Iceberg.ai – A platform for rapid development of legal and regulatory AI services

vLex.com

Abstract. In this work, Iceberg.ai is described. Iceberg.ai is a commercial data integration and AI augmentation platform specifically devised to ease the development of legal and regulatory AI application. This paper provides an overview of different services regarding Iceberg.ai: Input/Output, citation, keyword extraction, representation and deployment model. In addition, an example of the application of Iceberg.ai is also described: Vincent (short for Vlex insights). Vincent is a concrete application of the Iceberg.ai platform as a contextual search solution.

Keywords: Artificial Intelligence · Legal applications · Regulatory applications · Contextual search

1 Introduction

Iceberg.ai is a commercial data integration and AI augmentation platform designed to facilitate the development legal or regulatory AI applications. It also provides a data ingestion pipeline that includes the ability to harvest and parse data from the public web, upload structured private data and subscribe to curated regulatory data feeds from vLex.com [1] (an international legal research service). This combined data view can be enriched through pre-trained machine learning services focused on the legal and regulatory domain that provide citation discovery, legal document classification, topic extraction and recommendation services. Additionally, users can create custom machine learning models. Section 2 provides an overview of different services regarding the Iceber.ai platform.

On Section 3, Vincent [2], a specific application of Iceberg.ai, is described. It is a contextual search solution that analyses a legal document (such as a brief or a contract) and transforms it to a query that produces a list of relevant materials to it. The user can combine the generated query with additional keywords to obtain a set of search results that benefits both from the context provided by the document and the intent expressed by the keywords.

2 Overview of key Iceberg Services

2.1 Input/Output Services

Iceberg supports two input modes:

1. Import module: designed to ingest data from a variety of structured data formats (CSV, RDF, XML, SQL Databases, etc.) that can be transformed before ingestion with an in-place ETL pipeline. Imports can be scheduled to keep iceberg data view in sync with external systems or to subscribe to data feeds from vLex.com.
2. Crawl module: designed to harvest data from the public or private web. The crawl module breaks the harvesting task into three steps:
   (a) crawling, consisting of the actual retrieval of the original content by providing a list of seed URLs and a set of patterns that filter the URLs that should be followed. The output of crawling is stored into a deduplicated Crawl database.
   (b) parsing, which transforms the crawled data into a simple JSON-based data format with the assistance of file transformation services and custom data extractors
   (c) ingestion, where the output of the parsing is mapped to the organisations data model

Each step can be processed independently, and Iceberg keeps full provenance tracking of the origin of each data attribute. Likewise, an export module is available to export back the enriched data to CSV, XML RDF or SQL databases.

2.2 Citation Services

vCite is a legal citation service integrated within the Iceberg platform that can recognise citations to statutes, regulations, case law and administrative decisions of 16 jurisdictions[1] and resolve them to documents.

vCite combines a rules-based engine that produce candidate citation matches with a machine-learning model to resolve citations that contain some ambiguity. The disambiguation model takes into account the citation context and analyses the co-citations graph. vCite also weights the citations by the strength (citation strength is the models estimate of how significant is the cited document to the citation context discourse).

2.3 Representation services

Iceberg internal representation is a essentially a graphical model, inspired by RDF. Attributes of objects (such as documents or entities) and relationships are represented as triplets, which can be annotated with additional reified properties. This basic data representation is used as the central "repository of truth" from which additional representations can be generated.

Among those, representation services are of particular importance. Representation services bundle models that learn to transform objects (documents,

[1] Jurisdictions supported include full support for US (Federal and state), Brazil, Canada, India, Spain, México, Colombia, Chile, Argentina, Perú, Ecuador and UE Law, as well as more limited support for other jurisdictions like the UK.

events or entities) into dense or sparse vectors with computing resources that provide proximity queries on such representation spaces.

These proximity queries can be used as stepping stones to build more complex services. Concrete examples are:

- user-to-document or document-to-document recommendation engines,
- semantic search services,
- as an input to link disambiguation models
- as an element used within a graphical visualisation

Iceberg offers a pluggable facility that allows the user to pick representation generation algorithms, select the approximate nearest neighbourhood algorithm (to enable proximity queries) and define custom ranking formulas (to chose among different similarity measures, or combine the similarity ranking with other scoring factors)

Implementations are available to learn vectorial representations from the textual contents of the documents, from the structural properties of the relationships graph, from usage logs associated to objects (using matrix factorisation) or from a combination of those.

Among the models available, we want to remark that Iceberg provides a pre-trained deep learning model that has been trained to learn a link discovery goal from the contents of a 50-million documents multi-lingual legal collection. Such model has proved to be very useful on recommendation and link disambiguation tasks.

