Natural Language Processing & Sentiment Analysis in Trading

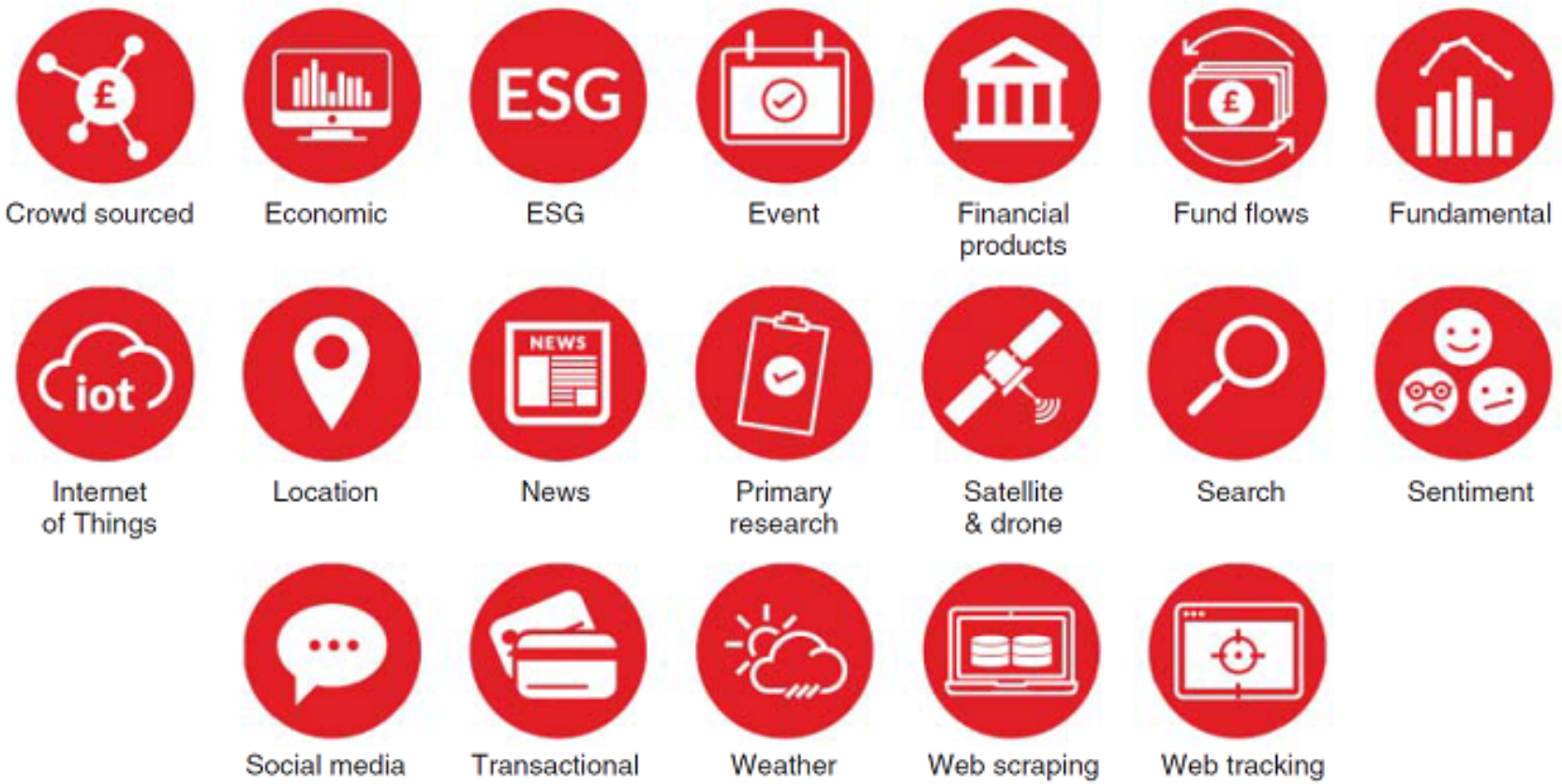
Radha Krishna

December 10, 2023

Introduction to Alt Data

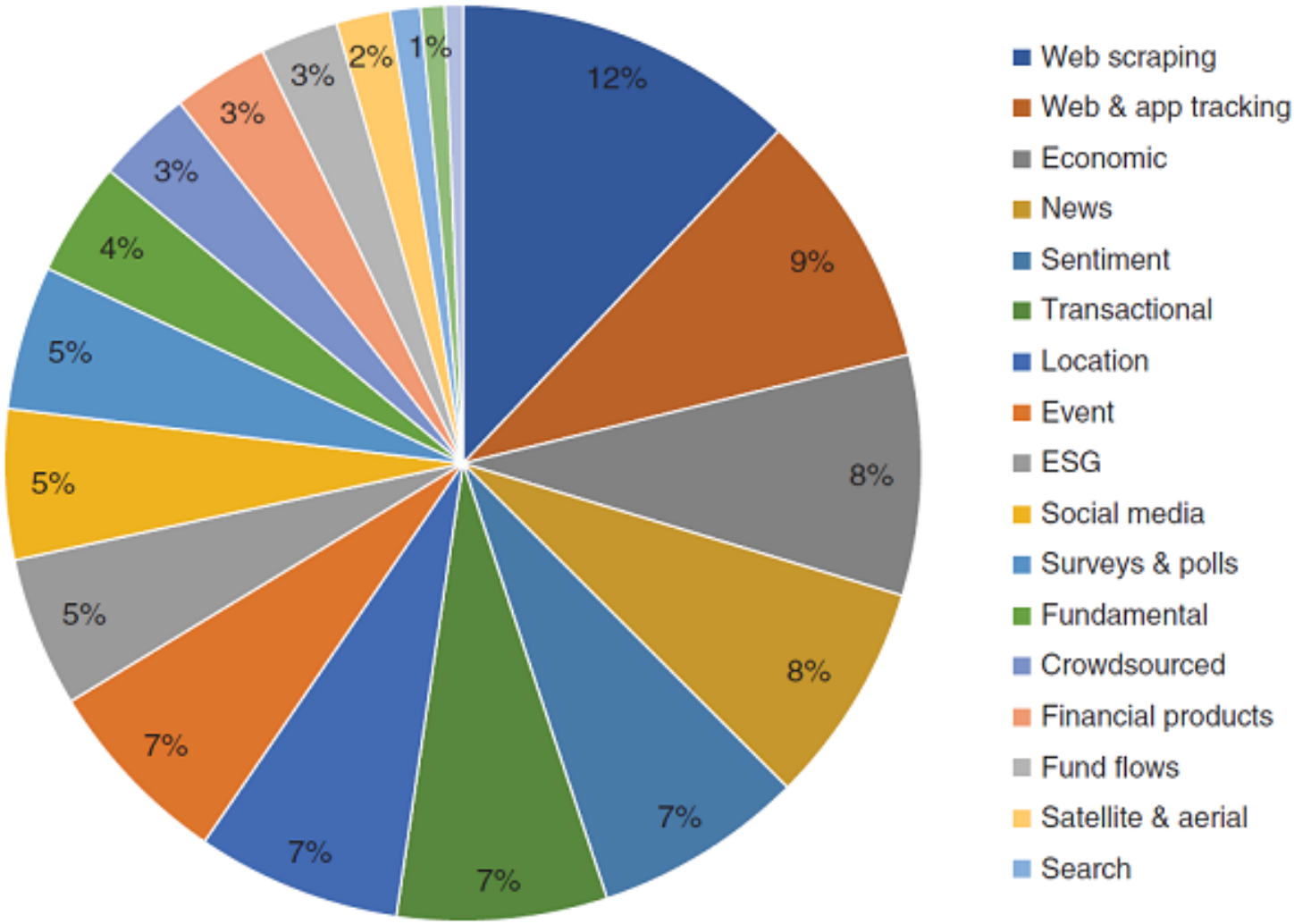
- *Novel Data sources* that can be used for investment management analysis and decision-making purposes in quant or quantamental investing
- Data sources that are beyond "market discovered" trades and quotes data
- Demand drivers
  - Ever shrinking alpha - Traditional asset managers are using Alt Data for investment ideas
  - Hedge Funds + Quant Funds + Algo Trading Houses - Using Alt Data to obtain *hidden state* information not available via traditional sources
  - Corporates, Venture Capitalists are also looking to get an edge on their decision making
- Supply drivers
  - More than *1000* alternate datasets available from various data vendors
  - FinTech Wave + Technological advances

