Uncover new insights in your data

NLP and Knowledge Graphs for Financial Markets

27 February 2018
London | Imperial College
Roundtable: NLP Machine-Learning - understanding practical applications in finance
Edin Zajmovic, Salman Jaffer & Yair Lieberthal
Importance of entity masters for high precision & recall
Entity extraction & resolution – why does it matter?

Typical Problems

- Legal Entity Identifiers & Counterparty Static and Ref Data
- ISINs, CUSIPs, SEDOLs, RICs and FIGIs
- Millions of companies, currencies, stocks and ETFs
- Connecting to industry standard taxonomies; TRBC and TRCS
- Connecting to industry standard ontology, TR Knowledge Graph
- Resolving instrument pricing curve data for Rates, FX, Credit Spreads and Equity Derivatives

Some available solutions for tagging

- CSE - Closed Set Extraction or Known Entity Extraction is used. By providing a list of People, Organizations, Instruments, Topics, Events and Relations that we are looking for in long and short news snippets, research reports and emails and tagging
- Lexicons - Keyword sand Aliases are used. Not only providing the words we are looking for but also creating lexicons of synonyms, AKAs, FKAs, tense sensitive instances using techniques such as stemming and lemmatization, word embedding and vectorization
- Signatures - Contextual Recognition. Using a technique similar to word embedding and n-gram tokenized strings to identify, with a statistical confidence interval, the words most likely to precede and succeed the words we are looking for in news snippets and research reports
Creation of an NLP solution - where is the time spent?

- Creation of Corpus: 15%
- Training Set Labelling: 30%
- Algorithm Creation: 30%
- Measure Models: 10%
- Run as a service: 5%
- Integrate with workflows: 5%
- Monitor: 5%
How do we reduce labeling time?
Option #1: negatives are everything that’s not positive
How do we reduce labeling time?
Option #2: exclude negatives that have a high likelihood of being positive
How do we reduce labeling time?
Option #3: iterative process with user
Appetizer # 2: NLP Optimization

Content type optimization: News, Research
- Language
- Structure
- Diversity of sources
- Length

Metadata optimization: Companies and Topics

Use case optimization
- Segmented classification
- Public companies in research report
- Ticker in research reports
Segmented Classification vs. Regular Classification

- **Brazil**
- **Dividends**
- **Hydroelectric Power St.**
- **Spain**
- **Debt Renegotiation**
- **Portugal**

**Topics**
- Business Finance
- Performance Results
- Western Europe
- Leverage Ratios
- Corporate Taxation
- Exchange Rate Impact
- Guidance Pre-Announcement
- Corporate Debt
- Scalability
- Corporate Debt Restructuring
- Broker Research
- Earnings Releases
- Margin Expansion
- Margin Pressure
- Equities
- Change of CEO

**Languages**
- English

**Document View**
- 28
- "Energias de Portugal EQUITIES ELECTRIC UTILITIES (EDP PL) Hold: Better year-end guidance on stronger hydro Portugal"

- Solid hydro conditions in Portugal have allowed EDP to raise its guidance for year-end
- Valuation metrics (mainly PE and yield) remain attractive but macro-sovereign risks in Portugal still persist
- CTP unchanged at EUR3.50/share
Roundtable: Linked-Data Models - merging knowledge-graphs with unstructured data to expose hidden insights and relationships

Geoff Horrell & Natasa Varytimou
Agenda

Customer Challenge
Linked Data Use Cases
Growth in Linked Data
Deep dive on the Semantic Web stack & why this matters
Knowledge Graphs – an idea whose time has come

Google led the way, Microsoft and others catching up

Apple once led the pack with its intelligent assistant Siri, but in just a few years, Amazon, Microsoft and Google have chipped away at its lead.

Siri is a critical component of Apple’s vision for the future, so integral that it was willing to spend $200 million to acquire Lattice Data over the weekend. The startup was working to transform the way businesses deal with paragraphs of text and other information that lives outside neatly structured databases. These engineers are uniquely prepared to assist Apple with building a next-generation internal knowledge graph to power Siri and its next generation of intelligent products and services.

Amazon acquires Santa Barbara start-up Graphiq to try to bolster Alexa
The Semantic Web was a precursor


39% of web pages now carry Linked Data.

November 2017 - http://webdatacommons.org/structureddata/2017-12/stats/stats.html
Customer Challenges

Organizing and connecting data for insight, action and savings

Integration across applications is impossible because of different data formats, semantics and a lack of common metadata

Process Automation requires data lineage and interactions from multiple data sets to allow automated actions

Requires data designed for machine consumption

Enterprise Insight systems for search and customer 360 view need to push insights

Predictive analytics requires deduction of non-obvious connections.
With the exponential increase in data analyzed, quantitative funds on the rise, and complex investment strategies, it is imperative to provide the infrastructure to quickly identify important data points relevant to their portfolio and potential new investment decisions.

