A System For Large Scale Data Exploration and Organization

Luis Rei, Blaž Fortuna, Marko Grobelnik, and Dunja Mladenić

Artificial Intelligence Laboratory, Jožef Stefan Institute
and
Jožef Stefan International Postgraduate School

http://github.com/lrei/elycite

Objectives

- Understand and organize large datasets
- Visual
- Semi-Automatic: Man + Machine
- Make Machine Learning easy to use
- Collaborative
- Interactive
- Easy to integrate with other software & services
- Powerful – give expert users fine-grained control

Elycite

- *Elycites* information from unlabelled textual data
- Organizes information as an ontology...
  - Displays it using multiple visualizations
- ... Using Machine Learning interactively
- Multiple clients, one server
- Web Interface (HTML + JavaScript)
- Everything (machine learning, data access, organization, ...) that can be done interactively can also be done using a REST API (Web Services)

Built with QMiner

- A data analytics platform for processing large-scale real-time streams containing structured and unstructured data
- In-memory database with indexing, search, aggregations and Machine Learning built-in
- Text Mining, Social Networks, Time Series, ...
- JavaScript API, C++ library
- JSON-like data: Strings, Numbers, Booleans, Dates, Vectors
- Open Source
- Runs the server component of Elycite
- **http://qminer.ijs.si**