Alt Data: Categories



Source: Neudata

Alt Data: Usage Patterns



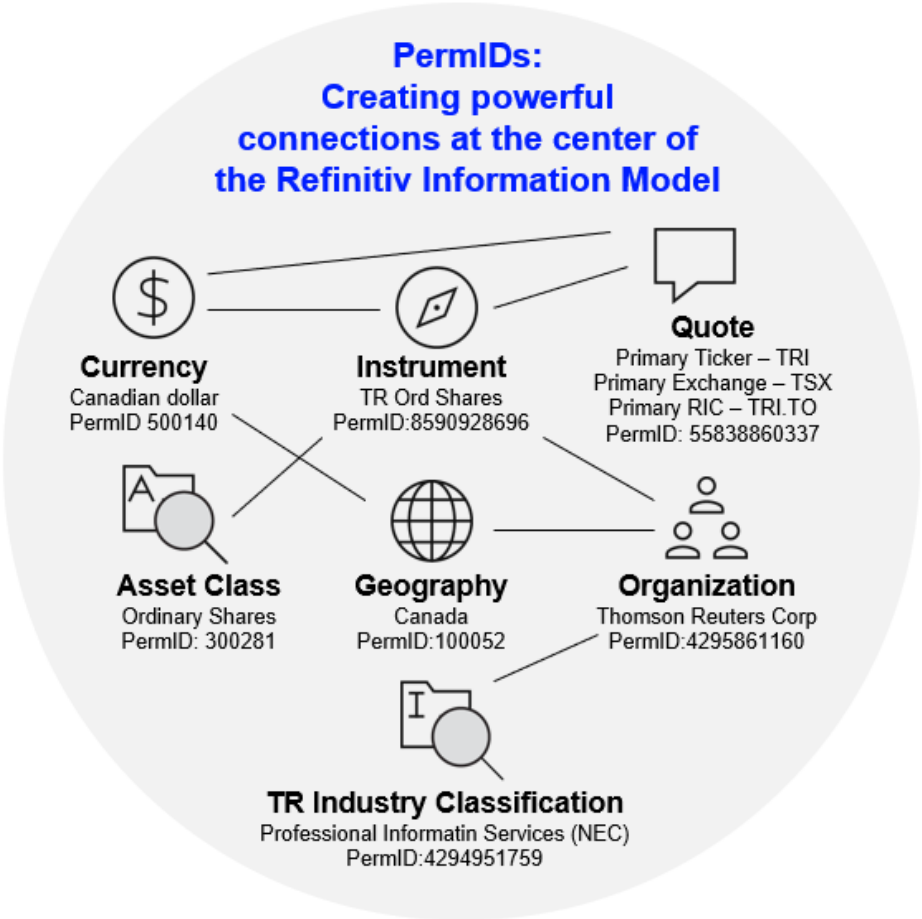
Source: Neudata

Alt Data - Sample Data

- Machine Readable News - *MRN-JSON-Sample.json*
- Sentiment Data for Equities, Commodities - *Sample-Company-News-File.txt, Sample-Company-Scores-File.txt*
- Market Psych Sentiment Data: Multi-dimensional view of Sentiment - *Sample-Commodities-Sentiment-File.txt*
- Transcripts : Verbatim reports of earnings, guidance, corporate conference calls - *Sample-Transcript(Deutsche Bank).XML*
- Briefs : Summaries of corporate conference calls including the full Q&A session - *Sample-Brief(BIDU).XML*
- ESG: Sample Data Walk through - *Eikon*
- Surveys and Polls: Reuters Polls, IBES - *Eikon*
- Advanced Events - *Eikon*
- Economic Monitor - *Eikon*

Download Link : `https://bit.ly/3Pr9HG5`

PermID - Open, Permanent, and Universal Identifiers



- Information on Organizations residing in different databases
- Market Data - mix of legacy and modern databases
- Reference data - Acquired and Internal data
- Need for consistency and scalable maintenance architecture
- PermID for various types of entities

PermID - Open, Permanent, and Universal Identifiers

- *Complementary* to the existing identifiers such as ISIN, RIC ,LEIs
- *Comprehensive:* provides identification across a wide variety of organizations, instruments, funds, issuers and people
- *Connected:* PermIDs connect all data sets in the Refinitiv information model, helping gain valuable insights
- *Machine-readable:* – a 64-bit number that operates beneath the surface to connect related information instantly and seamlessly
- *Open:* – Refinitiv open strategy is driving new opportunities, collaboration and innovation; Open Data Institute (ODI)-certified
- *Permanent:* – a never reused identifier is assigned to each information object; they don't change and allow you to trace object changes over time
- *Precise:* – points to each specific information object
- *Scalable:* – a vast number of PermIDs are available
- *Unique:* – each instance has its own PermID

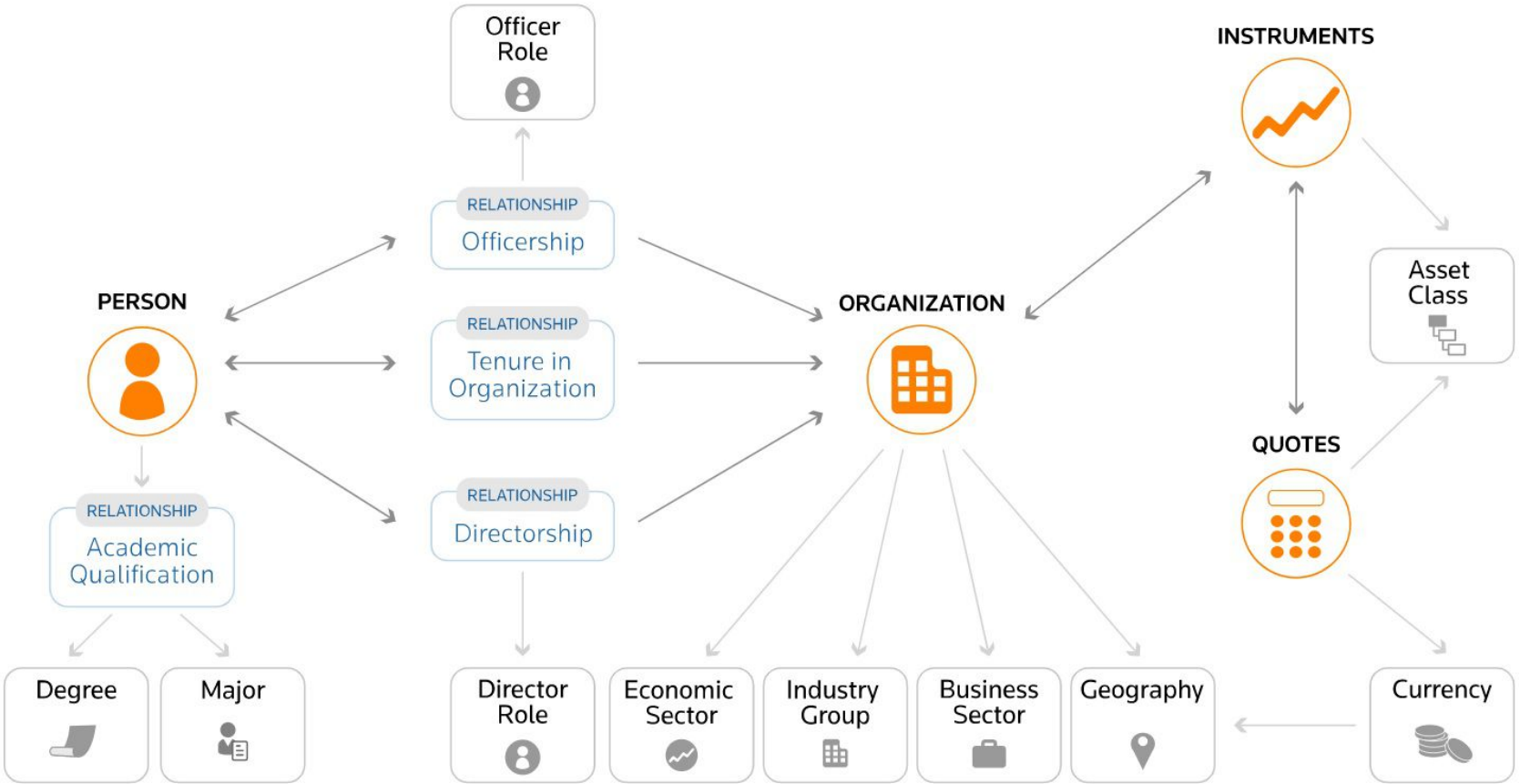
PermID - Open, Permanent, and Universal Identifiers

