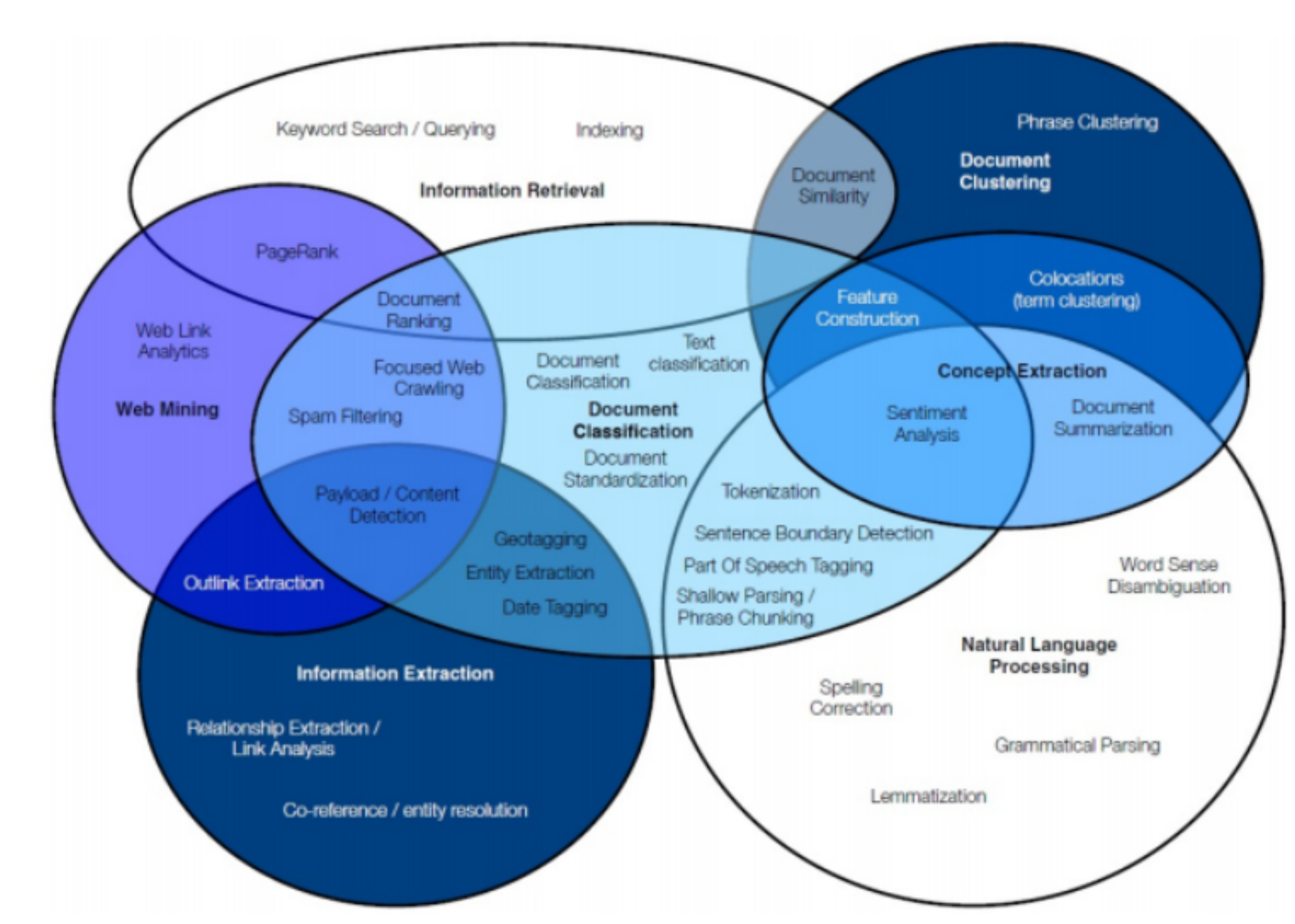
Information Extraction/Text Mining

-Rashmi Shivanna

What is Text Mining?

The Oxford English Dictionary defines text mining as the process or practice of examining large collections of written resources in order to generate new information, typically using specialized computer software. It is a subset of the larger field of data mining.

