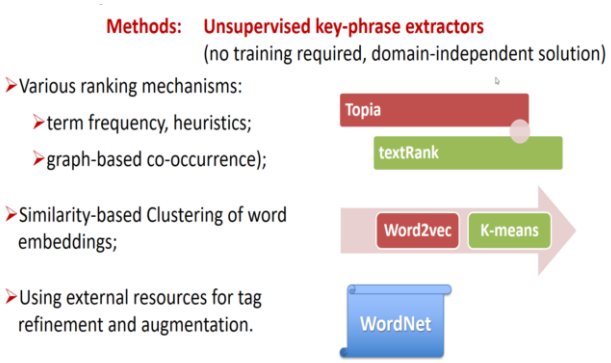
TagTell: Tag Detection in Large Scale Text Corpora

MARKET NEED

Many companies face an increase in data volume of unstructured text documents from customer feedback (surveys, customer support, chatlines, correspondence) or from company-specific documentation.

There is a need to quickly understand the content of such data and its main topics.

User-tagging of large volume of text data entries is expensive, time consuming and user-dependent.



TagTell approaches for automatic tag extraction using various key-phrase extraction methods.

A solution is to automatically assign tags to such datasets.

Tags can be extracted directly from the text data or can represent text content at a semantic level.

Tags can have multiple uses:

- Re-structuring data collections to support classification of data;
- Information retrieval based on new and more relevant attributes;
- Finding documents linked through common tags.

TECHNOLOGY SOLUTION

TagTell is an interactive web application based on unsupervised key-phrase extraction methods and Natural Language Processing.

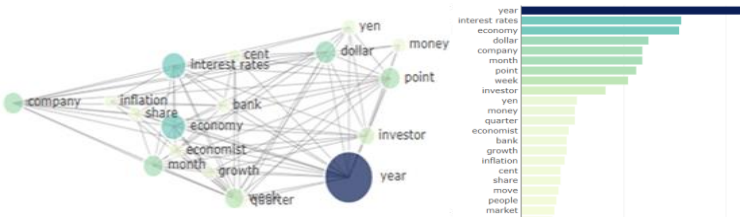
Three domain-independent, unsupervised approaches are implemented:

- Key-phrase extraction from text & various ranking mechanisms (term frequency, heuristics and graph-based co-occurrence);
- Similarity-based clustering of word embeddings;
- Using external resources for tag refinement and augmentation (Wordnet lexicon hierarchy).

APPLICABILITY

The TagTell demonstrator provides an interactive platform for automatic tag extraction enabling:

- Tag extraction, ranking and visualization as a network graph;
- Tags updates (select, delete, replace tags with user-defined tags);
- Drill-down to find all documents with a certain tag;
- Download the data with added tags.



TagTell visualization illustrating most frequent tags extracted from a dataset of financial news articles.

RESEARCH TEAM

Jayadeep Kumar Sasikumar DIT

Jimmy Doré , DIT

Dr. Tamara Matthews, DIT

Prof. Sarah Jane Delany, DIT

www.ceadar.ie