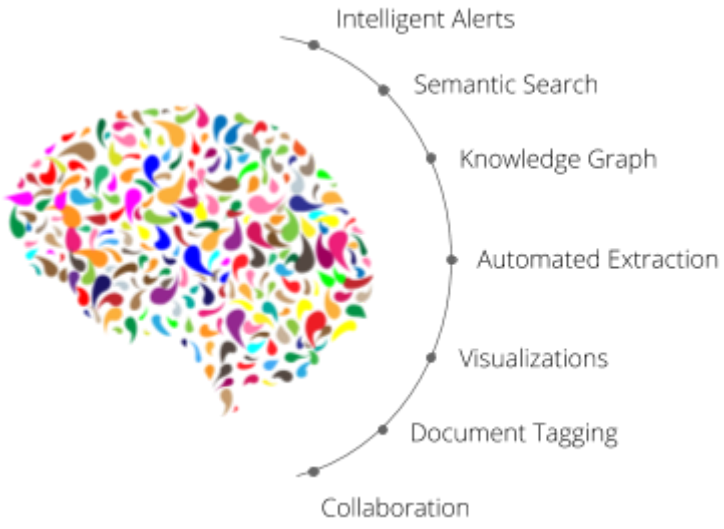
Salient

Machine Learning

Salient combines different state-of-the-art machine learning models, including several proprietary models developed by Lore, in order to quickly adapt to the workflow of different business users. Ingested documents are analysed extensively using both supervised methods (algorithms trained on labeled data) as well as unsupervised methods (algorithms that learn directly from data without requiring any labeling), providing a layer of intelligence that makes it easy for Salient to absorb additional knowledge with a minimal effort by the user.

Natural Language Processing (NLP)

All documents ingested in Salient are processed using a full NLP pipeline which includes extracting grammatical information and using it to identify entities (i.e. people, places, companies, etc). These entities are recognized by grammatical context so even unknown entities are tagged. This information forms the basis for much of the subsequent intelligence in the system.



Knowledge Graph (KG)

Salient incorporates a user-modifiable knowledge graph. A knowledge graph is a database of entities & concepts, along with their properties and relationships. Salient comes pre-loaded with ~6 million entities from Wikipedia. Any new entities identified in your documents are automatically added to the knowledge graph. It is also possible to upload additional entities representing internal organizational knowledge (e.g. employee or customer data, etc.). All documents are cross-referenced with the KG and every entity in the knowledge graph has a page where all its information is summarized.

Embedding Space

Salient uses a technique called "embedding" to represent all objects (documents, words, entities, etc.) as mathematical "vectors" in such a way that relationships between them become mathematical operations between vectors. This is an unsupervised technique that can be used to learn relationships directly from data without any user labeling. Such techniques expose important relationships between words, concepts and even documents without requiring any human intervention. The table below shows examples of similar concepts learned by a particular embedding space algorithm (word2vec) when analysing a large amount of text:

FRANCE	JESUS	XBOX	REDDISH	SCRATCHED	MEGABITS
AUSTRIA	GOD	AMIGA	GREENISH	NAILED	OCTETS
BELGIUM	SATI	PLAYSTATION	BLUISH	SMASHED	MB/S
GERMANY	CHRIST	MSX	PINKISH	PUNCHED	BIT/S
ITALY	SATAN	IPOD	PURPLISH	POPPED	BAUD
GREECE	KALI	SEGA	BROWNISH	CRIMPED	CARATS
SWEDEN	INDRA	PSNUMBER	GREYISH	SCRAPED	KBIT/S
NORWAY	VISHNU	HD	GRAYISH	SCREWED	MEGAHERTZ
EUROPE	ANANDA	DREAMCAST	WHITISH	SECTIONED	MEGAPIXELS
HUNGARY	PARVATI	GEFORCE	SILVERY	SLASHED	GBIT/S
SWITZERLAND	GRACE	CAPCOM	YELLOWISH	RIPPED	AMPERES

Embeddings are used throughout Salient to learn about the structure of the ingested data, be it similarities between entities in the text or in the knowledge graph, or between words, sentences or even whole documents.

Salient

Machine Learning

Clustering & Visualization

In addition to Machine Learning, Salient makes extensive use of visualizations such as:

- Relevance weighted word-clouds
- Map overlays
- Relationship graphs

These are combined with unsupervised clustering techniques that can group and characterize similar documents, words, and entities to give a quick visual overview.

Semantic Search

Salient includes a very sophisticated search that combines all the intelligence it extracts from the documents with a state-of-the-art search engine. It includes the ability to search using multiple terms, to exclude terms, to facet (i.e. refine) based on document metadata and other extracted intelligence (presence of entities, etc.). The advanced search also allows arbitrary filters to be combined using boolean logic (AND/OR, etc) and applied to the search.

Most of the intelligence of the system can be leveraged inside a search. For instance, it is possible to combine extracted grammatical knowledge (e.g. which sentences mention dates or monetary amounts), with unsupervised statistical knowledge (e.g. which documents mention a company like “deloitte”), and then apply a variety of different ranking algorithms (similarity to a word or concept or more standard word-frequency algorithms).

Smart Highlighters

Smart highlighters are at the core of Salient’s learning metaphor. They learn what users are looking for by example and can then be applied to scale up that knowledge. Highlighters are driven by proprietary neural networks developed by Lore. They combine the various unsupervised signals described above with supervised but very efficient models that have been designed to be highly “data efficient” (i.e. to learn quickly from small amounts of data).

Note that, in addition to highlighting specific sentences, Highlighters also allow users to select specific types of entities or other tagged data (e.g. monetary amounts, dates, etc.) from the sentences. These can be extracted in an excel report or can be used to generate metadata for the associated document or paragraph. For instance, the jurisdiction of a contract is a location that can be extracted from the relevant sentences and added as a metadata field to automatically build a contract database from unlabeled PDFs.

Patterns

Patterns are a powerful data extraction tool. They provide direct access to Salient’s accumulated knowledge inside a sophisticated pattern matching algorithm. Using patterns like “{ENT.ORGANIZATION} hired {ENT.PERSON}” users can very quickly extract intelligence from documents and convert it to tabular data (a list of companies and the name of the people they hired). This allows domain experts to interactively and quickly collect information by combining their specialized knowledge of the documents with all the intelligence Salient has already extracted.