- **Hedge Funds/Private Equity: Research**: Look beyond basic fundamental indicators in order to see what outstanding investments that company may have.

- **Traditional Asset Management & Wealth Management**: Optimize client relationships & risk management through graph driven analysis.

- **Alpha Generation**: Use graph to overlay historical entity relationship patterns.

Description: Uber’s embattled Chief Executive Travis Kalanick told that he will take time away from the company he helped to found.

Chief Executive Travis Kalanick will temporarily leave his ride-hailing to impose some order. The indefinite span of absence and putting a Travis Kalanick resigns under investor pressure

Instrument Topic: Executive Change Article x3 Company

Sentiment: Negative Portfolio A

Portfolio B
As banks begin to reevaluate how risk is calculated in trading and banking books, graph databases enable advanced capabilities when it comes to calculating these numbers. Additionally, it is extremely important for major financial institutions to be able to have a single source of truth when it comes to managing and analyzing relationships within the data in terms of entities and financial instruments.

- **Customer 360**: Optimize client relationships & opportunity identification through graph driven analysis. Knowing a customer is using products from other parts of the bank the graph can enable advanced predictive services for high value clients.

- **Counterparty Risk**: Graph databases can drive significantly deeper insight by being able to extrapolate all the potential risks associated with counterparties of positions held at the bank.
Sales and Trading Desks

Sales & Trading desks are built around speed and execution. Linked data approaches can help create automation processes that combine external and internal data to improve the efficient making of trades and suggestion of trade ideas.

Traditional Data Sources

- Manage Data
- Performs Analytic
- Relevant data
- Appropriate data sources
- Transforms data into a graph

Predictive S&T Models

- Analytics Engine
- Customer Pattern Recognition & News Filtering

- By leveraging advanced relational data as provided by a merged internal and external data graph, a trader can create high confidence predictions of behavior accounts ahead of time based on stimulus.

- Examples of relationships shown above can highlight key interdependencies that are not available through standardized financial data resources.

- Standard financial reporting data & pricing data does not include deeper (or predictive) insight to the external data ecosystem of connect to customer information. Those systems are expensive to rebuild.

Source: TR, GraphFeed
Semantic Web – The initial concept has been evolving

Some Semantic Web Languages

- **Open World Assumption**
  - OWL
    - Restrictions, union, disjoint, inverse, symmetric etc.
  - SPIN
    - Common constraints
    - Constraints, rules, templates, functions
  - RDF Schema
    - Classes, subclasses, subproperties

- **Closed World Assumption**
  - Shapes
  - Other SPIN Libraries

**RDF**
- Resources, types, properties, triples, named graphs, datatypes

**Tim’s Semantic Web Stack (2005)**
Turning the semantic web concept into practice for financial applications

The interaction of these standards creates a powerful toolkit for data management

**URI**: Global unique identifier for everything

**RDF Schema** allows description of direct & indirect knowledge around entities

**OWL** allows more advanced inferencing and relationships

**SHACL** is a data modelling language to describe constraints (data quality) on RDF data without the need of programatically effort

**SPARQL** is a popular query language for RDF graphs

**RDF** enables the automated processing of information by restructuring individual statements in the form of triples

**RDF Schema** allows description of direct & indirect knowledge around entities

**OWL**
- Restrictions, union, disjoint, inverse, symmetric etc

**SHACL CONSTRAINTS**
- Rules, Targets, Functions

**SPARQL**
UNIFORM RESOURCE IDENTIFIER (URI)
The Power of the Universal Resource Identifier (URI)

Global unique identifier for everything \(\rightarrow\) Agreement that facilitates integration and information discovery

URI

- The “same thing” is represented the same with the same unique id in the context of an organization
- Consistent statements
- Merge metadata and information since they will all refer to the same “thing”

FULL URIs

Consists of 2 parts: Namespace + Localname

\(<http://id.thomsonreuters.com/kg/data/>\text{MyOrganization}\>

Namespace refers to the domain the URI lives in

Example

\(<http://id.thomsonreuters.com/kg/data/>\text{myOrganization}\>
\(<http://id.thomsonreuters.com/df/data/>\text{myOrganization}\>

are two different “things”

QNAMES

Prefix = Agreement to refer to the namespace in a more...human friendly way

For example we declare

@prefix org-data: \text{http://id.thomsonreuters.com/kg/data/}

@prefix df-data: \text{http://id.thomsonreuters.com/df/data/}

so we now can say

\text{org-data:myOrganization}
\text{df-data:myOrganization}

What if 3 different URIs represent the same organization, the same “real world” thing?

And now comes RDF...
RESOURCE DESCRIPTION FRAMEWORK (RDF)
RDF enables the automated processing of information by structuring individual statements in a form of triple.