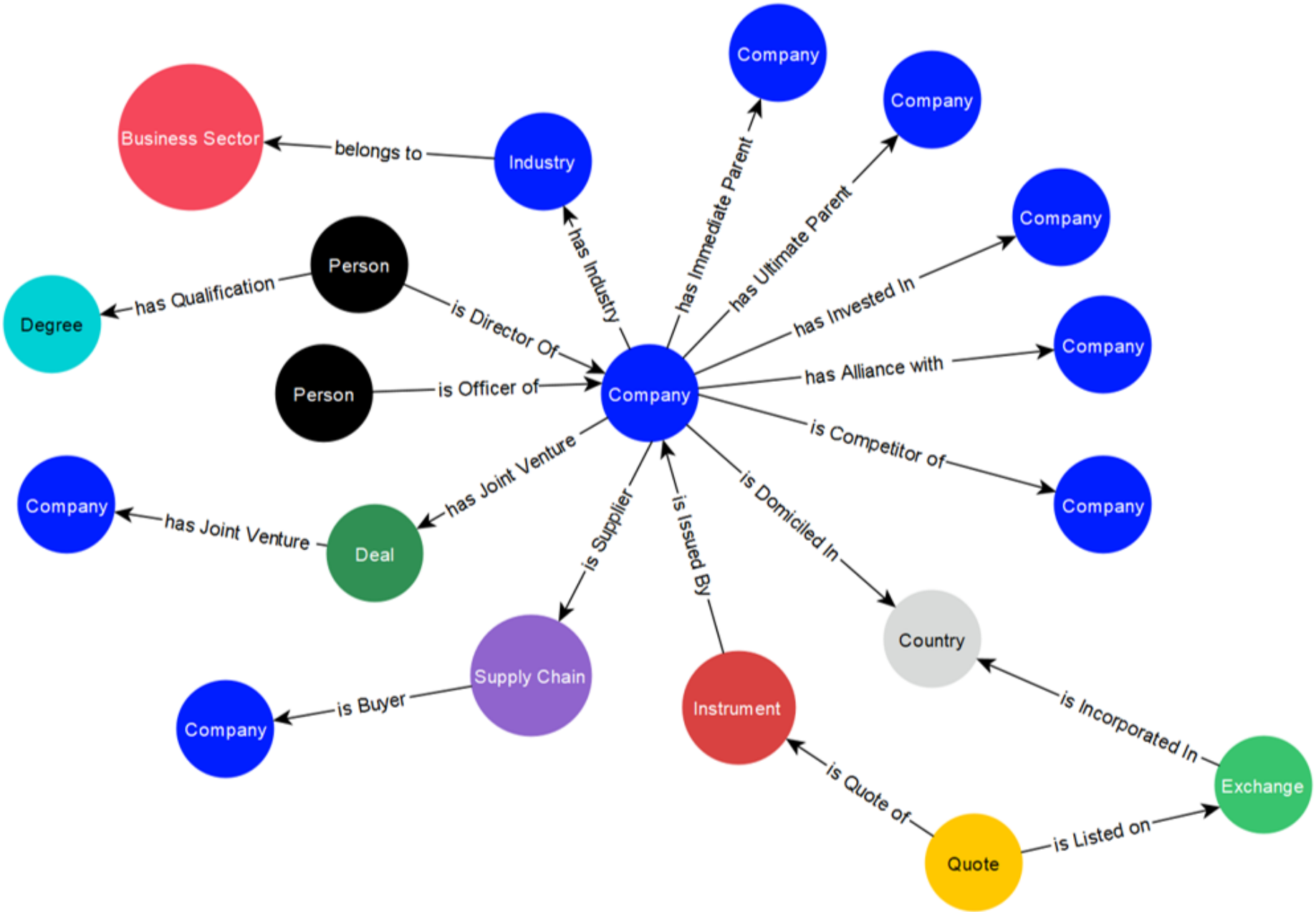
Geography	Financial Analyst	Financial Markets Event	Financial Research Contributor
Industry	Deals JV	Index Record	Deal Basic Information
Currency	Bankruptcy	Index Constituent	Proxy Fight
Commodity	Change Event	Index Value	Repurchase
Language	Dividend	Fund Admin Status	Financial Markets Calendar
Script	M&A Offer	Fund Class Admin Status	Economic Indicator Metadata
Time Zone	Meeting	Person	Object of Forecast
Holiday	Restructure	Relationship	Financial Filing Document
Holiday Event	Trading Status	Event	Relationship
Warrant Exercise	Unit	EAN Instrument	Market Attributable Source

Usecases

- Search smarter develop platforms to easily search and connect your data, uncovering the right connections
- Tagging all of your data and exposing powerful linkages for unique insights
- Comingle internal and external data to manage risks

Knowledge Graph




Business Sector
belongs to
Industry
Company
Company
Company
Company
Company
Person
has Qualification
Degree
is Director Of
Person
is Officer of
has Industry
has Immediate Parent
has Ultimate Parent
has Invested In
has Alliance with
is Competitor of
Company
Company
Company
has Joint Venture
Deal
has Joint Venture
is Supplier
is Issued By
is Domicilled In
Supply Chain
is Buyer
Company
Instrument
Country
is Incorporated In
Exchange
is Quote of
Quote
is Listed on

Knowledge Graph - Node Statistics

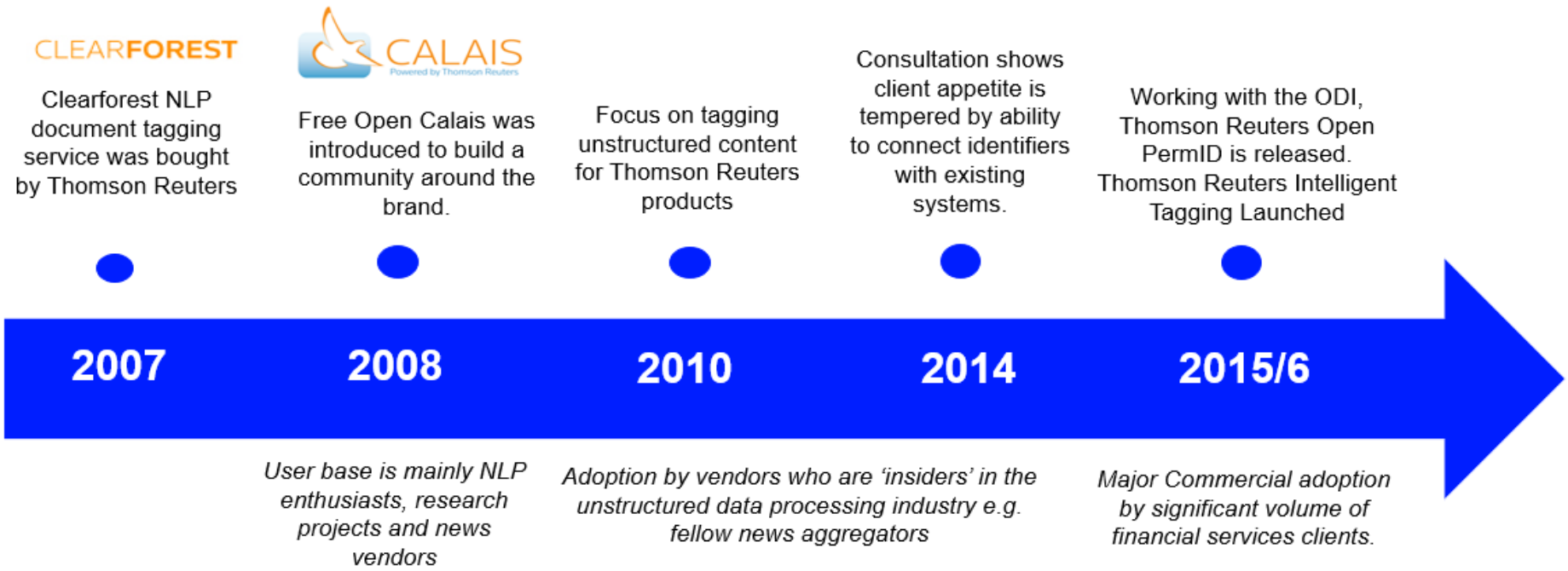
Entity Type	# of Nodes in the graph
Quote	115 Million
Instrument	43 Million
Organizations	5.2 Million
Person	2.9 Million
Officers	2.7 Million
Directors	1.6 Million
Mergers and Acquisitions	860,000
Strategic Investments	225,000
Competitor Information	58,000
Supply Chain Agreements	52,000
Business Activities	837
Asset Categories	145

- All the Graph data is serialized is available in `RDF` format
- Billions of triples representing node attributes and node relationships
- Many commercially available databases such as Neo4j, TigerGraph, TitanDB, Amazon Neptune available to store the data

