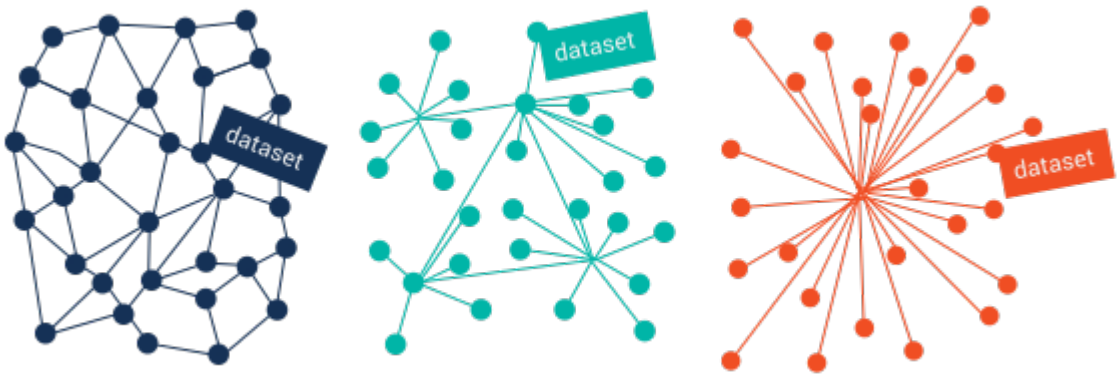
Graph databases and Text Mining work well together because you can apply Natural Language Processing to extract meaning from free flowing text and then store the results in a graph database to be further used for knowledge discovery and analysis.

Why Graph Databases?

Graph databases, and especially semantic graph databases (also called RDF triplestores), can smoothly integrate heterogeneous data from multiple sources and make it all interlinked. They can store hundreds of billions of facts (in the form of RDF triples) about any concept imaginable.

Today, there are many freely available interlinked facts from sources such as DBpedia, GeoNames, Wikidata and so on, and their number continues to grow as we speak. Some estimate their total between 150 and 200 billion right now. This freely available data is called Linked Open Data (LOD) and it can be a good source of information to power your graph database.

However, the real power of LOD comes when you transform your own data into RDF triples and then connect your proprietary knowledge to open world knowledge.



Another equally important functionality of graph databases is inference where new knowledge can be created from already existing facts. Here is a simple example of inference using two pre-existing facts: *Fido is a dog* and *a dog is a mammal.* From these facts, we can

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

infer that *Fido is also a mammal*. When such new facts are materialized and stored in a graph database, your search results become much more relevant, opening new avenues for actionable insights.

But if you want to add even more power to your data, you can use Text Mining techniques to extract the salient facts from free-flowing texts and then add them to the facts in your graph database. The resulting dataset is richer and much more useful to analyze, visualize, aggregate and report on.

Unfortunately, the market for this type of technology is still fragmented. Some vendors only sell graph databases and leave it up to you to determine how to do the Text Mining part. Others, offer just Text Mining and let you figure out where to store the results.

We at Ontotext, however, believe that graph databases and Text Mining should go hand in hand. Share on X Both of these technologies power our Ontotext Platform and are at the core of our Technology Solutions.

More About Text Mining

But what does a typical Text Mining process look like?

