UNSTRUCTURED DATA EXTRACTION VIA NATURAL LANGUAGE PROCESSING (NLP)

Presented by Alex Wu, Partner, Sagence, Inc.

2nd Annual INFORMS Midwest Practice of Analytics Conference
University of Chicago’s Gleacher Center
Agenda

Intro
  – About Sagence

Why NLP?

What is NLP?

Use cases & patterns

Q&A
Sagence is a specialized firm designed to help organizations drive their businesses with information and insight through better data and analytics practices.

**Our Value**

- **Improving Decision Making & Innovation:**
  Helping clients create better business outcomes by using their data more effectively

- **Asking the Right Questions:**
  Identifying the high impact analytics for a client's combination of industry, marketplace, and strategy

- **Applying Decision Science:**
  Generating hypotheses, collecting data, analyzing, testing, and informing new decisions

- **Developing Information Assets:**
  Identifying, unleashing, delivering and maintaining the hidden value in data

- **Delivering Innovative Solutions:**
  Helping clients with pragmatic innovation that acknowledges business, technical, economic, and cultural constraints

**Our People**

- Multi-disciplinary teams
- Led by professionals with 25+ years of experience
- Providing objective thinking combined with an interdisciplinary perspective
- Working collaboratively with clients
- Focused on business results

**Our Position**

- **Services**
  Data Specialists
  Thinkers and Doers
  Strategic and Objective

---

**We assist companies with everything from broad strategic challenges to tackling point problems with expert resources and solutions**
Quick Survey on Natural Language Processing (NLP)

Where would you rate your NLP knowledge?

- **NEW TO NLP**
  Start with the fundamentals

- **FAMILIAR WITH NLP**
  Experience with tools and techniques

- **SUBJECT MATTER EXPERT**
  In-depth discussions
This presentation represents a solution to solving a commercial problem.

**WHAT WE WILL COVER**

- Commercial problem and context
- Survey of frameworks/tools
- Select challenges encountered along with implemented solutions

**WHAT WE WON’T**

- Discussion on pros/cons of theory
- Detailed analysis of solutions
- In-depth discussion on topics in Natural Language Processing
Agenda

Intro

Why NLP?
- Growth of unstructured data
- Existing solutions & inefficiencies
- Value proposition

What is NLP?

Use cases & patterns

Q&A
The rapid growth of unstructured data magnifies the importance of finding the business value in that unstructured information.

Growth of Unstructured Data

61.8% CAGR for unstructured data

23.7% CAGR for structured data
Why NLP?
Utilizing Natural Language Processing (NLP) as an approach to distilling business value from unstructured data yields many benefits.
**Why NLP?**
Existing solutions largely fail at addressing the challenge of distilling business value from unstructured data.

**Unstructured Data Management Landscape**
- Fragmented data management solutions
  - Leads to Complex tool landscape & data quality issues
  - Leads to Operational inefficiencies & inability to respond to emerging business needs

**Continuum of Techniques**
- Manual Entry
- Manual Data Entry
- Template Driven
- Advanced Search
- Natural Language Processing (NLP)
- Highly Automated
Why NLP?
In one client example, utilizing NLP realized large efficiency gains compared to existing manual processes.

<table>
<thead>
<tr>
<th>Speed (minutes)</th>
<th>Accuracy (%)</th>
<th>Consistency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLP 240</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Manual 2</td>
<td>90%</td>
<td>LOW</td>
</tr>
</tbody>
</table>

Accuracy - Extraction: NLP 100% vs Manual 90%
Accuracy - Classification: NLP 80% vs Manual 90%
Agenda

Intro

Why NLP?

What is NLP?
  - Definition
  - Approach
  - Tools & Landscape

Use cases & patterns

Q&A
What is NLP?