PermID Demo + Hands on

https://permid.org/

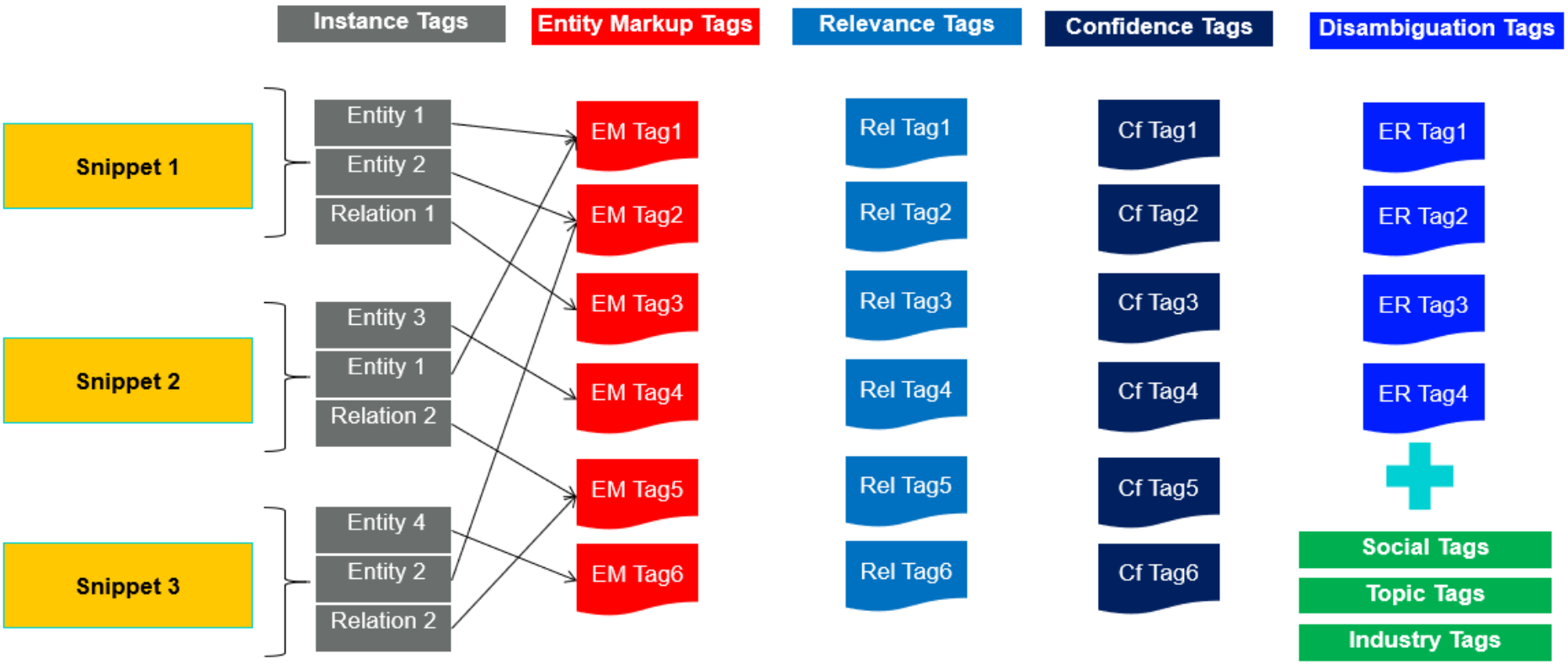
Intelligent Tagging



Intelligent Tagging

- Turning qualitative, unstructured text into quantitative and actionable insight
- Tag the unstructured data with "metadata"
- What kind of metadata is attached ?
  - Entities
  - Relations
  - Industries
  - Topics
  - Social Tags
- Input Content types : text/html, text/xml, application/pdf, text/raw
- Languages supported : English, French and Spanish
- Authentication
  - Hosted : Need to send the API key for every request
  - On Premise : No need to send the API key. Unlimited access for "On Premise" deployment

Intelligent Tagging - Under the hood



Intelligent Tagging - Entities

Entities that can be tagged by the engine:

Anniversary	FaxNumber	MusicGroup	PowerStation	SportsGame
City	Holiday	NaturalFeature	Product	SportsLeague
Company	IndustryTerm	OperatingSystem	ProgrammingLanguage	Technology
Continent	Journalist	Organization	ProvinceOrState	TVShow
Country	MarketIndex	Person	PublishedMedium	TVStation
Currency	MedicalCondition	PharmaceuticalDrug	RadioProgram	URL
Editor	MedicalTreatment	PhoneNumber	RadioStation	Vessel
EmailAddress	Mine	PoliticalEvent	Refinery	
EntertainmentAwardEvent	Movie	Port	Region	
Facility	Music Album	Position	SportsEvent	

- Specific Algorithms have been built for carrying out Entity disambiguation tasks
- Think of the above entities as *common nouns*. For example, `Company`, the engine can recognize 5.7 Million companies, i.e. $70,000$ public companies and 5 Million private companies.

Intelligent Tagging - Relations

Relations that can be tagged by the engine:

Acquisition	CompanyForceMajeure	Conviction	PatentIssuance
Alliance	CompanyFounded	Deal	PersonAttributes
AnalystEarningsEstimate	CompanyInvestigation	DebtFinancing	PersonCareer
AnalystRecommendation	CompanyInvestment	DelayedFiling	PersonCommunication
AnalystRevision	CompanyLaborIssues	DiplomaticRelations	PersonEducation
ArmedAttack	CompanyLayoffs	Dividend	PersonEmailAddress
ArmsPurchaseSale	CompanyLegalIssues	EmploymentChange	PersonLocation
Arrest	CompanyListingChange	EnvironmentalIssue	PersonParty
Acquisition	CompanyLocation	EquityFinancing	PersonRelation
Alliance	CompanyMeeting	Extinction	PersonTravel
BusinessRelation	CompanyNameChange	FamilyRelation	PoliticalEndorsement
Buybacks	CompanyProduct	FDAPhase	PoliticalRelationship
CompanyAccountingChange	CompanyRelationship	IndicesChanges	PollsResult
CompanyAffiliates	CompanyReorganization	JointVenture	ProductIssues
CompanyCompetitor	CompanyRestatement	ManMadeDisaster	ProductRecall
CompanyCustomer	CompanyTechnology	Merger	ProductRelease
CompanyEarningsAnnouncement	CompanyTicker	MilitaryAction	Quotation
CompanyEarningsGuidance	CompanyUsingProduct	MovieRelease	StockSplit
CompanyEmployeesNumber	ConferenceCall	MusicAlbumRelease	Trial
CompanyExpansion	ContactDetails	PatentFiling	VotingResult

- Specific Algorithms have been built for carrying out Relationship mining
- Most of the tagged relationships belong to financial sector.

RDF-Primer

- Not a Data Format . It is a data model for describing resources and relationships
- RDF statement : Two things and a relationship between them
  - Subject - Predicate - Object
  - Subject and Predicate represented as URI
  - Object can be another Subject or a value
- Subject - Predicate - Object Patterns
  - URI-URI-value
  - URI-URI-URI
- RDF Data model can be serialized in a variety of standard formats
  - N-Triples, Turtle, RDF/XML, Notation 3, JSON-LD, RDFa

Examples

http://test.org/1-7303 http://test.org/CommonName “StarHub”.
http://test.org/1-7303 http://test.org/isChildOf http://test.org/1-7304.

Intelligent Tagging Demo + Hands on

https://permid.org/onecalaisViewer

Text Processing - Common Tasks

- Case conversion
- Removing Punctuation Symbols
- Removing Numbers
- Stopword removal
- Stemming
- Lemmatization
- Tokenization
- Part-of-speech tagging
- Dependency parsing
- Named Entity Recognition
- Sentence segmentation

Text Processing

Strip White Spaces:

Original Text	**Processed Text**
TAIPEI, April 15 (Reuter) - Taiwan's money rates finished mixed on Monday, dealers expecting overnight to rise further amid current bullish stock market and income tax payments. Overnight ended at 6.134 percent against Saturday's 5.949, while 30-day commercial paper fell to 7.00 from 7.10-7.15. Though Taiwan share prices hit a new 11-month high on Monday, attracting liquidity into the stock market, bond traders did not expect significantly tighter conditions as the central bank monetary policy remains relaxed.	TAIPEI, April 15 (Reuter) - Taiwan's money rates finished mixed on Monday, dealers expecting overnight to rise further amid current bullish stock market and income tax payments. Overnight ended at 6.134 percent against Saturday's 5.949, while 30-day commercial paper fell to 7.00 from 7.10-7.15. Though Taiwan share prices hit a new 11-month high on Monday, attracting liquidity into the stock market, bond traders did not expect significantly tighter conditions as the central bank monetary policy remains relaxed.

Text Processing

Change to lower case:

Original Text	**Processed Text**
TAIPEI, April 15 (Reuter) - Taiwan's money rates finished mixed on Monday, dealers expecting overnight to rise further amid current bullish stock market and income tax payments. Overnight ended at 6.134 percent against Saturday's 5.949, while 30-day commercial paper fell to 7.00 from 7.10-7.15. Though Taiwan share prices hit a new 11-month high on Monday, attracting liquidity into the stock market, bond traders did not expect significantly tighter conditions as the central bank monetary policy remains relaxed.	taipei, april 15 (reuter) - taiwan's money rates finished mixed on monday, dealers expecting overnight to rise further amid current bullish stock market and income tax payments. overnight ended at 6.134 percent against saturday's 5.949, while 30-day commercial paper fell to 7.00 from 7.10-7.15. though taiwan share prices hit a new 11-month high on monday, attracting liquidity into the stock market, bond traders did not expect significantly tighter conditions as the central bank monetary policy remains relaxed.

Text Processing

Remove Punctuation:

Original Text	**Processed Text**
TAIPEI, April 15 (Reuter) - Taiwan's money rates finished mixed on Monday, dealers expecting overnight to rise further amid current bullish stock market and income tax payments. Overnight ended at 6.134 percent against Saturday's 5.949, while 30-day commercial paper fell to 7.00 from 7.10-7.15. Though Taiwan share prices hit a new 11-month high on Monday, attracting liquidity into the stock market, bond traders did not expect significantly tighter conditions as the central bank monetary policy remains relaxed.	TAIPEI April 15 Reuter Taiwans money rates finished mixed on Monday dealers expecting overnight to rise further amid current bullish stock market and income tax payments Overnight ended at 6134 percent against Saturdays 5949 while 30day commercial paper fell to 700 from 710715 Though Taiwan share prices hit a new 11month high on Monday attracting liquidity into the stock market bond traders did not expect significantly tighter conditions as the central bank monetary policy remains relaxed

Text Processing

Remove Numbers:

Original Text	**Processed Text**
TAIPEI, April 15 (Reuter) - Taiwan's money rates finished mixed on Monday, dealers expecting overnight to rise further amid current bullish stock market and income tax payments. Overnight ended at 6.134 percent against Saturday's 5.949, while 30-day commercial paper fell to 7.00 from 7.10-7.15. Though Taiwan share prices hit a new 11-month high on Monday, attracting liquidity into the stock market, bond traders did not expect significantly tighter conditions as the central bank monetary policy remains relaxed.	TAIPEI, April (Reuter) - Taiwan's money rates finished mixed on Monday, dealers expecting overnight to rise further amid current bullish stock market and income tax payments. Overnight ended at . percent against Saturday's ., while -day commercial paper fell to . from .-.. Though Taiwan share prices hit a new -month high on Monday, attracting liquidity into the stock market, bond traders did not expect significantly tighter conditions as the central bank monetary policy remains relaxed.