Inter-relationship among different text mining techniques and their core functionalities:-
Ref https://thesai.org/Downloads/Volume7No11/Paper_53-Text_Mining_Techniques_Applications_and_Issues.pdf



Information Extraction(IE):

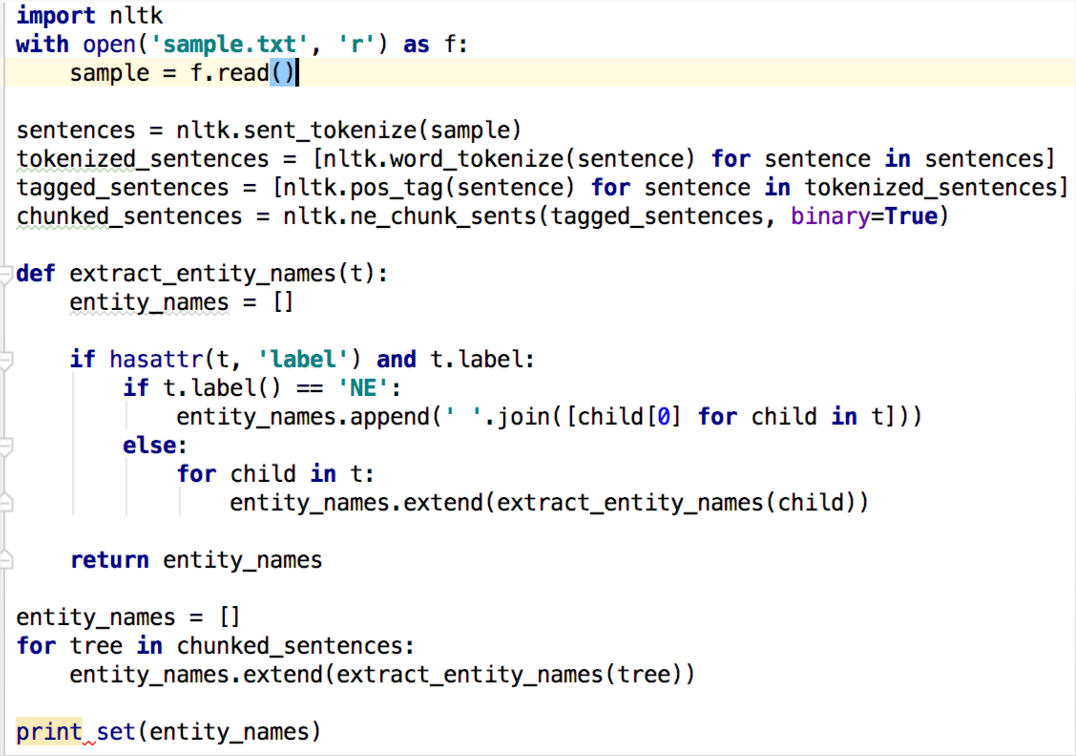
One of the techniques used in text mining. It involves automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. In most of the cases this activity concerns processing human language texts by means of natural language processing (NLP). Ref https://en.wikipedia.org/wiki/Information_extraction

Tasks involved in Information Extraction:

- **Named entity extraction:**
  - **Named entity recognition:** Identifying items in text that belong to predefined categories, such as the names of people, organizations, location, and monetary value. For example, in the sentence "Bill Smith acquired 51% of the outstanding shares of XYZ Inc." - Bill Smith would be identified as the name of a person - 51% would be identified as a percentage - XYZ Inc. would be identified as the names of an organization
  - **Coreference resolution:** Deriving the correct interpretation of text by connecting pronouns to the right individuals. For example, Gary is an investor. He invests all the time. "He" should be connected to "Gary."
  - **Relationship extraction:** Finding links between previously extracted named entities. For example, all entities including dates, locations, and people associated with the same meetings.
- **Semi-structured information extraction**
  - Table extraction: finding and extracting tables from documents.
  - Comments extraction : extracting comments from actual content of article in order to restore the link between author of each sentence
- **Language and vocabulary analysis**
  - Terminology extraction: Finding the relevant terms for a given corpus
- **Audio extraction:**
  - Template-based music extraction: finding relevant characteristic in an audio signal taken from a given repertoire;

NLTK for named entity extraction:

- NLTK provides packages which can be used to extract named entities from a given text
- The following program tokenizes each line from a file, splits it into chunks and looks for NE (named entity) labels for every chunk recursively.
- For e.g., it was given the input “French journalist Paul Legall reported that all three hostages arrived safely at Athens International Airport” among which it identifies the named entities Paul Legall, Athens International Airport and French. Ref https://gist.github.com/onyxfish/322906