Some Definitions

“By speech and language processing, we have in mind those computational techniques that process spoken and written human language, as language.” ~ Daniel Jurafsky and James H. Martin

- Phonetics and Phonology – The study of linguistic sounds.
- Morphology – The study of the meaningful components of words.
- Syntax – The study of the structural relationships between words.
- Semantics – The study of meaning.
- Pragmatics – The study of how language is used to accomplish goals.
- Discourse – The study of linguistic units larger than a single utterance.

Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition
**What is NLP?**

**NLP Steps**

**Pre-processing**
- Convert content to images and use OCR to convert images to text
- Utilize proprietary algorithms to adjust for artifacts and discrepancies
- Differentiate discourse from tabular structures and identify financial table boundaries

**Syntax**
- Tokenize text, assign X,Y coordinate structure to each token and assemble as in-memory data grid as reference
- Perform Part-of-Speech (POS) tagging and assemble Treebank (grammatical relationships of tokens within a sentence)
- Traverse data grid and assemble ‘sentence’ structure for each data point

**Semantics**
- Normalize extracted terms to key concepts
- Apply Named Entity Recognition (NER) to ‘sentence’ to classify tokens
- Use Semantic Vector rules (SV rules) to identify key mnemonics

**Pragmatics**
- Assign semantic Predicate/Subject/Object (PSO) relationship to relate data point (Money) with surrounding context (Legal Entity, Statement Type, Statement Date, Line Item, Fund…etc.)
The landscape consists of both open source and commercial solutions. No single solution addresses the full data lifecycle.
Agenda

Intro

Why NLP?

What is NLP?

Use cases & patterns
  – Use cases
  – Solution

Q&A
The solution serves as the enterprise data collection and management pipeline into the ratings analytical process.

- Global credit ratings agency
- 450 person organization dedicated to data collection and governance
- Automation of data collections results in
- ~$20 million in savings over five years

Data Lifecycle for Financial Data Extraction
Document artifacts require preprocessing prior to NLP processing

Table spans across multiple pages

Complex nested line items and subsections

Poorly scanned documents
The solution must detect tables, identify boundaries, parse tokens and assign coordinate metadata.
Targeted extraction of data points within table structures requires preservation of data lineage and context. By combining the Row Context + Value + Column Context, a ‘sentence’ is formed for NLP.
Financial data extraction requires precise targeting of content with related context and data lineage

- Automatically crawl and retrieve documents
- Extract and classify key mnemonics with preservation of context and lineage

### Table 2 - Changes in Net Position (dollars in thousands)

<table>
<thead>
<tr>
<th></th>
<th>2013</th>
<th>2012</th>
<th>2013</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Program revenues:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Charges for services</td>
<td>$37,438</td>
<td>$41,823</td>
<td>$49,603</td>
<td>$49,591</td>
</tr>
<tr>
<td>Operating grants &amp; contribs</td>
<td>5,824</td>
<td>6,911</td>
<td>138</td>
<td>181</td>
</tr>
<tr>
<td>Capital grants &amp; contribs</td>
<td>7,660</td>
<td>9,260</td>
<td>6,498</td>
<td>5,175</td>
</tr>
<tr>
<td><strong>General revenue:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sales &amp; use taxes</td>
<td>104,151</td>
<td>108,771</td>
<td>118</td>
<td>119</td>
</tr>
<tr>
<td>Property taxes</td>
<td>29,434</td>
<td>1,976</td>
<td>1,952</td>
<td></td>
</tr>
<tr>
<td>Other taxes</td>
<td>21,184</td>
<td>20,964</td>
<td>904</td>
<td>836</td>
</tr>
<tr>
<td>Interest &amp; investment</td>
<td>755</td>
<td>1,051</td>
<td>301</td>
<td>384</td>
</tr>
<tr>
<td>Miscellaneous revenue</td>
<td>1,678</td>
<td>1,645</td>
<td>211</td>
<td>974</td>
</tr>
<tr>
<td>Gain on sale of capital</td>
<td>3,878</td>
<td>2,173</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Total Revenues</strong></td>
<td>211,957</td>
<td>210,014</td>
<td>56,239,612</td>
<td>58,212</td>
</tr>
<tr>
<td></td>
<td>271,586</td>
<td>269,226</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Program expenses (includes indirect expenses allocation)