Text Processing

Stop Word Removal:

Original Text	**Processed Text**
TAIPEI, April 15 (Reuter) - Taiwan's money rates finished mixed on Monday, dealers expecting overnight to rise further amid current bullish stock market and income tax payments. Overnight ended at 6.134 percent against Saturday's 5.949, while 30-day commercial paper fell to 7.00 from 7.10-7.15. Though Taiwan share prices hit a new 11-month high on Monday, attracting liquidity into the stock market, bond traders did not expect significantly tighter conditions as the central bank monetary policy remains relaxed.	TAIPEI, April 15 (Reuter) - Taiwan's money rates finished mixed Monday, dealers expecting overnight rise amid current bullish stock market income tax payments. Overnight ended 6.134 percent Saturday's 5.949, 30-day commercial paper fell 7.00 7.10-7.15. Though Taiwan share prices hit new 11-month high Monday, attracting liquidity stock market, bond traders expect significantly tighter conditions central bank monetary policy remains relaxed.

Text Processing

Stemming:

Original Text	**Processed Text**
TAIPEI, April 15 (Reuter) - Taiwan's money rates finished mixed on Monday, dealers expecting overnight to rise further amid current bullish stock market and income tax payments. Overnight ended at 6.134 percent against Saturday's 5.949, while 30-day commercial paper fell to 7.00 from 7.10-7.15. Though Taiwan share prices hit a new 11-month high on Monday, attracting liquidity into the stock market, bond traders did not expect significantly tighter conditions as the central bank monetary policy remains relaxed.	TAIPEI, April 15 (Reuter) - Taiwan money rate finish mix on Monday, dealer expect overnight to rise further amid current bullish stock market and incom tax payments. Overnight end at 6.134 percent against Saturday 5.949, while 30-day commerci paper fell to 7.00 from 7.10-7.15. Though Taiwan share price hit a new 11-month high on Monday, attract liquid into the stock market, bond trader did not expect signific tighter condit as the central bank monetari polici remain relaxed.

Text Processing

Lemmatization:

Original Text	**Processed Text**
TAIPEI, April 15 (Reuter) - Taiwan's money rates finished mixed on Monday, dealers expecting overnight to rise further amid current bullish stock market and income tax payments. Overnight ended at 6.134 percent against Saturday's 5.949, while 30-day commercial paper fell to 7.00 from 7.10-7.15. Though Taiwan share prices hit a new 11-month high on Monday, attracting liquidity into the stock market, bond traders did not expect significantly tighter conditions as the central bank monetary policy remains relaxed.	TAIPEI, April 15 (Reuter) - Taiwan's money rates finished mixed on Monday, dealers expecting overnight to rise further amid current bullish stock market and income tax payments. Overnight ended at 6.134 percent against Saturday's 5.949, while 30-day commercial paper fell to 7.00 from 7.10-7.15. Though Taiwan share prices hit a new 11-month high on Monday, attracting liquidity into the stock market, bond traders did not expect significantly tighter conditions as the central bank monetary policy remains relaxed.

Text Processing

Strip White Spaces, Lower casing, Removing Punctuation, Removing Numbers, Stop Word Removal, Lemmatization

Original Text	**Processed Text**
TAIPEI, April 15 (Reuter) - Taiwan's money rates finished mixed on Monday, dealers expecting overnight to rise further amid current bullish stock market and income tax payments. Overnight ended at 6.134 percent against Saturday's 5.949, while 30-day commercial paper fell to 7.00 from 7.10-7.15. Though Taiwan share prices hit a new 11-month high on Monday, attracting liquidity into the stock market, bond traders did not expect significantly tighter conditions as the central bank monetary policy remains relaxed.	taipei april reuter taiwans money rates finished mixed monday dealers expecting overnight rise amid current bullish stock market income tax payments overnight ended percent saturdays day commercial paper fell though taiwan share prices hit new month high monday attracting liquidity stock market bond traders expect significantly tighter conditions central bank monetary policy remains relaxed

Turning words to numbers

- Term Frequency Matrix (TF)
- Term Frequency-Inverse Document Frequency Matrix (TF-IDF)
- Word2vec models
  - Continous Bag-of-Words (CBOW)
  - Skip-gram based
- Pre-trained Models
  - Universal Language Model Fine-Tuning(ULMFiT)
  - Embedding from Language Models(ELMo)
  - Transformer
  - Universal Sentence Encoder(USE)
  - GPT-2
  - Bidirectional Encoder Representations from Transformers(BERT)
  - Transformer-XL
  - XLNet

Text Processing - Hands on

A little bit of history...

- Human annotation all the documents is too slow
- Identify the themes across unstructured data and explore across themes
- Unsupervised machine learning techniques
  - Latent Semantic Analysis(LSA) - *Cognitive Psychology*
  - Probabilistic Latent Semantic Analysis(PLSA)
  - Latent Dirichlet Allocation(LDA)
- LSA(1988) : Organize documents as matrix (Document Term Matrix) and analyze the matrix; Turned in to a *Linear Algebra Problem(Factorization)*
- PLSA(2000) : Embed LSA in a statistical model; *Probabilistic model of word counts using concepts from language modeling and LSA*. Work directly with the counts instead factorizing a matrix.
- LDA(2003) was introduced by David Blei, Andrew Ng, and Michael Jordan.
  - PLSA Topic Simplex vs Alternative idea : Andrew Ng drew on a napkin at *Brewed Awakening*, a Coffee shop at Berkeley
  - *Variational Inference (Before 2010)*; max # of documents = *200k-500k*
  - *Stochastic Optimization(Post 2010)*; millions of documents is routine
- Features are fed in to downstream machine learning algorithms

Latent Semantic Analysis(LSA)

- Assumptions
  - Meaning of a sentence depends on the *which* words occur and not *how* they occur
  - Associations between words are not explicitly stated; they are latent and can be culled from the corpus
- Main Steps
  - Represent the data as a matrix X, *TFIDF Matrix(Term Frequency Inverse Document Frequency)*;

$$X_{ij} = \log(1 + f_{t,d}) \cdot \log(N/n_t)$$

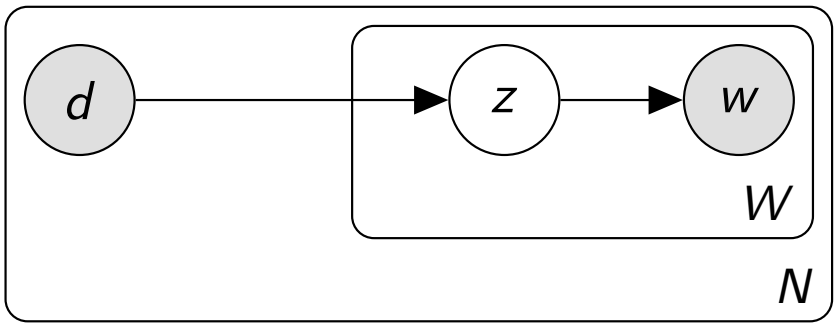
    where $f_{t,d}$ is the term frequency and $\frac{n_t}{N}$ is the document frequency
  - Dimensionality reduction using SVD(Singular Value Decomposition)

$$X = U\Sigma V^T$$

  - Each of the components of *SVD* factorization represents a latent dimension,i.e. a topic
- Applications
  - Essay grading, querying documents that are similar to the input document, parsing related concepts across documents

Probabilistic Latent Semantic Analysis(PLSA)

- Frames LSA in a probabilistic setting - sort of a Generative Model
- Key Idea:
  - Frame a graphical model that generates the data obseverved in the *Document-Term Matrix*
  - Introduce unobserved class variable z for each observation
  - Plate diagram for the graphical model



$$P(d, w) = P(d) \sum_{z} P(w|z)p(z|d)$$

- Fit the model using *Expectation-Maximization*
- Limitation : For large datasets, it is computationally intractable and parameters are unstable

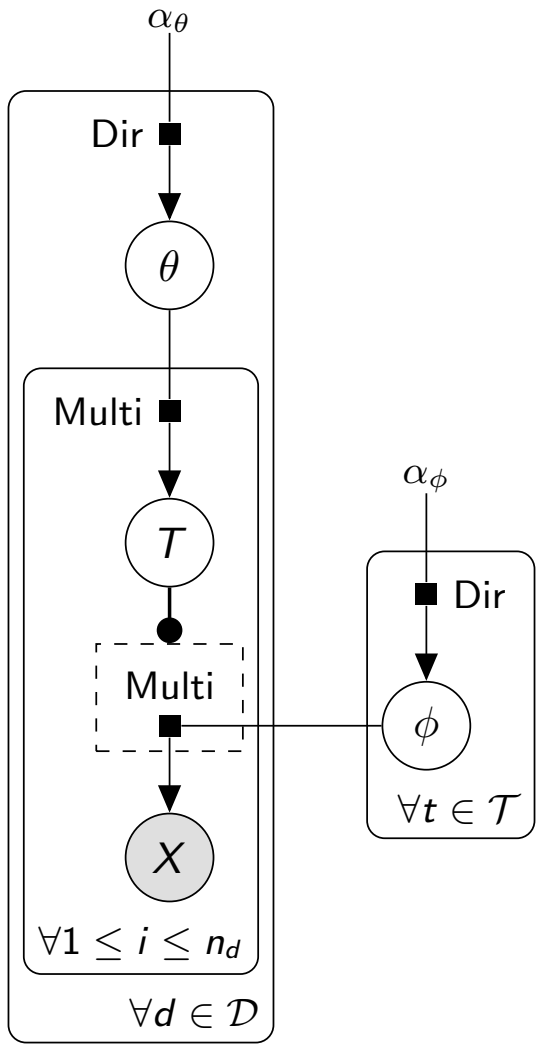
Latent Dirichlet Allocation(LDA)