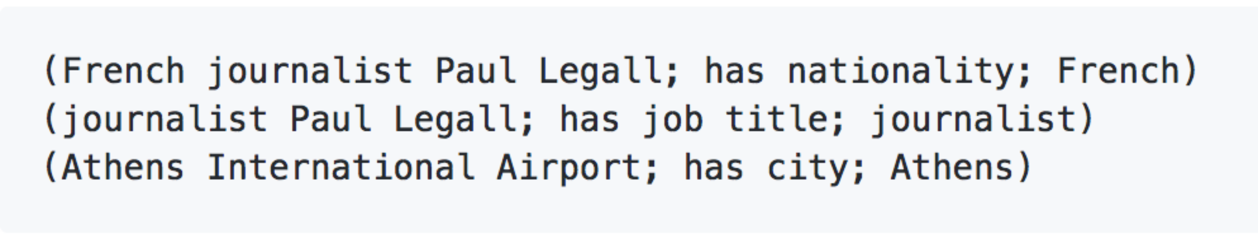
```
import nltk
with open('sample.txt', 'r') as f:
    sample = f.read()

sentences = nltk.sent_tokenize(sample)
tokenized_sentences = [nltk.word_tokenize(sentence) for sentence in sentences]
tagged_sentences = [nltk.pos_tag(sentence) for sentence in tokenized_sentences]
chunked_sentences = nltk.ne_chunk_sents(tagged_sentences, binary=True)

def extract_entity_names(t):
    entity_names = []

    if hasattr(t, 'label') and t.label:
        if t.label() == 'NE':
            entity_names.append(' '.join([child[0] for child in t]))
        else:
            for child in t:
                entity_names.extend(extract_entity_names(child))

    return entity_names

entity_names = []
for tree in chunked_sentences:
    entity_names.extend(extract_entity_names(tree))

print set(entity_names)
```

Output: set(['Paul Legall', 'Athens International Airport', 'French'])

Open Source tools for IE:

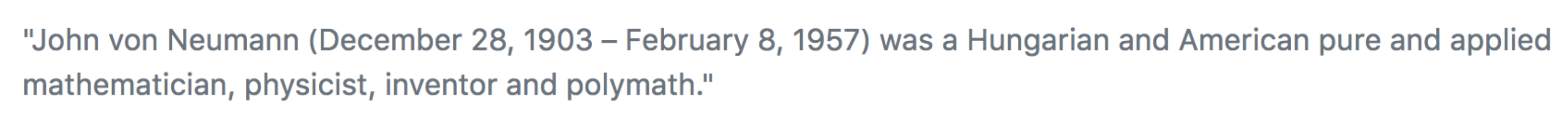
Many APIs have been developed for performing information extraction and few of them are discussed here.

- **MITIE: MIT Information Extraction**
  - Provides free (even for commercial use) state-of-the-art information extraction tools.
  - Includes tools for performing named entity extraction and binary relation detection as well as tools for training custom extractors and relation detectors. Ref https://github.com/mit-nlp/MITIE
- **MPLIE (IMPLicit relation Information Extraction)**
  - Extracts binary relations from English sentences where the relationship between the two entities is not explicitly stated in the text
  - Supports the following target relations out-of-the-box: *has nationality*, *has job title*, *has province*, *has city*, and *has religion*. Ref https://github.com/knowitall/implie . Following is an example of how MPLIE extracts relations
    - Sentence given: French journalist Paul Legall reported that all three hostages arrived safely at Athens International Airport.
    - The out-of-the-box IMPLIE system extracts the following binary relations:



```
(French journalist Paul Legall; has nationality; French)
(journalist Paul Legall; has job title; journalist)
(Athens International Airport; has city; Athens)
```

- **IEPY**
  - IEPY is an open source tool for information extraction focused on Relation Extraction. Ref https://github.com/machinalis/iepy
  - To give an example of Relation Extraction, if we are trying to find a birth date in



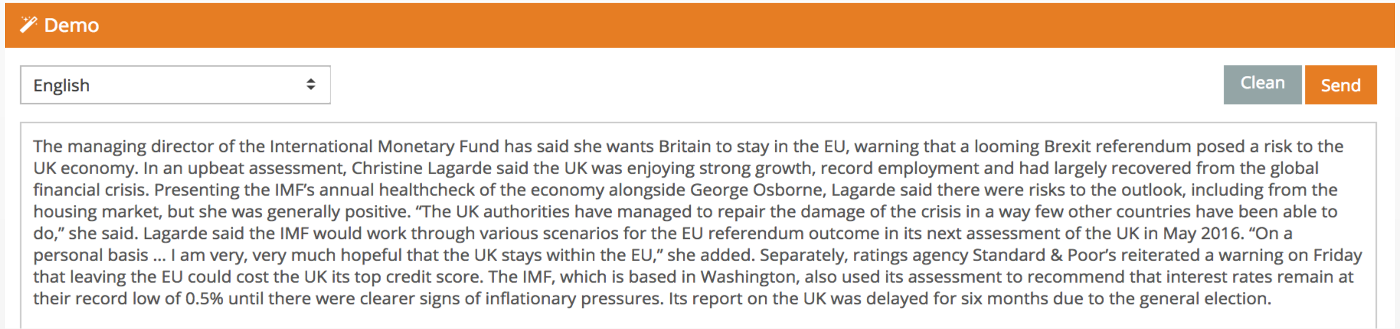
"John von Neumann (December 28, 1903 – February 8, 1957) was a Hungarian and American pure and applied mathematician, physicist, inventor and polymath."

  - Then IEPY's task is to identify "John von Neumann" and "December 28, 1903" as the subject and object entities of the "was born in" relation.