| Governmental activities: |       |       |       |       |
| General Government       |       |       |       |       |
| Administrative Services  |       |       |       |       |
| Public Safety            |       |       |       |       |
| Public Works             |       |       |       |       |
| Planning & Development   |       |       |       |       |
| Culture & Recreation     |       |       |       |       |
| Open Space & Mountain    |       |       |       |       |
| Housing & Human Services |       |       |       |       |
| Interest on long-term    |       |       |       |       |
| Business-type activities:|       |       |       |       |
| Water Utility            |       |       |       |       |
| Wastewater Utility       |       |       |       |       |
| Stormwater & Flood       |       |       |       |       |
| Parking Services         |       |       |       |       |
| Property & Facility      |       |       |       |       |
| Total expenses           |       |       |       |       |

Excess before extraordinary transfers

- Extraordinary items
- Transfers
- Increase in net position
- Net position, beginning of year
- Restatement for change in accounting principle
- Net position, beginning of year, as restated

**Extraction Results**
Other use cases include legal discovery and document data mining.

- ~$500 trillion dollars in notional amounts outstanding
- Extract and classify key clauses based on NLP

http://www.isda.org/statistics/historical.html
Other use cases include legal discovery and document data mining

- Automatically crawl and retrieve online postings
- Extract and classify demand based on machine learning
A solution that addresses the full data lifecycle consists of four primary components:

**Sourcing**
- Crawl
- Upload
- Feeds

**Extraction**
- Preprocessing
- NLP Processing
- Post Processing

**Workflow & QA**
- Assignment
- Review

**Distribution**
- File
- Database
The Sagence solution offers end-to-end integration with enterprise systems to allow data usability and exchange.

**USE CASES & PATTERNS**

**BUSINESS CAPABILITIES:**
- End-to-end workflow
- Handle multiple extraction use cases within a single document
- Configurable user dictionary
- Capture document context using metadata tagging
- User configurable rules and knowledge base
- Multiple data sourcing options (Web, FTP, internal feeds, manual upload)
- Multiple document types (PDF, scanned documents)
- Pre-extraction document data cleanup and adjustment
- User dashboard to manage workflow and ensure quality assurance
- Data distribution to key systems and consumers
- Stores documents within content management system

**TECHNICAL FEATURES:**
- Modular architecture
- Open source components
- Extensibility to additional analytical tools
- Integrated data sourcing and retrieval (Web, FTP)
- Integration with reference data systems and custom feeds
- Intelligent algorithms for data preprocessing
- Distribution via XML or data services
- Customizable load balancing and work routing
- Conforms with enterprise standards
- Deployable via custom environments, virtual machines, or cloud-based platforms

---

**Diagram Description:**

- **Document**
  - File Mount
  - Web/FTP Sourcing
  - Backend Engine (JAVA + Other Open Source Components)
  - Front End UI
  - NLP Engine
  - Data Services Layer
  - Entity Referential Data Services
  - Document Repository Services
  - Entity Data Store
  - Document Repository
  - Relational Database
  - XML
  - FTP
  - Target Distribution System

**Legend:**
- NLP Core Product
- Enterprise System
THANK YOU

Alex Wu | Partner | Sagence
mobile: +1.860.680.7913
alex.wu@sagenceconsulting.com

Roger Moore | Senior Principal | Sagence
mobile: +1.312.543.1319
roger.moore@sagenceconsulting.com

http://info.sagenceconsulting.com/natural-language-processing-isda
The NLP engine utilizes out-of-the-box as well as user defined lexicons and rules to identify, classify, and extract targeted values.

http://rosoka.com/home/products/rosoa/rosoa-nlp/