Directed Factor Graph

- For each document, draw a Topic distribution θ from *Dirichlet* with hyperparameter α_θ
- Draw a Topic from *Multinomial* distribution with θ as the hyperparameter
- For specific topic chosen, draw a Word distribution ϕ from *Dirichlet* with hyperparameterα_ϕ
- Draw a word from *Multinomial* distribution with ϕ as the hyperparameter

Estimating the model parameters

- Sampling methods: Gibbs Sampling methods
- Variational Inference methods : Assume a parametrized family of distributions and then find the member in the family that fits the data



Topic Modeling - Hands on

Machine Readable News

Machine Readable News

- Understanding News Story structure
- Sample Output File
- Metadata Description
- Volume distribution of the news items
- News Source Providers
- Languages covered
- Topic Codes
- Named Item Codes
- Product Codes
- MRN content summary snapshot
- How are commericals structured ?

Evolution of a story on 2017-07-19		UTC	Δt
ALERT	MCCORMICK TO ACQUIRE RECKITT BENCKISER'S FOOD DIVISION $\langle MKC.N \rangle, \langle RB.L \rangle$	01:15:02	

Evolution of a story on 2017-07-19		UTC	Δt
ALERT	MCCORMICK TO ACQUIRE RECKITT BENCKISER'S FOOD DIVISION $\langle MKC.N \rangle, \langle RB.L \rangle$	01:15:02	
TAKE 2	MCCORMICK & COMPANY INC $\langle MKC.N \rangle$ - AGREED TO ACQUIRE RECKITT BENCKISER'S FOOD DIVISION FROM RECKITT BENCKISER GROUP PLC FOR $ 4.2 BILLION $\langle MKC.N \rangle, \langle RB.L \rangle$	01:18:00	2 min 58 sec

Evolution of a story on 2017-07-19		UTC	Δt
ALERT	MCCORMICK TO ACQUIRE RECKITT BENCKISER'S FOOD DIVISION $\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:15:02	
TAKE 2	MCCORMICK & COMPANY INC $\langle MKC.N \rangle$ - AGREED TO ACQUIRE RECKITT BENCKISER'S FOOD DIVISION FROM RECKITT BENCKISER GROUP PLC FOR $ 4.2 BILLION $\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:18:00	2 min 58 sec
TAKE 3	MCCORMICK & COMPANY INC - COMBINED PRO FORMA 2017 ANNUAL NET SALES ARE EXPECTED TO BE APPROXIMATELY $5 BILLION WITH SIGNIFICANT MARGIN ACCRETION$\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:18:28	3 min 26 sec

Evolution of a story on 2017-07-19		UTC	Δt
ALERT	MCCORMICK TO ACQUIRE RECKITT BENCKISER'S FOOD DIVISION $\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:15:02	
TAKE 2	MCCORMICK & COMPANY INC $\langle MKC.N \rangle$ - AGREED TO ACQUIRE RECKITT BENCKISER'S FOOD DIVISION FROM RECKITT BENCKISER GROUP PLC FOR $ 4.2 BILLION $\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:18:00	2 min 58 sec
TAKE 3	MCCORMICK & COMPANY INC - COMBINED PRO FORMA 2017 ANNUAL NET SALES ARE EXPECTED TO BE APPROXIMATELY $5 BILLION WITH SIGNIFICANT MARGIN ACCRETION$\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:18:28	3 min 26 sec
TAKE 4	MCCORMICK & COMPANY INC - WILL INTEGRATE RB FOODS INTO ITS CONSUMER AND INDUSTRIAL SEGMENTS AND WILL RETAIN BRAND NAMES OF FRENCH'S, FRANK'S REDHOT AND CATTLEMEN'S $\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:18:32	3 min 30 sec

Evolution of a story on 2017-07-19		UTC	Δt
ALERT	MCCORMICK TO ACQUIRE RECKITT BENCKISER'S FOOD DIVISION $\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:15:02	
TAKE 2	MCCORMICK & COMPANY INC $\langle MKC.N \rangle$ - AGREED TO ACQUIRE RECKITT BENCKISER'S FOOD DIVISION FROM RECKITT BENCKISER GROUP PLC FOR $ 4.2 BILLION $\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:18:00	2 min 58 sec
TAKE 3	MCCORMICK & COMPANY INC - COMBINED PRO FORMA 2017 ANNUAL NET SALES ARE EXPECTED TO BE APPROXIMATELY $5 BILLION WITH SIGNIFICANT MARGIN ACCRETION$\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:18:28	3 min 26 sec
TAKE 4	MCCORMICK & COMPANY INC - WILL INTEGRATE RB FOODS INTO ITS CONSUMER AND INDUSTRIAL SEGMENTS AND WILL RETAIN BRAND NAMES OF FRENCH'S, FRANK'S REDHOT AND CATTLEMEN'S $\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:18:32	3 min 30 sec
TAKE 5	MCCORMICK & COMPANY INC $\langle MKC.N \rangle$ - MCCORMICK EXPECTS TO ACHIEVE COST SYNERGIES OF APPROXIMATELY $50 MILLION $\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:18:46	3 min 44 sec

Evolution of a story on 2017-07-19		UTC	Δt
ALERT	MCCORMICK TO ACQUIRE RECKITT BENCKISER'S FOOD DIVISION $\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:15:02	
TAKE 2	MCCORMICK & COMPANY INC $\langle MKC.N \rangle$ - AGREED TO ACQUIRE RECKITT BENCKISER'S FOOD DIVISION FROM RECKITT BENCKISER GROUP PLC FOR $ 4.2 BILLION $\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:18:00	2 min 58 sec
TAKE 3	MCCORMICK & COMPANY INC - COMBINED PRO FORMA 2017 ANNUAL NET SALES ARE EXPECTED TO BE APPROXIMATELY $5 BILLION WITH SIGNIFICANT MARGIN ACCRETION$\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:18:28	3 min 26 sec
TAKE 4	MCCORMICK & COMPANY INC - WILL INTEGRATE RB FOODS INTO ITS CONSUMER AND INDUSTRIAL SEGMENTS AND WILL RETAIN BRAND NAMES OF FRENCH'S, FRANK'S REDHOT AND CATTLEMEN'S $\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:18:32	3 min 30 sec
TAKE 5	MCCORMICK & COMPANY INC $\langle MKC.N \rangle$ - MCCORMICK EXPECTS TO ACHIEVE COST SYNERGIES OF APPROXIMATELY $50 MILLION $\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:18:46	3 min 44 sec
TAKE 6	MCCORMICK - MCCORMICK HAS OBTAINED COMMITTED BRIDGE FINANCING; EXPECTS TO PERMANENTLY FINANCE TRANSACTION THROUGH COMBINATION OF DEBT AND EQUITY $\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:19:11	4 min 9 sec

Evolution of a story on 2017-07-19		UTC	Δt
ALERT	MCCORMICK TO ACQUIRE RECKITT BENCKISER'S FOOD DIVISION $\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:15:02	
TAKE 2	MCCORMICK & COMPANY INC $\langle MKC.N \rangle$ - AGREED TO ACQUIRE RECKITT BENCKISER'S FOOD DIVISION FROM RECKITT BENCKISER GROUP PLC FOR $ 4.2 BILLION $\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:18:00	2 min 58 sec
TAKE 3	MCCORMICK & COMPANY INC - COMBINED PRO FORMA 2017 ANNUAL NET SALES ARE EXPECTED TO BE APPROXIMATELY $5 BILLION WITH SIGNIFICANT MARGIN ACCRETION$\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:18:28	3 min 26 sec
TAKE 4	MCCORMICK & COMPANY INC - WILL INTEGRATE RB FOODS INTO ITS CONSUMER AND INDUSTRIAL SEGMENTS AND WILL RETAIN BRAND NAMES OF FRENCH'S, FRANK'S REDHOT AND CATTLEMEN'S $\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:18:32	3 min 30 sec
TAKE 5	MCCORMICK & COMPANY INC $\langle MKC.N \rangle$ - MCCORMICK EXPECTS TO ACHIEVE COST SYNERGIES OF APPROXIMATELY $50 MILLION $\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:18:46	3 min 44 sec
TAKE 6	MCCORMICK - MCCORMICK HAS OBTAINED COMMITTED BRIDGE FINANCING; EXPECTS TO PERMANENTLY FINANCE TRANSACTION THROUGH COMBINATION OF DEBT AND EQUITY $\langle MKC.N \rangle$, $\langle RB.L \rangle$	01:19:11	4 min 9 sec
ARTICLE	BRIEF - McCormick to acquire Reckitt Benckiser's food division for $4.2 Billion $\langle MKC.N \rangle$, $\langle RB.L \rangle$ McCormick & Company Inc.$\langle MKC.N \rangle$, a global leader in flavor, today announced that it has signed a definitive agreement to acquire Reckitt Benckiser's Food Division from Reckitt Benckiser Group plc $\langle RB.L \rangle$ . . .	01:30:48	15 min 46 sec