Online tool demo for IE

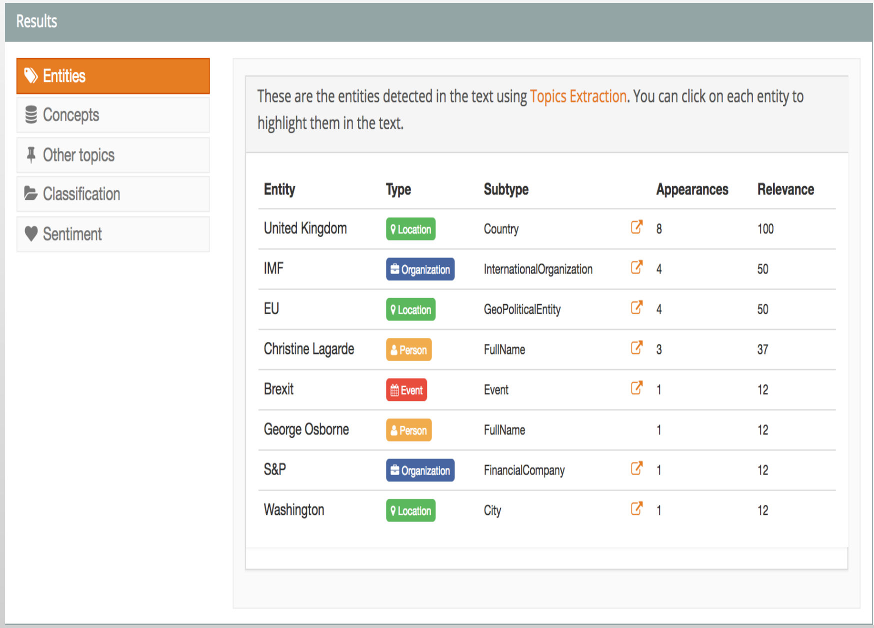
MeaningCloud provides a tool for information extraction. A demo of it is provides using screenshots below. Ref https://www.meaningcloud.com/demos/text-analytics-demo

Text:



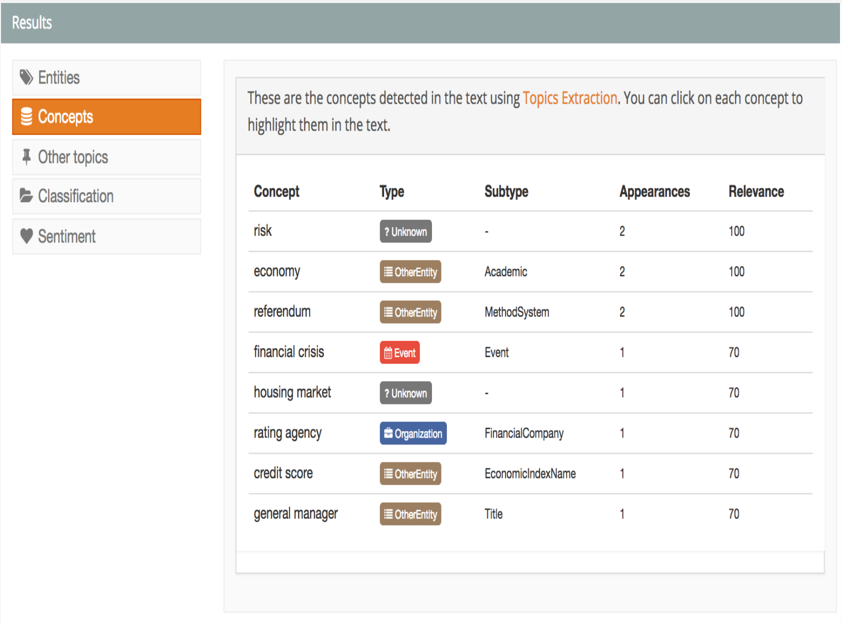
The managing director of the International Monetary Fund has said she wants Britain to stay in the EU, warning that a looming Brexit referendum posed a risk to the UK economy. In an upbeat assessment, Christine Lagarde said the UK was enjoying strong growth, record employment and had largely recovered from the global financial crisis. Presenting the IMF's annual healthcheck of the economy alongside George Osborne, Lagarde said there were risks to the outlook, including from the housing market, but she was generally positive. "The UK authorities have managed to repair the damage of the crisis in a way few other countries have been able to do," she said. Lagarde said the IMF would work through various scenarios for the EU referendum outcome in its next assessment of the UK in May 2016. "On a personal basis … I am very, very much hopeful that the UK stays within the EU," she added. Separately, ratings agency Standard & Poor's reiterated a warning on Friday that leaving the EU could cost the UK its top credit score. The IMF, which is based in Washington, also used its assessment to recommend that interest rates remain at their record low of 0.5% until there were clearer signs of inflationary pressures. Its report on the UK was delayed for six months due to the general election.