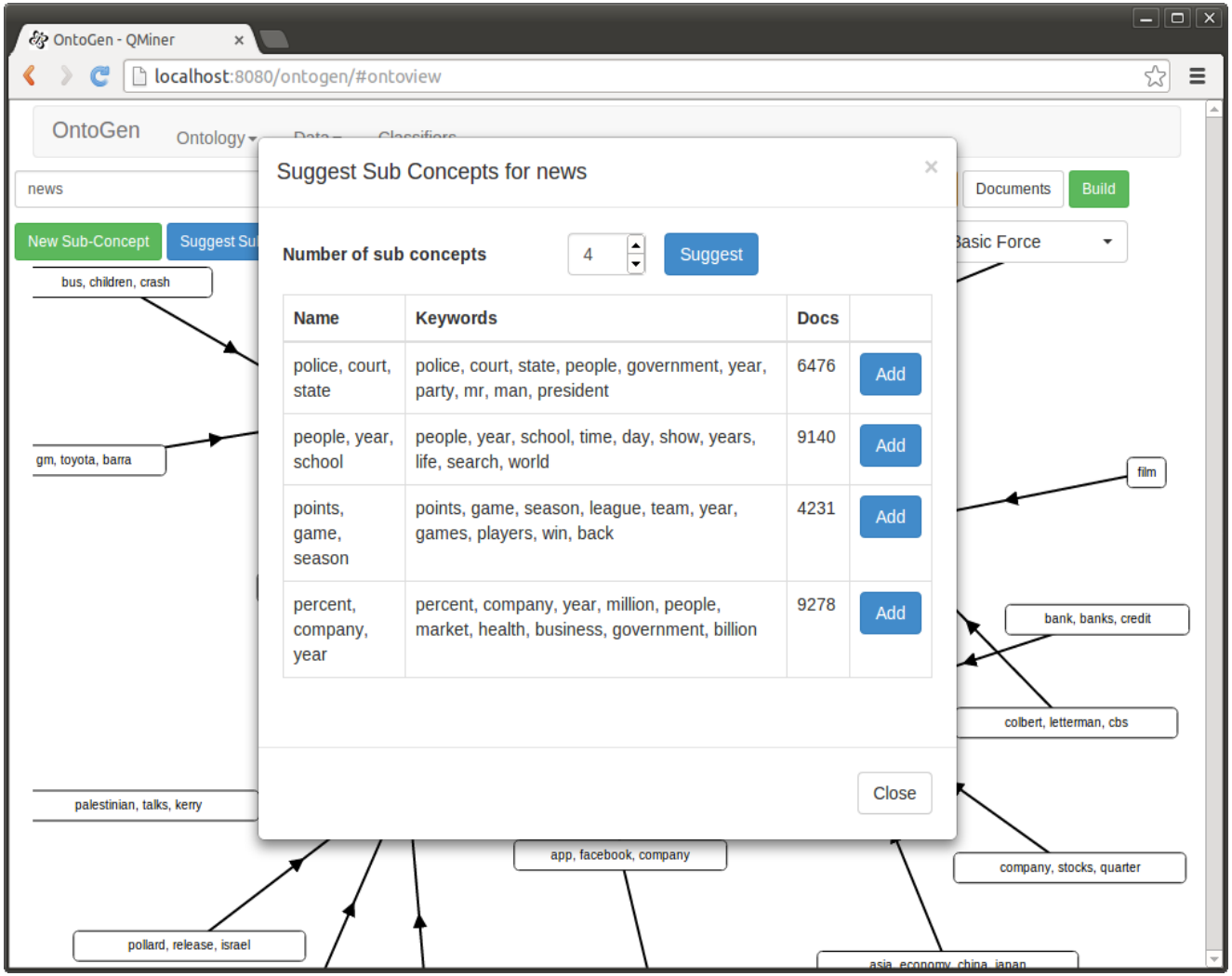
Elycite Basics

- **Documents** (Qminer Records)
- Organized into **Concepts** **(sets of documents)**
- Organized with relationships **- Ontology**
- Semi-automatic or manual
- E.g.:
  - **Documents**: News articles
  - **Concepts**: Politics, Economy, Entertainment, Sports, ...
  - **Hierarchy**: News -> Sports -> Football -> World Cup, Injuries, Commentary, FC Porto, ...

Unsupervised Methods

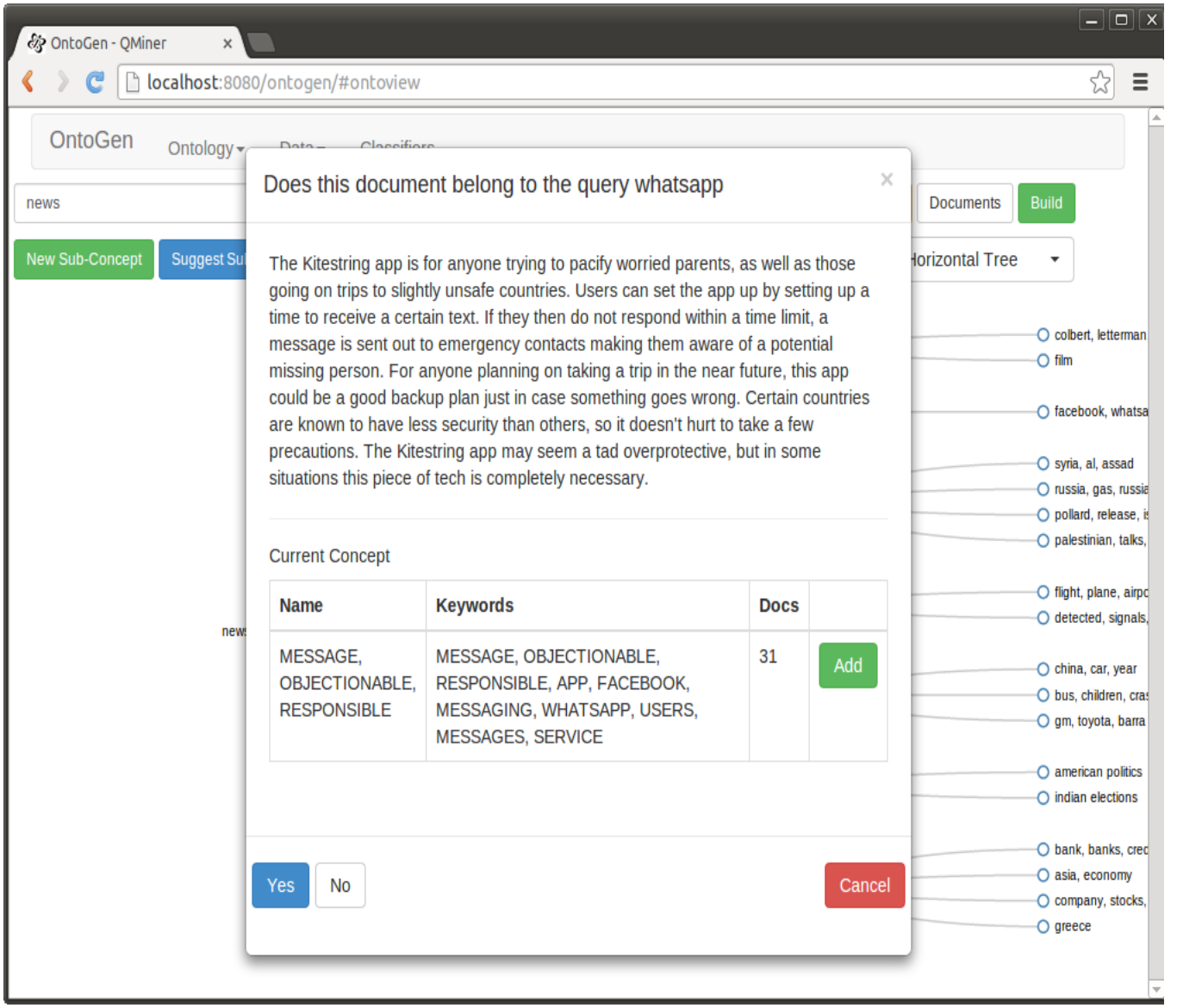
TF/IDF keyword extraction is used throughout the application to provide a summary of a set of documents (e.g. a concept).