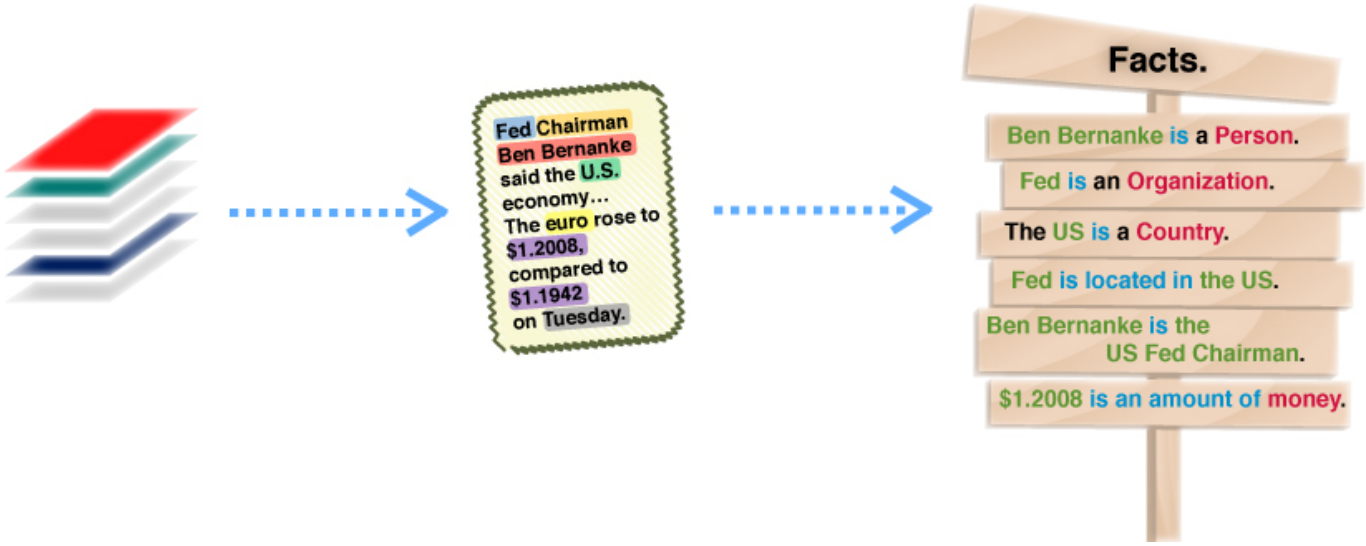
Here's a simplified explanation in 5 steps:

Step 1 – Extracting Concepts (Facts) from Free Flowing Text

In this step, the text from your unstructured data (such as documents, blogs, news articles, research reports, etc.) is processed by a Text Mining pipeline.

There are, of course, different stages that take place during this step such as sentence-splitting, tokenization, Named Entities Recognition (NER), etc. But, basically, it's about subjecting the text of the document to a series of actions, which break it down into smaller data pieces in order to add a layer of meaning to the raw content.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com



Step 2 – Enriching Concepts by Using Linked Open Data

As we have already mentioned, there are many freely available datasets that describe places, people, events, music, etc. You can use these sources of Linked Open Data to identify the concepts you have extracted from the text and get additional information about each of them.

For example, if your text mentions Bruce Springsteen, you can link the extracted concept to the MusicBrainz Linked Open Dataset. As a result, you will derive a wealth of additional information about the singer including songs he has written, albums, concerts, biographical information and much more.

In this way, enriching your original data with additional data opens even greater possibilities for Text Mining.

Step 3 – Identifying Relationships Between Concepts

The next step is to identify the relationships between the different concepts in your text. Here are some examples:

Sally worked at Banking Corp.
Gary lives in Tampa.