- **Topic extraction:** This tool extracts various types of entities from the text which includes named entities(people, organizations, places, etc., concepts(significant keywords), Time and money expressions**,** Quotes and Relations.



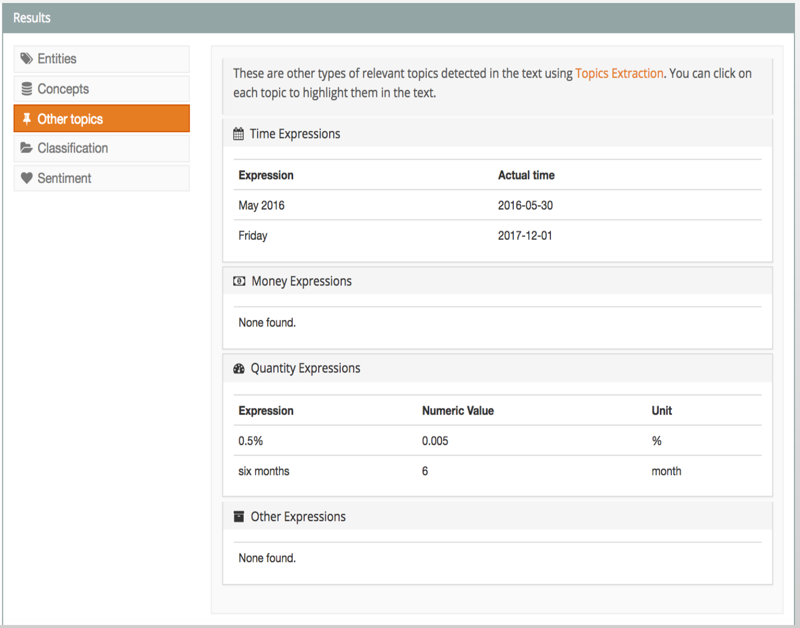
These are the entities detected in the text using Topics Extraction. You can click on each entity to highlight them in the text.

Entity	Type	Subtype		Appearances	Relevance
United Kingdom	Location	Country	↗	8	100
IMF	Organization	InternationalOrganization	↗	4	50
EU	Location	GeoPoliticalEntity	↗	4	50
Christine Lagarde	Person	FullName	↗	3	37
Brexit	Event	Event	↗	1	12
George Osborne	Person	FullName		1	12
S&P	Organization	FinancialCompany	↗	1	12
Washington	Location	City	↗	1	12



These are the concepts detected in the text using Topics Extraction. You can click on each concept to highlight them in the text.

Concept	Type	Subtype	Appearances	Relevance
risk	? Unknown	-	2	100
economy	OtherEntity	Academic	2	100
referendum	OtherEntity	MethodSystem	2	100
financial crisis	Event	Event	1	70
housing market	? Unknown	-	1	70
rating agency	Organization	FinancialCompany	1	70
credit score	OtherEntity	EconomicIndexName	1	70
general manager	OtherEntity	Title	1	70



These are other types of relevant topics detected in the text using Topics Extraction. You can click on each topic to highlight them in the text.

Time Expressions

Expression	Actual time
May 2016	2016-05-30
Friday	2017-12-01

Money Expressions

None found.

