

Overview

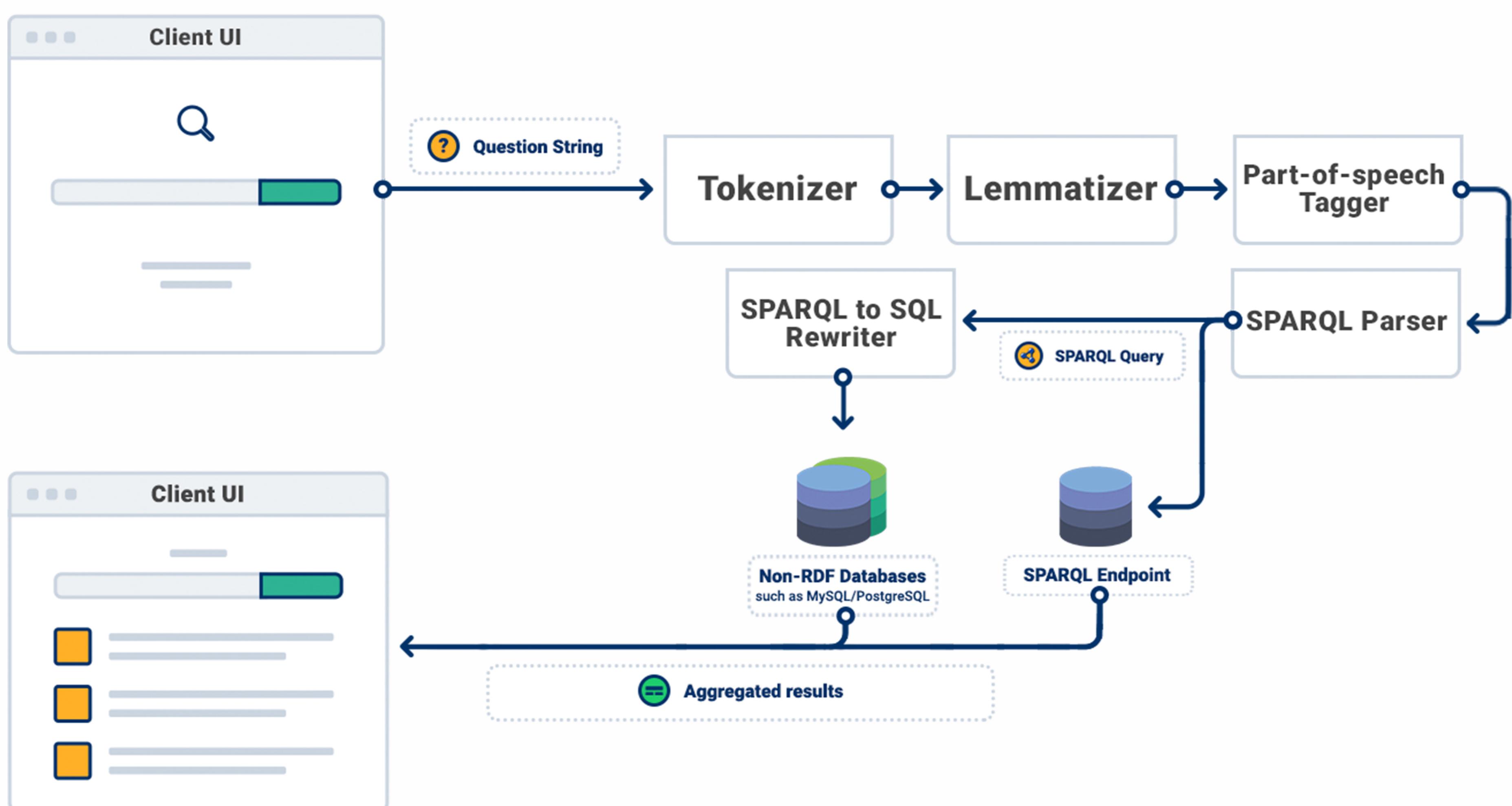
Organizations have access to vast amounts of data regarding their operations, either in a structured format such as spreadsheets and tables, or in an unstructured one such as documents. This data will be typically stored in detached entities, which makes day to day operations such as searching or filtering overly complicated.

Ideally, all available data should be part of a single coherent unit that is easy to query and interact with.

Objectives

- The primary objective is to develop a platform capable of receiving natural language questions from the user and returning data resulted from application of semantic-preserving queries to the provided datasets.
- The project should be able to parse a valid natural question into a valid SPARQL query for the provided SPARQL endpoint(s).
- The system should be able to query the provided non-RDF databases using SPARQL.
- A user interface system is to be provided for seamless integration with the underlying data.
- Optionally, the platform could support additional data formats such as CSV, and parse them into relational databases to be subsequently queried.

Implementation



Challenges

- Searching using the exact expression typed by the user:** The user should be provided with relevant results not only for their exact question string, but also for other possible forms of the words used.
- Queries cannot be parsed without semantics:** When translating from English to French, simply translating each word individually and then concatenating the results does not suffice. The same applies for SPARQL to SQL; parsing SPARQL (which queries an RDF graph) to SQL (which queries a relational database) requires an understanding of meaning behind the provided query and data.
- Natural Language Processing limitations apply:** The platform initially processes the user's question into an intermediate representation and matches it to a question regex before parsing it into a SPARQL query. Consequently, the project is limited by all the challenges applicable to state-of-the-art NLP projects.