Interface 2015

NLP Entity Analytics and Logo Recognition in the Cloud

Military and Commercial Use Cases

Jack H. Davenport

Solving complex problems... in a manner judged to be truly superior
Agenda

• Overview of Data Types and Analytics
• Cloud Architectural Overview
• Use Cases
  – NLP Entity Analytics
  – Image/Video Analytics
• Q & A
Data Types and Analytics

• Text Documents
  – Military: M3, CIDNE, OSC, blogs
  – Commercial: Financial filings, news articles, blogs

• Video & Imagery
  – Military: Surveillance, handheld, collected in expeditionary warfare setting
  – Commercial: Broadcast news, sitcom/drama, websites

• NLP Entity Analytics
  – Extract & resolve people and companies
  – Extract relationships

• Logo Recognition and Text Extraction
  – Identify “graffiti” in urban areas
  – Brand coverage and brand placement
Architectural Overview

External Data Sources

Data Feeds for NRT/Streaming Ingest

Streaming Ingest
Data Processing Framework (DPF)
Artifact Persistence

Real Time Advanced Analytic Pipeline (RTAAP)
Streaming Analytics using Storm Framework
Processing Structured and Unstructured Data

Bulk Ingest Map Reduce Framework
On-Demand Ingest MR Jobs

Data Cloud

Custom Analytics Tables
Common Tables

Bulk Data Exports for Initial Load
and Occasional Batch Update

Analytics Map Reduce Framework
Periodic Analytic MR Jobs

UX Ozone Widget Framework (OWF-UX)

User Authentication and Authorization

Web Service Applications

Accumulo

JBoss

Ozone
Streaming Ingest and Analytics

**External Data Sources**

- Data Feeds for NRT/Streaming Ingest

**Streaming Ingest**

- Data Processing Framework (DPF)
- Artifact Persistence

**Real Time Advanced Analytic Pipeline (RTAAP)**

- Streaming Analytics using Storm Framework
- Processing Structured and Unstructured Data

**Data Cloud**

- Custom Analytics Tables
- Common Tables

- Accumulo

**Bulk Ingest Map Reduce Framework**

- On-Demand Ingest MR Jobs

**Analytics Map Reduce Framework**

- Periodic Analytic MR Jobs

**Web Service Applications**

- User Authentication and Authorization
- UX Ozone Widget Framework (OWF-UX)

**Ingestion and Analytics**

- Bulk Data Exports for Initial Load and Occasional Batch Update
NLP Entity Analytics

• Extract people, organizations, and locations from unstructured text.
  – People are most interesting for our DoD clients.
  – Organizations (corporations) are most interesting for our commercial clients.

• Extract relationships between entities.
  – Map relationship types into the existing relationship ontologies.
  – This mapping requires rich relationships instead of generic relationships such as associatedWith.
President Hamed Karzai ran for office. He won the election. President Karzai traveled to Kabul. Karzai met with the US President.

1, 2, 4 are extracted. 2 and 3 are resolved. 1-4 are resolved.

Many entity management tools require analysts to manual tag entities at this stage. DAC’s entity resolution capability does not require analyst action until this stage.

Based on Biometrics and other structured data sources.
Through Amir Bajric, Bajro Ikanovic purchased from Senad Hasanovic over 19 kilos of explosives.

Flores Cena said he killed Sanchez Rivera.

Chavez addresses his comments to Bolivian Armed Forces Chief Luis Trigo.

Qa'ida leader Mustafa Sitmarin, arrested in Pakistan, is in CIA custody.
SRL Pipeline

Raw Data
(text, transcribed voice, …)

Tokenization
Break text into word/punctuation tokens

Sentence Boundary Detection
Sentence boundaries identified statistically

Part-Of-Speech Tagging
Generates parse trees

Lemmatization
Automated training builds probabilistic model of role assignment

Word Sense Disambiguation
Language model based on FrameNet

Semantic Role Labeling

semantic model
Mapping into Intel Workflow

• Entity Relationship Extraction
  – Extracted entities are mapped into semantic roles.
  – When two entities in the same sentence are mapped to semantic roles, then a rich relationship is extracted.

• Entity Management tools in the intel space typically have a relationship/attribute ontology.
  – Employed_by, Lives_with, Arrested, Communicates_with, Killed_by, Purchased_from

• These relationships and attributes are typically populated by humans and from structured databases which have been previously vetted (e.g., biometrics databases).
Intel Entity Mgt

Entity Manager GUI

Vetted Entities & Relationships
Analyst creates & updates entities

Structured Datasets

Can NLP extract enough structure to enhance the definition of entities?

Unstructured Text
Integrate with Entity Workflow

Entity Manager GUI

Vetted Entities & Relationships

Structured Datasets

API

Analyst validates and exports to Entity Manager

NLP Entity & Relationship GUI

NLP Entities & Relationships

Entity & Relationship Extraction

Unstructured Text
The semantic relationships extracted with the breadth of FrameNet were down-selected and mapped to the relationship types of an entity manager used in an Army system.

The analyst validates and sends the relationship to the vetted entity manager.
Through Amir Bajric, Bajro Ikanovic purchased from Senad Hasanovic over 19 kilos of explosives.

Flores Cena said he killed Sanchez Rivera.

Chavez addresses his comments to Bolivian Armed Forces Chief Luis Trigo.

Qa'ida leader Mustafa Sitmarin, arrested in Pakistan, is in CIA custody.
• Entity Management tools in the financial services space also have a relationship/attribute ontology.
  – Competitor, Customer, Supplier, Investor

• These relationships and attributes are typically populated by humans who read periodically published semi-structured reports.

• However, news articles contain breaking business intelligence that will not be published for months in the semi-structured reports.
Fuse NLP Results with Financial Datastore

Platform Equinix provides Front Porch Digital with the network connectivity and availability demanded by customers.

Hudson Energy Solar Inc. is a subsidiary of Just Energy Group Inc., a publicly traded company (NYSE:JE and TSX:JE).

Elan's investment in Newbridge has been recorded as an equity method investment on the balance sheet.
• Recognize logos
  – Commercial applications: advertising coverage, brand monitoring, copyright infringement

• Recognize images
  – Commercial applications: brand monitoring, copyright infringement
  – Military applications: brand coverage in urban areas

• Text Extraction
  – Commercial applications: search enhancement, brand monitoring in news
  – Military applications: find identifiers on vehicles and equipment
Validating Sponsorship in Sporting Events
PR firm represents a landmark. Brand monitoring.
Exemplar Search

Analyst provides an exemplar image. →

System finds matches.
Text in video and imagery is searchable.
Expeditionary warfare and urban settings.
Multi-modal Correlation

Textual ID extracted from image. Can be correlated to a structured DB and unstructured text mentions of vessel.