KMeans++ based unsupervised concept suggestion.



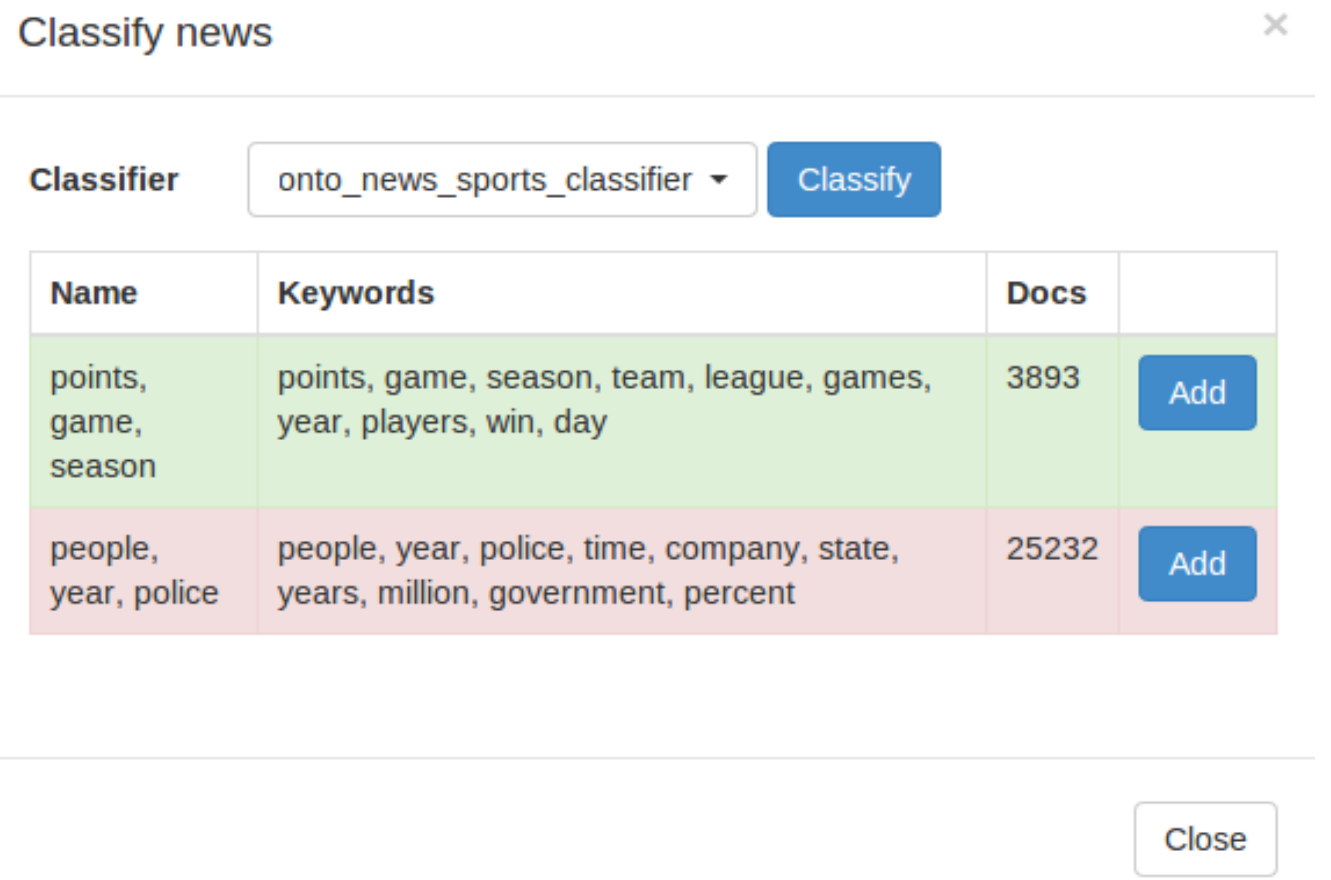
Semi-Supervised Methods

SVM based Active Learning with user supervision provided initially by a query followed by a sequence of Yes or No questions. The concept is refined after each answer, the user can decide when to stop depending on how satisfied he is with the current suggestion (based on the keywords an number of documents).

