Master thesis ideas @ Findwise 2014

## For more information contact Svetoslav Marinov (Svetoslav.marinov@findwise.com)

# Unsupervised co-reference resolution

Exploit unsupervised machine learning methods for co-reference resolution. Given information such a Part-of-Speech tags on words, dependency-based syntactic parse trees, word clusters, Named Entities, etc, propose a system that will be able to extract co-reference chains from free text. Test data for Swedish and English can be provided to test the performance.

# Supervised co-reference resolution

Using already annotated data for Swedish and English, create a system to extract co-reference chains from text. The system should rely on the state-of-the-art as demonstrated at CoNLL Shared Task 2012 and possibly improve it. Alternative route can be to exploit the possibilities which Apache OpenNLP provides and extend these.

# Document summarization

Findwise already has a document summarization model, built relying on the TextRank algorithm. The task is to improve it by incorporating additional information such as Named Entities, co-reference, temporal expressions, domain, document class, etc.

# Vector space models for Information Retrieval

Evaluate Googles Word2vec & Semanticvectors för search med Solr

# Query language definition and query parsing

Most often queries are short, 2-3 words long. However, in some applications and scenarios it is required that people write a longer query or even a whole sentence. An example application is Facebook’s graph search. The aim of this thesis is to build a model and possibly a fast and robust query analyzer. It should be able to extract different kinds of entities from queries, as well as have fallback when the query is unparsable.

# Extraction of semantic relations

Design a method to extract relations between entities from free text. Evaluate dependency parsers such as MaltParser. The ultimate goal is to be able to identify the entities and which relation holds between these (e.g. Obama – visit – France, from the sentence “the US President Obama will be visiting France next week.”) and store these as RDF triples.

# Named Entity disambiguation

Names of people, places and products may be written identical but in reality may refer to different real world entities. The task is to use domain specific and contextual information to disambiguate which real world entity they may refer to.

# Multilingual search

Search in one language but get results from different languages. Alternatively present some of the results as a summary, which is translated based in the language of the query.

# Sentiment analysis

Create a module to do sentiment analysis on Twitter & FB and try to extract trends on that.

# Expertise extraction

Can we discover experts from free text? People in companies tend to respond, comment and involve in a lot of projects regarding different expert areas, however in the person database all this information is missing. The task is create a module that given some text sources of different authors will be able to identify what expertise the authors have.

# 11. OCR

Evaluate software and models for Optical Character Recognition. A lot of data nowadays is available in scanned documents. How much structured information can we extract from such documents with existing software (open source and proprietory)? What is the certainty of finding the title, author and date in a scanned document? What else can we extract with high precision?