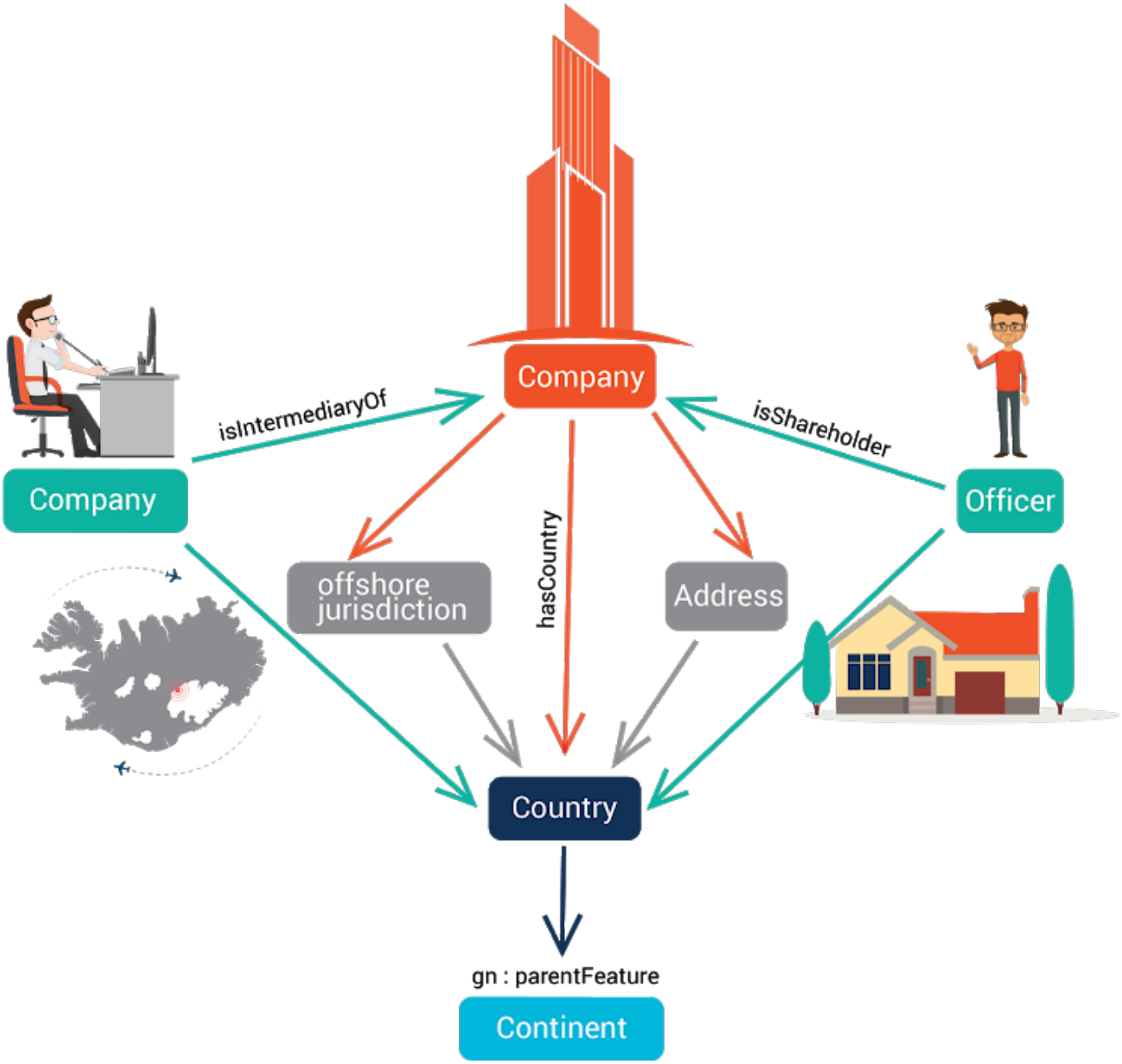
Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

Tamps is a city in Florida.
Gary worked with Sally.
Sally plays golf.

Some of these relationships are explicitly described in the text but some have been added through the process of LOD enrichment.

Because graph databases represent relationships in a Knowledge Graph structure, they can easily express not only the relationship between two facts in a statement but also multiple relationships across the whole dataset. Therefore, graph databases are ideal for storing facts extracted from text. It also makes them a powerful tool for relationship-centered analytics and knowledge discovery.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com



Step 4 – Disambiguating One Concept from Another

In this step, the concepts detected in the text are disambiguated and linked to their specific instances in the graph database. In other words, here you have to determine whether mentions in the text that appear to be similar refer to the same specific instance or not.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

For example, figuring out whether “Paris” in a text refers to the capital of France, or to Paris, Texas, or to Paris Hilton, or maybe even to the movie “Paris, Texas” is crucial for the correct understanding of this text. And without a good understanding, no reliable analysis can happen.

As ambiguity is a common problem in free-flowing text, the ability to instruct a computer how to learn which is the correct instance is very powerful.

Step 5 – Semantically Indexing Everything

In this final step, all facts and their original reference to the documents are indexed and stored in your graph database (together with the ontology – the model used for classifying the concepts and the relationships between them).

This provides robust capabilities for search and knowledge discovery. For example, after indexing your concepts and the quotation relationships in a dataset of news, you can ask what Trump has said about Putin.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

Important Takeaways

When you think of Text Mining in these simple terms, it's not too difficult to understand how the free-flowing text in your documents can be transformed into meaningful insights.

The ability to process unstructured data, transforming it into structured intelligence, and storing the results in a graph database along with a classification system, will give your business a huge competitive advantage. Share on X

Today, as organizations make the critical decision to discover the hidden meaning behind the massive amounts of unstructured data that lies in their different legacy systems, Text Mining is becoming increasingly important. And when they couple this process with the power of a graph database like Ontotext's GraphDB to query, aggregate, report and visualize this data, this fuels faster knowledge discovery and smarter decision-making.

Want to learn more about graph databases like Ontotext's GraphDB?