Sample Output File


```
"data": {
        "altId"         : "nASB0B9KO",
        "audiences"     : ["NP:DNP","NP:E","NP:EMK","NP:FMA",...,"NP:UKI"],
        "body"          : "McCormick to acquire Reckitt Benckiser's ...",
        "firstCreated": "2017-07-19T01:15:01.000Z",
        "headline"      : "BRIEF-McCormick to acquire Reckitt Benckiser's ..."
                          <MKC.N><RB.L>",
        "id"            : "tr:ASB0B9KO__1707192k",
        "instancesOf"   : [],
        "language"      : "en",
        "mimeType"      : "text/plain",
        "provider"      : "NS:RTRS",
        "pubStatus"     : "stat:usable",
        "subjects"      : [..."MKC.N", "RB.L",...],
        "takeSequence": 1,
        "versionCreated": "2017-07-19T01:15:01.000Z",
        "urgency"       : 1
    }
```


JSON - lighter, faster on-the-wire format

Metadata for a News Item

Metadata for a News Item

- Take Sequence with the appropriate timestamp

Metadata for a News Item

- Take Sequence with the appropriate timestamp
- Topic Codes
  - Describe the news item's subject matter. These cover asset classes, geographies, events, industries/sectors and other types
  - **2,000** types of topic codes

Metadata for a News Item

- Take Sequence with the appropriate timestamp
- Topic Codes
  - Describe the news item's subject matter. These cover asset classes, geographies, events, industries/sectors and other types
  - **2,000** types of topic codes
- Product Codes
  - Identify which desktop news product(s) the news item belongs to. Tailored to Specific audiences
  - **649** types of product codes

Metadata for a News Item

- Take Sequence with the appropriate timestamp
- Topic Codes
  - Describe the news item's subject matter. These cover asset classes, geographies, events, industries/sectors and other types
  - **2,000** types of topic codes
- Product Codes
  - Identify which desktop news product(s) the news item belongs to. Tailored to Specific audiences
  - **649** types of product codes
- Named Item Codes
  - Identify news items that follow a pattern,also called recurring report codes
  - **2,354** types of named item codes

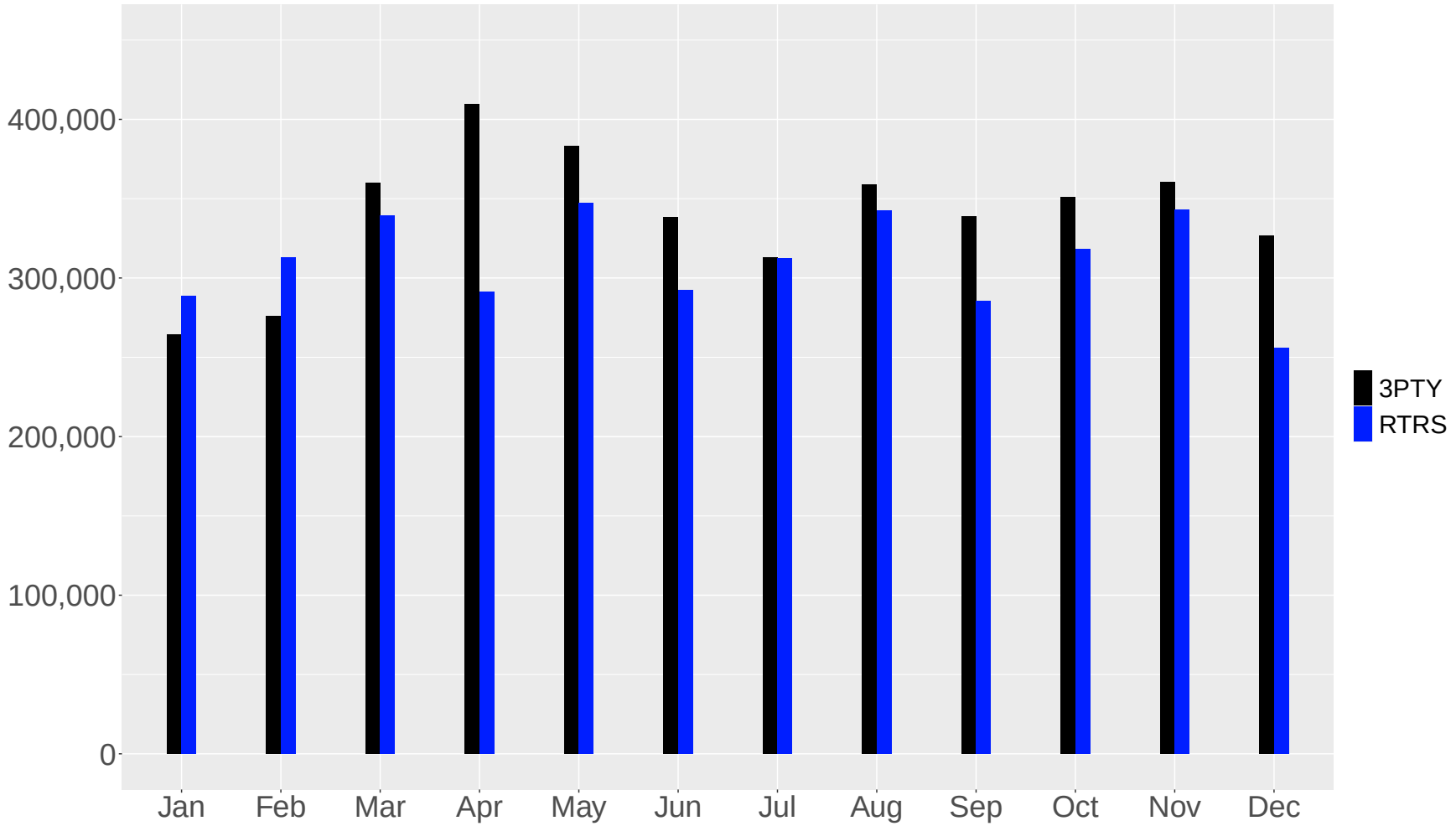
Metadata for a News Item

- Take Sequence with the appropriate timestamp
- Topic Codes
  - Describe the news item's subject matter. These cover asset classes, geographies, events, industries/sectors and other types
  - **2,000** types of topic codes
- Product Codes
  - Identify which desktop news product(s) the news item belongs to. Tailored to Specific audiences
  - **649** types of product codes
- Named Item Codes
  - Identify news items that follow a pattern,also called recurring report codes
  - **2,354** types of named item codes
- Attributions
  - Organizations that published the news item
  - Includes **49** Third Party Sources

Metadata for a News Item

- Take Sequence with the appropriate timestamp
- Topic Codes
  - Describe the news item's subject matter. These cover asset classes, geographies, events, industries/sectors and other types
  - **2,000** types of topic codes
- Product Codes
  - Identify which desktop news product(s) the news item belongs to. Tailored to Specific audiences
  - **649** types of product codes
- Named Item Codes
  - Identify news items that follow a pattern,also called recurring report codes
  - **2,354** types of named item codes
- Attributions
  - Organizations that published the news item
  - Includes **49** Third Party Sources
- RICs and Company PermIDs

Average # of News Items in a Month



- Average # RTRS News Items in a month is **310,756**
- Average # 3PTY News Items in a month is **339,981**

Third Party Sources

- There are **49** Third Party Service providers in the MRN offering.

#	Source
1	Asian Corporate Newswire
2	Actus News
3	Aktiengesellschaft für Wirtschaftspublikationen
4	Borsa Italia
5	Bank Negara Malaysia
6	Bombay Stock Exchange
7	Business Wire
8	Cision
9	Shenzen Securities Information Co. Ltd.
10	Comisión Nacional del Mercado de Valores
11	Canada Newswire
12	Copenhagen Stock Exchange
13	DGAP
14	DealWatch Debt
15	DealWatch Equity
16	News Aktuell
17	Economic Cycle Research Institute
18	eDaily News
19	eDaily Market Plus
20	Equity Story
21	Filing Services Canada Newswire
22	Globe Newswire
23	Globe Newswire Europe
24	Hong Kong Stock Exchange IIS
25	Helsinki Stock Exchange

#	Source
26	Iceland Stock Exchange
27	IFR Markets News
28	Japan Corporate News
29	London Stock Exchange
30	Market Wire
31	Money Today (Korean)
32	National Bank of Denmark
33	Swedish National Debt Office
34	NASDAQ OMX Saxess System Messages
35	PR Newswire
36	Swedish National Bank
37	Romeike
38	Riga Stock Exchange
39	Reuters News
40	Hong Kong Stock Exchange
41	Thailand Stock Exchange
42	Stockholm Stock Exchange
43	Thai Bond Market Association
44	Teikoku Databank
45	Tensid
46	Tijd Nieuwslijn (Dutch)
47	Tallinn Stock Exchange
48	Vilnius Stock Exchange
49	Weather Services Corporation