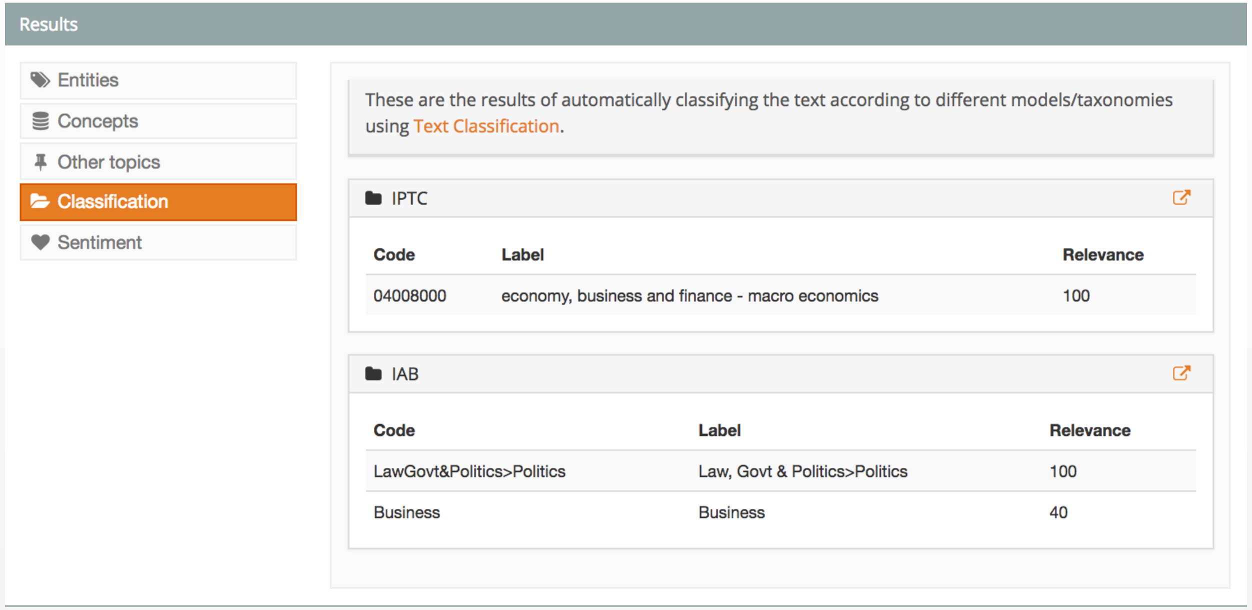
Quantity Expressions

Expression	Numeric Value	Unit
0.5%	0.005	%
six months	6	month

Other Expressions

None found.

- **Text Classification** is MeaningCloud's solution for automated document classification. It assigns one or more categories to a text, using standard domain-specific taxonomies (e.g., IPTC. IAB, ICD-10) or user-defined categories.



These are the results of automatically classifying the text according to different models/taxonomies using Text Classification.

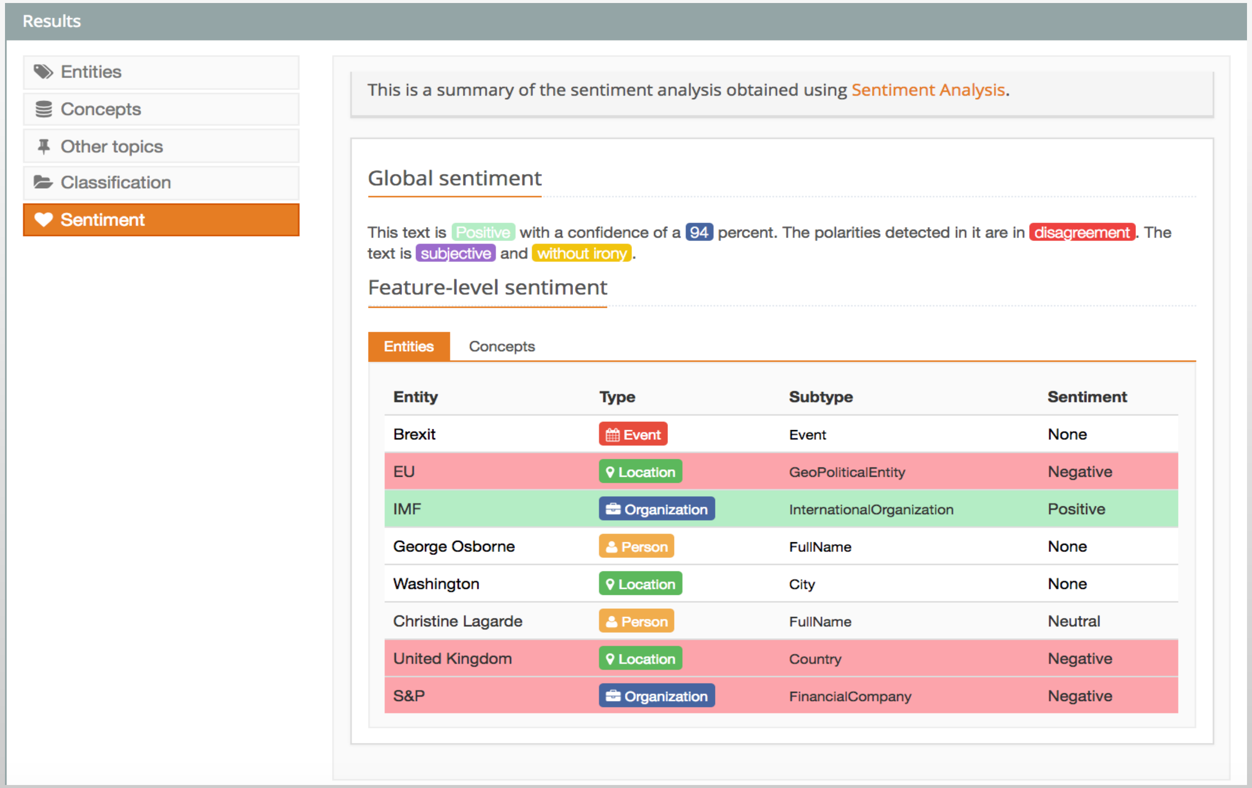
IPTC

Code	Label	Relevance
04008000	economy, business and finance - macro economics	100

IAB

Code	Label	Relevance
LawGovt&Politics>Politics	Law, Govt & Politics>Politics	100
Business	Business	40

- **Sentiment Analysis:** It identifies the positive, negative, neutral polarity in any text, including comments in surveys and social media.



