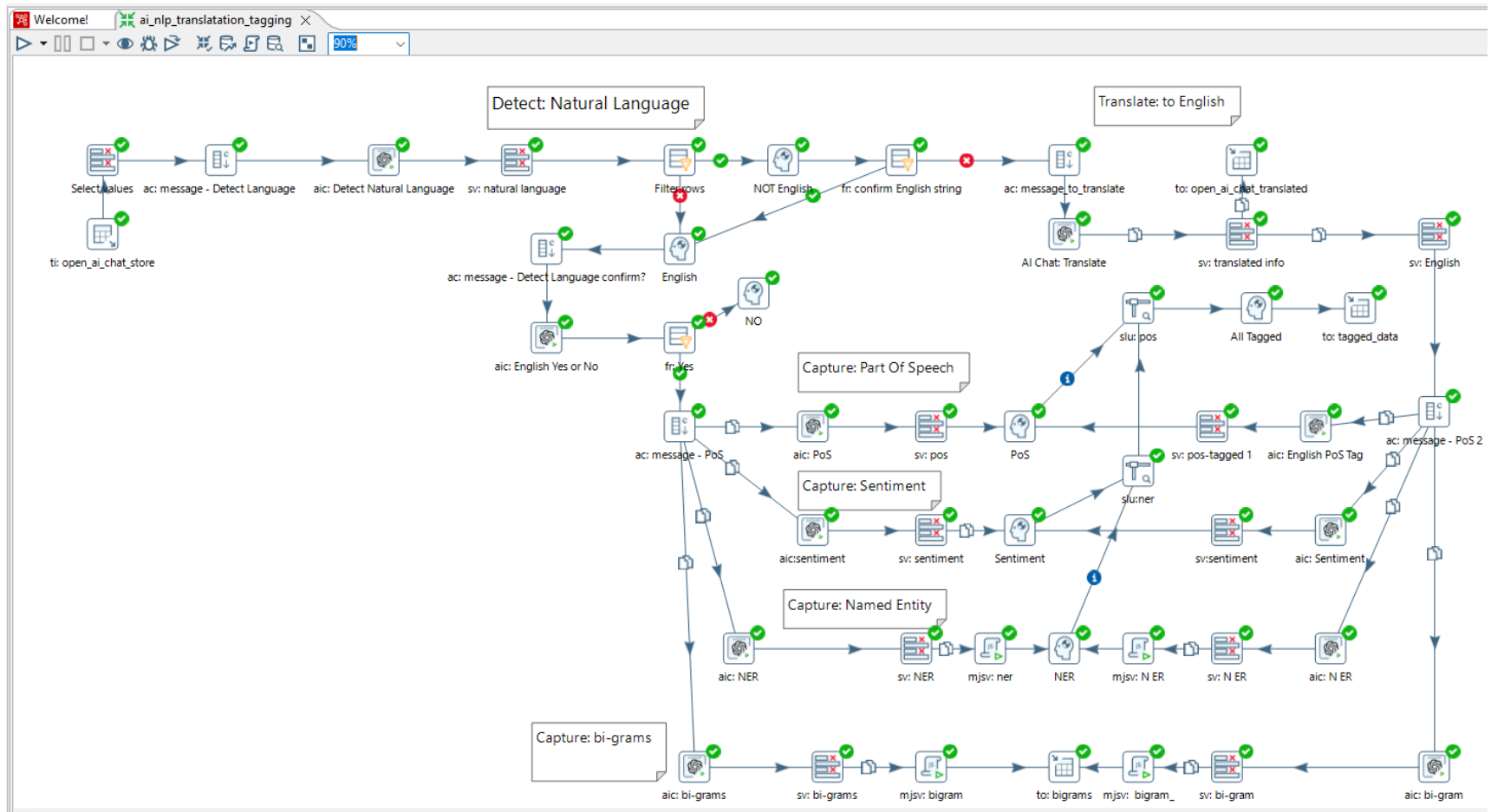
Execute for each row? ☐

Limit size: 0

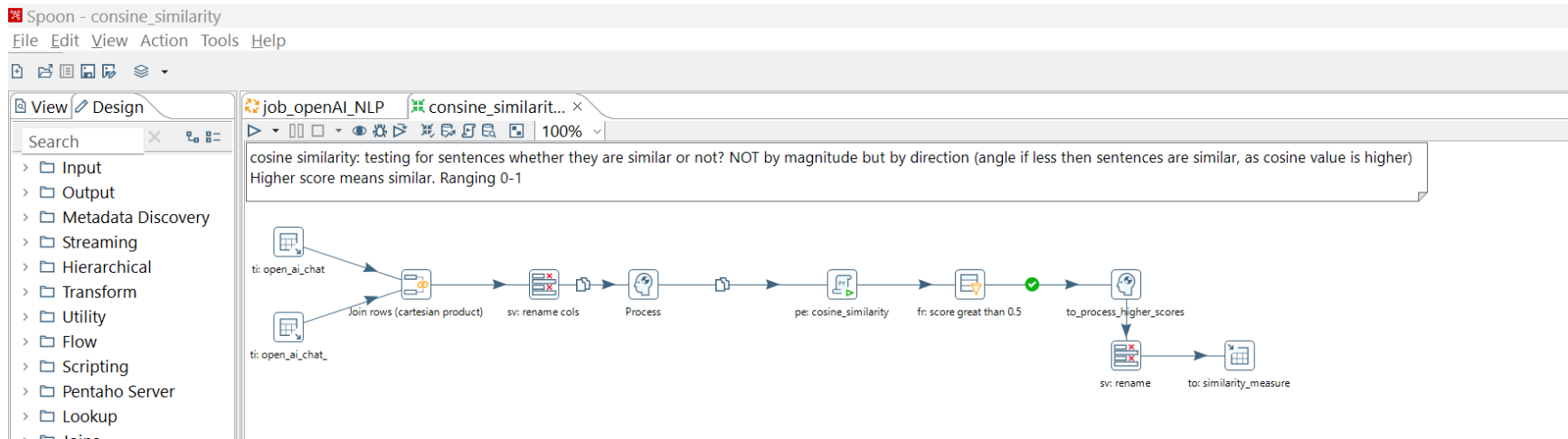
OK Preview Cancel

Help

#Translation, Sentiment, NER, bigrams, PoS



ChatGPT couldn't provide a score through prompt, so having this via Python while using data processing in PDI transformation (data in flight).



--Co-Reference