URI
The “thing” you want describe

URI
Usually a verb, a term to describe the relationship. Taken from an ontology

URI or datatype

<table>
<thead>
<tr>
<th>org-data:Volkswagen</th>
<th>org-schema:hasName</th>
<th>“Volkswagen”</th>
</tr>
</thead>
<tbody>
<tr>
<td>org-data:Volkswagen</td>
<td>org-schema:isParentof</td>
<td>car-data:Audi</td>
</tr>
</tbody>
</table>
RDF GRAPHS are set of triples that can be created and stored separately but can be easily merged for integration of information.

Graph A

http://tr.com/Employee

http://tr.com/Person

Natasa Varytimou

Graph B

http://tr.com/NatasaVarytimou

http://tr.com/GiuseppeSaltini

Varytimou

Merge Graph A and Graph B

http://tr.com/Person

Natasa Varytimou

http://tr.com/GiuseppeSaltini

Varytimou

http://tr.com/NatasaVarytimou

http://tr.com/Employee

http://tr.com/NatasaVarytimou

http://tr.com/Person

Natasa Varytimou

http://tr.com/GiuseppeSaltini

Varytimou

http://tr.com/NatasaVarytimou
RDF SCHEMA (RDFS) & OWL
RDFS enables more knowledge around my entities and their relationships

- Subclasses
- Domains & Ranges
- rdfs:label, rdfs:comment etc.
And if I want to describe more complicated relationships we have OWL ...

- Restrictions, inverseOf, Functional Properties
- owl:sameAs
Owl:SameAs enables integration in cases that 2 URIs represent the same “thing”

If we cannot use the power of one, global uri...

We declare that these 2 are the same real “thing”
SHACL ( SHAPES CONSTRAINT LANGUAGE )
SHACL is a data modelling language
Describes constraints on RDF without the need of programmatically effort

- Defines constraints, “shapes”, on the data
- Checks the data based on these constraints and creates violation messages
Data Quality Example with SHACL

- Example: A quote is active and has exchanges but the instrument itself is inactive.

- We can create our own messages.

```
> Error ** Exchange Exchange YYY-2 of Inactive instrument. The Quote-YYY seems to be active
Resource: Quote-YYY
Value: Quote-YYY

Error ** Exchange Exchange YYY-1 of Inactive instrument. The Quote-YYY seems to be active
Resource: Quote-YYY
Value: Quote-YYY
```
SHACL in a glance!

Not everything in the W3C standard yet but yet useful

Overview of SHACL with Advanced Features
Taken from https://www.topquadrant.com/2017/09/13/shacl/
SHACL Rules to create new information out of my data
Like inferencing but in a more controlled way

- Creates new information, new triples based on conditions

```turtle
example:Product
    rdf:type owl:Class ;
    rdf:type sh:NodeShape ;
    sh:rule [
        rdf:type sh:TripleRule ;
        sh:condition [
            sh:property [
                sh:path example:size ;
                sh:in (1
                    2
                    3
                ) ;
            ] ;
        ] ;
        sh:object example:SmallProduct ;
        sh:predicate rdf:type ;
        sh:subject sh:this ;
    ] ;
```

Example: If a product has a size in a particular range then it is a small product
TR PermID – A Barcode for Information

Currency
Canadian Dollar
PermID: 500140

Instrument
Ordinary shares
PermID: 300281

Quote
Primary Ticker – TRI
Primary Exchange – TSX
Primary RIC – TRLTO
PermID: 85909928696

Asset Class
Ordinary shares
PermID: 100052

Geography
Canada
PermID: 100052

Organization
Thomson Reuters Corp
PermID: 4295861160

Unique
Permanent
Key identifier

TR Industry Classification
Professional Information Services (NEC)
PermID: 4294951759
"The Resource Description Framework (RDF) is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model. It has come to be used as a general method for conceptual description or modeling of information that is implemented in web resources ... » -Wikipedia.org
Actelion CEO says Basilea purchase would make no sense: paper

ZURICH (Reuters) - Swiss drug maker Actelion's chief executive quashed speculation over its possible interest in purchasing Basilea Pharmaceutica.

<Node> Actelion Ltd
Node Annotations
Location = Zurich
Identifier 44272

<Property>
Is Acquirer Of

<Node> Basilea Pharmaceutica AG
Node Annotations
Location = Basel
Identifier = 43936

Property Annotations
Status = Rumour
Iscore = 0.57
Date = Jan 15th 2016
RDF is a part of a comprehensive set of open standards designed for sharing &
merging data – property graphs are optimized for traversal for specific use cases.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Property Graph</th>
<th>RDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreed Standard</td>
<td>No – many standards</td>
<td>Yes - W3C standard</td>
</tr>
<tr>
<td>Sharing</td>
<td>Customized for query performance, harder to share</td>
<td>Standards designed for sharing and merging</td>
</tr>
<tr>
<td>Database Support</td>
<td>Proprietary Property Graph Databases</td>
<td>Can be stored in document, key value, dedicated triple stores or property graphs.</td>
</tr>
</tbody>
</table>