Topic Codes

TRNA supports 1,400 Topic codes

Topic Code Distribution

Topic Code	Description	TRCS Concept Type	# occurrences	%
LEN	English	Language	5,220,924	67
CMPNY	Company News	Broad News Topic	4,224,378	54
EMRG	Emerging Market Countries	Geography	3,083,588	40
ASIA	Asia / Pacific	Geography	3,027,083	39
NEWR	News Announcements	Genre	2,486,461	32
EUROP	Europe	Geography	2,011,069	26
BACT	Corporate Events	Event Type	1,977,579	25
AMERS	Americas	Geography	1,796,762	23
REG	Regulatory Corporate News Announcements	Genre	1,761,565	23
WEU	Western Europe	Geography	1,707,222	22
US	United States	Geography	1,688,541	22
STX	Equities Markets	Asset Class / Property	1,580,817	20
CN	China (PRC)	Geography	1,557,514	20
FINS	Financials (TRBC)	Business Sector	1,457,894	19
HK	Hong Kong	Geography	1,118,456	14
LZH	Chinese	Language	1,014,339	13
RES	Performance / Results / Earnings	Event Type	1,001,061	13
GEN	General News	Broad News Topic	925,128	12
INDS	Industrials (TRBC)	Business Sector	894,126	11
CYCS	Cyclical Consumer Goods & Services (TRBC)	Business Sector	871,101	11
BISV	Banking & Investment Services (TRBC)	Business Sector	837,687	11
TMT	Technology / Media / Telecoms	Business Sector	805,253	10
GB	United Kingdom	Geography	733,114	9
EZC	Euro Zone	Geography	692,590	9
REP	Reports	News Flag / Status	686,735	9
CEEU	Central / Eastern Europe	Geography	667,646	9
INDU	Corporate Events	Event Type	626,247	8
TECH	Technology (TRBC)	Business Sector	614,756	8

Topic Code Distribution

Topic Code	Description	TRCS Concept Type	# occurrences	%
MCE	Economic News	Event Type	583,052	7
POL	Government / Politics	International Affairs	549,331	7
BMAT	Basic Materials (TRBC)	Business Sector	492,496	6
DBT	Debt / Fixed Income Markets	Asset Class / Property	479,585	6
BSVC	Banking Services (TRBC)	Business Sector	455,001	6
DEAL1	Deals	Event Type	453,924	6
ENER	Energy (TRBC)	Business Sector	421,846	5
COM	Commodities Markets	Asset Class / Property	414,756	5
INVS	Investment Banking & Investment Services (TRBC)	Business Sector	414,226	5
HECA	Healthcare (TRBC)	Business Sector	400,946	5
ISER	Industrial Services (TRBC)	Business Sector	398,257	5
TEEQ	Technology Equipment (TRBC)	Business Sector	379,312	5
INDG	Industrial Goods (TRBC)	Business Sector	378,345	5
CEN	Central Banks / Central Bank Events	Broad News Topic	368,115	5
MIN	Mining	Business Sector	366,603	5
CCOS	Cyclical Consumer Services (TRBC)	Business Sector	350,966	4
MINE	Mineral Resources (TRBC)	Business Sector	332,452	4
LKO	Korean	Language	330,385	4
CDTY	Commodities Markets	Asset Class / Property	324,429	4
RESF	Results Forecasts / Warnings	Event Type	321,673	4
MEAST	Middle East	Geography	318,906	4
IN	India	Geography	318,303	4
MRG	Mergers / Acquisitions / Takeovers	Event Type	315,993	4
MTAL	Metals & Mining (TRBC)	Business Sector	315,329	4
ECON	Economy	Event Type	314,892	4
CA	Canada	Geography	314,204	4
CYCP	Cyclical Consumer Products (TRBC)	Business Sector	313,626	4

Commercial Packages

Commercial Packages

- Real-Time-News : Reuters News
  - Company News: AMERS, EMEA, APAC
  - Political / General / Economic News

Commercial Packages

- Real-Time-News : Reuters News
  - Company News: AMERS, EMEA, APAC
  - Political / General / Economic News
- Real-Time-News : Third-Party Corporate and Regulatory News
  - 36 third party sources, like Business Wire, PR Newswire, Filing Services Canada
  - Top five Press wires

Commercial Packages

- Real-Time-News : Reuters News
  - Company News: AMERS, EMEA, APAC
  - Political / General / Economic News
- Real-Time-News : Third-Party Corporate and Regulatory News
  - 36 third party sources, like Business Wire, PR Newswire, Filing Services Canada
  - Top five Press wires
- Real-Time-News : Reuters News, Third-Party Corporate and Regulatory News

Commercial Packages

- Real-Time-News : Reuters News
  - Company News: AMERS, EMEA, APAC
  - Political / General / Economic News
- Real-Time-News : Third-Party Corporate and Regulatory News
  - 36 third party sources, like Business Wire, PR Newswire, Filing Services Canada
  - Top five Press wires
- Real-Time-News : Reuters News, Third-Party Corporate and Regulatory News
- News Archive with millisecond timestamps back to 1996 for Reuters, 2003 for third parties

Commercial Packages

- Real-Time-News : Reuters News
  - Company News: AMERS, EMEA, APAC
  - Political / General / Economic News
- Real-Time-News : Third-Party Corporate and Regulatory News
  - 36 third party sources, like Business Wire, PR Newswire, Filing Services Canada
  - Top five Press wires
- Real-Time-News : Reuters News, Third-Party Corporate and Regulatory News
- News Archive with millisecond timestamps back to 1996 for Reuters, 2003 for third parties
- Economic Indicators from government & private agencies
  - Packaged by region: Global, US, Europe, Canada, APAC
  - High-value content: ISM, Markit PMIs, API, IPSOS

Commercial Packages

- Real-Time-News : Reuters News
  - Company News: AMERS, EMEA, APAC
  - Political / General / Economic News
- Real-Time-News : Third-Party Corporate and Regulatory News
  - 36 third party sources, like Business Wire, PR Newswire, Filing Services Canada
  - Top five Press wires
- Real-Time-News : Reuters News, Third-Party Corporate and Regulatory News
- News Archive with millisecond timestamps back to 1996 for Reuters, 2003 for third parties
- Economic Indicators from government & private agencies
  - Packaged by region: Global, US, Europe, Canada, APAC
  - High-value content: ISM, Markit PMIs, API, IPSOS
- Delivery Options : Real Time Delivery and Historical News Archive

MRN Demo

Usecases

- FX Trading based on Sentiments model using real-time Macro news and Economic Events
- Building Sentiment on Bond Issuers using Macro news.
- Generating Corporate Client Insights for banking RM's from News and other unstructured data sources
- Market Surveillance/Abuse – Insider Dealing, Benchmark Monitoring etc
- Adverse Media Check – For AFT/CFT Compliance
- News Event Detection

MRN Hands-on

Exploring Sample MRN Data

Write sample code to answer the following questions:

- How many days of the news items are covered in the dataset ?
- List down the top 5 languages by item count
- List down the top 4 Topics by item count
- List down the top 5 Sources by item count
- What is the proportion of alerts to articles in the dataset ?
- What is the maximum number of alerts reported for any story in the dataset?

MRN Hands-on-Solutions

http://bit.ly/2YOtFQQ

Q&A

References

- BattleFin

  https://www.battlefin.com/

- Alternate Data Trends

  https://www.neudata.co/

- Machine Readable News

  https://www.refinitiv.com/en/products/reuters-and-third-party-news

- News Analytics - Sentiment Data for Equities and Commodities

  https://www.refinitiv.com/en/products/world-news-data

- Market Psych Sentiment Indices

  https://www.marketpsych.com/

- Transcripts and Briefts

  https://www.refinitiv.com/en/financial-data/company-data/company-events-coverage

- ESG

  https://www.refinitiv.com/en/financial-data/company-data/esg-research-data

References

- Surveys and Polls

  `https://www.refinitiv.com/en/financial-data/market-data/economic-data`
- PermID

  `https://permid.org/`
- Intelligent Tagging

  `https://permid.org/onecalaisViewer`, `https://developers.refinitiv.com/open-permid`
- Knowledge Graph

  `https://www.refinitiv.com/en/products/knowledge-graph-feed`,

  `https://developers.refinitiv.com/open-permid/intelligent-tagging-restful-api`
- spaCy

  `https://spacy.io/api`
- BERT

  `https://github.com/google-research/bert`
- XLNet Pretrained model

  `https://github.com/zihangdai/xlnet`

References

- Latent Semantic Analysis

  https://www.asc.ohio-state.edu/reidy.16/LSAtutorial.pdf

- Probabilistic Latent Semantic Analysis

  https://arxiv.org/ftp/arxiv/papers/1301/1301.6705.pdf

- Latent Dirichlet Allocation

  http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

- Time Series Analysis by State Space Methods

  https://www.amazon.com/Time-Analysis-State-Space-Methods/dp/019964117X

- Dynamic Linear Models in R

  https://www.amazon.com/Dynamic-Linear-Models-Giovanni-Petris/dp/0387772375

- `statsmodels.tsa.statespace` package in Python

  https://www.statsmodels.org/dev/statespace.html