This is a summary of the sentiment analysis obtained using Sentiment Analysis.

Global sentiment

This text is Positive with a confidence of a 94 percent. The polarities detected in it are in disagreement. The text is subjective and without irony.

Feature-level sentiment

Entity	Type	Subtype	Sentiment
Brexit	Event	Event	None
EU	Location	GeoPoliticalEntity	Negative
IMF	Organization	InternationalOrganization	Positive
George Osborne	Person	FullName	None
Washington	Location	City	None
Christine Lagarde	Person	FullName	Neutral
United Kingdom	Location	Country	Negative
S&P	Organization	FinancialCompany	Negative

Companies providing text mining solutions:

- Due to its capability of easing text processing, text mining is being extensively used across different sectors of IT. Many companies have developed software solutions for text mining.
- Text mining solutions are used in various fields like improving customer service of a company, improving efficiency in scanning documents for business critical data, etc.
- I have introduced 4 such companies such as Rocket Software, UchiData, iLexIR and SoftLaw, here which are currently providing text mining solutions for different kinds of customers.

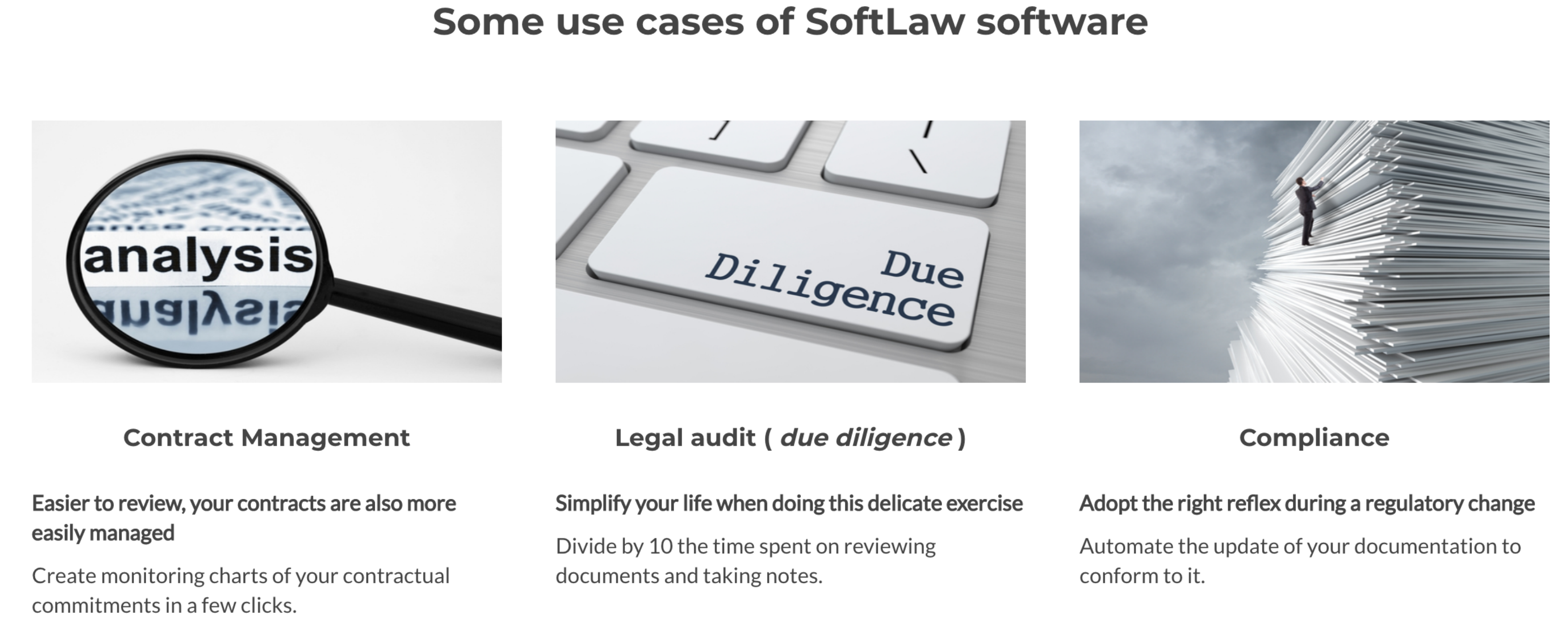
Rocket software

- Rocket® Enterprise Search and Text Analytics solutions make it easy for teams to find the information and data they need to get their work done, and our intuitive, flexible interface helps to ensure that they don't get bogged down in complex queries and incomprehensible results.
- Why not just use a text search engine or a relational database system to find the information you need to take action?