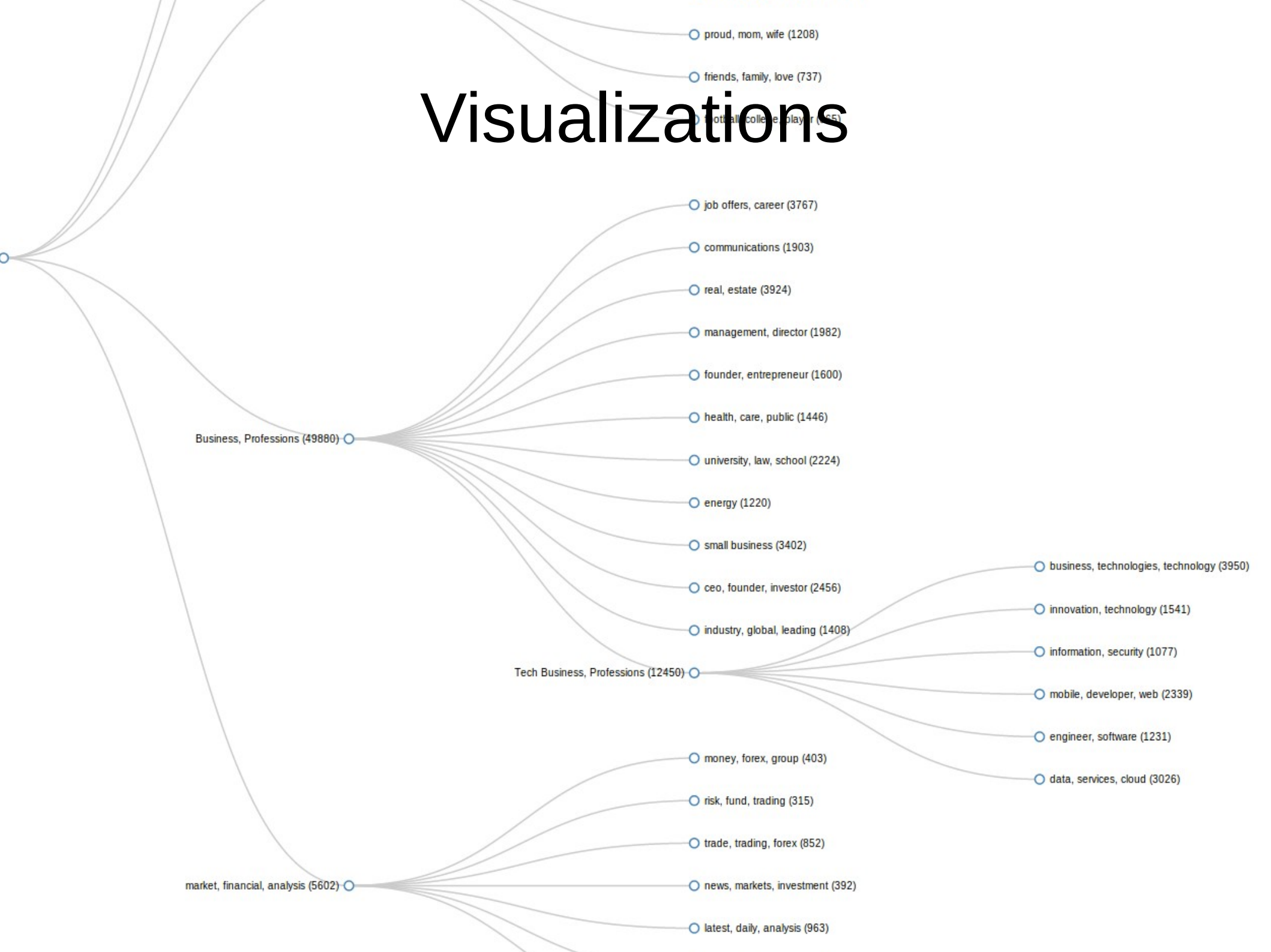


Supervised Methods

A **SVM** classifier can be built from a concept and used to classify other sub-concepts in the same ontology, create concepts in another ontology, or as service to classify new documents for another application.



Visualizations



Future Work

- Interactively handle larger datasets
- More visualizations
- Improve text preprocessing (n-grams, hashing)
- Guided Learning
- Pre-built Classifiers (DMOZ topics, sentiment)
- Data: Numerical, Graph, Image, Time Series
  - Preprocessing
  - Feature extraction/learning
  - Machine Learning (Unsupervised, Semi-supervised, supervised)
  - Visualizations