2.4 Keyword extraction service

Automated keyword (or keyphrase) extraction can be very useful to to assist on ontology management, supplement or substitute human summaries and enrich recommendations.

Iceberg provides a pre-trained keyword extraction service for English and Spanish languages that has been trained with over 200.000 tagged legal documents.

Iceberg keyword extractor is essentially an extractive summariser. A pre-processing part-of-speech tagger identifies candidate phrases, followed by a neural network that's trained to recognise the most relevant terms from a document taking into account features such as the term frequency, it's location and span, the term-to-document semantic similarity, the context of the term occurrences and additionally out-of-document features associated to the term.

Users of Iceberg that have access to pre-existing datasets of documents tagged with valuable keywords, can fine-tune (or train from scratch) a custom keyword extractor to better represent their data.

2.5 Deployment model

Although Iceberg is commonly deployed as a platform-as-service offering on the cloud; legal and regulatory data has quite frequently high privacy and confidentially requirements. Caretakers of such data commonly have a strong preference

(or a legal requirement) to reduce the quantity of data processors that have access to the raw data.

To support these use cases, Iceberg supports the deployment of an on-premises data extractor delivered as a Docker image (that can run locally on a standalone server or within a Kubernetes cluster). The service can work in a isolated network environment, pre-downloading the models and other supporting artifacts required to evaluate the different predictions.

3 Example application: Vincent

We now describe Vincent (short for vlex insights) [2], a concrete application of the Iceberg platform. As described on the Introduction, Vincent is a contextual search solution that analyses a legal document (such as a brief or a contract, or a passage within it) and transforms it to a query that produces a list of relevant materials to it. Vincent can suggest materials from vLexs Legal Research collection and from private data collections that have been indexed into a private index.

3.1 Pre-computed legal graph

Vincent relies on a pre-computed directed legal graph that includes nodes for primary materials (such as regulatory, legislative and case law) and secondary materials (such as practical notes, journal articles and legal forms). Edges between nodes are added when:

- A document cites another document
- Two documents are strongly topically related (semantic edges)
- Two documents are frequently cited together (co-citation edges)

Edges are weighted by the strength of the relationship between the nodes.

3.2 Document Analysis

At document analysis time, Vincent evaluates a pipeline with the following steps:

1. File reception (documents can be uploaded by a web interface o directly analysed from a Microsoft Word plug-in)
2. File format transformation into text
3. Temporary insertion of the received document as a node within the legal graph by:
   (a) Insertion of citation edges (after citation extraction and disambiguation)
   (b) Insertion of semantic edges (by producing both a dense and sparse vector representation and evaluating proximity searches)
4. Exploration of the neighbourhood of the received document on the graph to obtain a list of strongly connected documents, which will be included on the result set

Additionally, a keyphrase extractor processes the received text and obtains a list of topics that are likely to describe the document content. Vincent then attempts to select among those the sublist of keyphrases that, when interpreted as joint searches, produce results that are strongly connected to the document. In other words: Vincent goal at this stage is to generate a list of search terms that serve as a good generalisation of the graph exploration task.

As a result of the analysis stage, Vincent obtains a list of topics, a placement of the received document into the legal graph and a ranked list of highly connected nodes. The contents of the document are then discarded (Vincent does not store at any point the original document both for privacy reasons and to reduce the security concerns related to keeping potentially highly confidential information).

3.3 Query Time Evaluation

At query time (which is usually performed immediately after document analysis) Vincent transforms the ranked list of connected nodes and search terms into a joint search thats then processed by a conventional full-text search engine. The obtained results are ranked according to a combination of their Vincent relevancy score and their pre-existing authority score.

The user can then explore the search results as it if were produced by a conventional keyword search, applying additional filters (such as date, material type, among others) or refining with additional search terms.

On addition to conventional filters, the user is provided with a Vincent-specific search filter ("Direct Vincent") that can be used to restrict the type of connections that Vincent takes into account. For instance, the user can decide to ignore materials already cited on the sources or, on the contrary, to ignore semantic analysis and only focus on proximity on the citation graph.

Results are presented with a summary explanation of what motivates them being recommended (for instance: describing that an item is recommended because its frequently cited together with one of the source's cited authorities). We've found this explanations very useful on building trust on the users and helping them focus on the the highest quality recommendations.

3.4 Alerting

A particularly relevant use case of Vincent is associated to alerting. Although the contents of the analysed document are not stored, the output of the analysis is saved and users can create alerts queries that monitor new regulatory developments that are relevant to the analysed document.

References

1. vLex.com Home Page, http://vlex.com. Last accessed 4 January 2019
2. Introducing Vincent: the first intelligent legal research assistant of its kind, https://blog.vlex.com/introducing-vincent-the-first-intelligent-legal-research-assistant-of-its-kind-bf14b00a3152. Last accessed 5 February 2019