  In both cases, what you'll get is a very wide search versus one that is very deep. The search will be "wide" in terms of the number of items searched (and possibly found) but not very "deep" in terms of intelligence gained — i.e., in terms of learning new information you might want to know about these items.
- Key text analytics functions present:
  - Entity extraction: Explained before
  - Relationship Extraction: Explained before
  - Sentiment analysis: Classifying text based on emotional tone, such as positive, negative, or neutral.
  - Faceted Search. Progressively narrow your search via guided navigation and category drill down. For example, all competitors who sell products in various categories at various price points.
  - Clustering. Finding all documents that are related in some way without necessarily knowing ahead of time why they are related. For example, all documents related to a class action lawsuit.

SoftLaw

Develops text mining software that facilitates the review, analysis and processing of data contained in legal and administrative documents for legal professionals and legal and administrative services of companies of all sizes



Some use cases of SoftLaw software

Contract Management

Easier to review, your contracts are also more easily managed

Create monitoring charts of your contractual commitments in a few clicks.

Legal audit (*due diligence*)

Simplify your life when doing this delicate exercise

Divide by 10 the time spent on reviewing documents and taking notes.

Compliance

Adopt the right reflex during a regulatory change

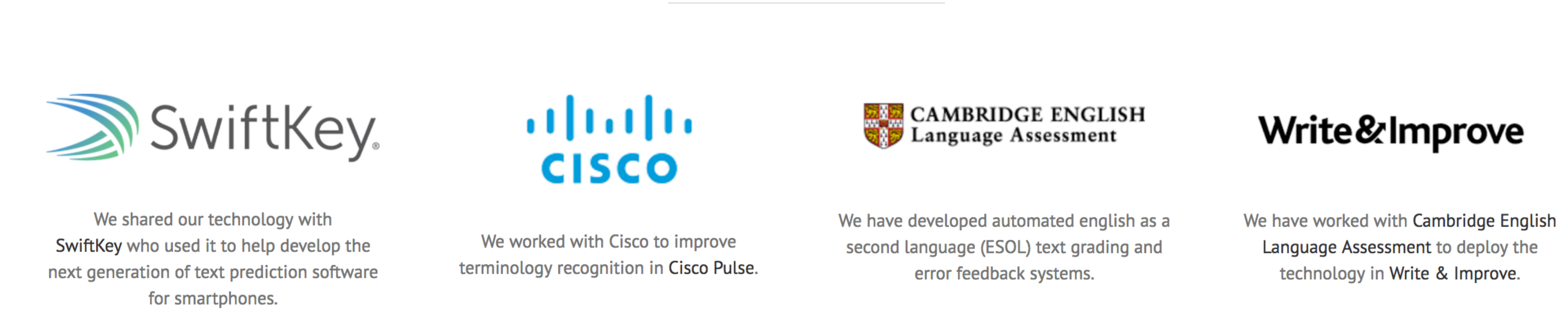
Automate the update of your documentation to conform to it.

Benefits of SoftLaw's software package

- Increased efficiency: Your documents are reviewed and analyzed by your teams in a timely manner.
- Improved productivity: Your teams can focus on higher value-added tasks.
- Gain of precision and serenity: You get a second look at your documents and check the information found in one click.

iLexIR:

- Specializes in text analytics, mining, classification, and search applications.
- Owns the sole commercial rights to the RASP system and search technology for developing such applications.
- They are a consultancy developing applications in partnership with technology providers and end users.
- Experience of about a century of research in natural language processing, computational linguistics, and its applications.”
- **Clients & Partnerships:**



We shared our technology with SwiftKey who used it to help develop the next generation of text prediction software for smartphones.

We worked with Cisco to improve terminology recognition in Cisco Pulse.

We have developed automated english as a second language (ESOL) text grading and error feedback systems.

We have worked with Cambridge English Language Assessment to deploy the technology in Write & Improve.

UchiData:

- The Uchidata API makes it easy to analyze, classify and tag texts in real time. Their state-of-the-art algorithms help us understand and extract valuable information from any documents.
- Solutions provided:
  - Reviews analysis: Analyze your customers' satisfaction and the quality of your service.
  - Sentiment mining: Determine customers preferences and intents by analyzing social conversations and web contents.
  - Topic extraction: Extract topics of documents or analyze users' profiles to guess their preferences.
  - Product classification: Analyze description of ecommerce products to classify them into your own referential

References:

- https://ischool.syr.edu/infospace/2013/04/23/what-is-text-mining/
- https://thesai.org/Downloads/Volume7No11/Paper_53-Text_Mining_Techniques_Applications_and_Issues.pdf
- https://gist.github.com/onyxfish/322906
- https://www.ventureradar.com/keyword/Text%20mining?#!
- http://www.rocketsoftware.com/about-us?rcid=mm-au-aboutus
- https://ilexir.co.uk/nlp-applications/index.html
- http://uchidata.com/#about-us
- https://www.softlaw.digital/index.html
- https://github.com/knowitall?utf8=%E2%9C%93&q=&type=&language=