Topic 1: Automated Classification of Estonian Online Media Articles

Alexander Tkachenko
alex.tk.fb@gmail.com
Lihtööstuse juht: eestlased söövad peamiselt välismaist liha

Atria Eesti juhi Olle Hormi kinnitusel söövad enam kui pooled tarbijad välimaist liha ning ei oska seda kodumaisest eristada.

Savisaare välzel kulub tal korteri üürimiseks 700 eurot kuus

Tallinn loodab elanike kasvult neli miljonit eurot puhastulu saada (5)

Tallinn televisiooni juht registreeris Krossi nime sisaldava domeeni (4)

Röövel viis Jõhvi juveelipoest kuldeheit

Väidetavalt esitas Süuria Venemaale tõendid mässuliste seotusest keemiarünnakuga (6)

Ministeerium: ilmselt saavad ka haiglad üldarstiabi osutamise õiguse
Topic 1: Automated Classification of Estonian Online Media Articles

• Goal: explore different forms of document representation with respect to classification accuracy:
  – words, lemmas, n-grams, tf-idf weighting, article section weighting, named entities, tags, LSI/LDA/random indexing, clustering, etc.

• Input: a corpus collected from major Estonian online news outlets.

• The student will need to:
  – topic-tag a set of articles for training and evaluation purposes,
  – implement/compare several document representation schemes,
  – write a report summarizing experimental results.
Topic 2: Multi-label document classification

Alexander Tkachenko
alex.tk.fb@gmail.com
Box aims for NSA-resistant cloud security with customers holding the keys

After eight years of existence, file sharing service Box has built a huge user base—claiming 180,000 businesses, including 97 percent of the Fortune 500—by offering cloud storage and collaboration tools with top-notch security and regulatory compliance.

But while Box may be resistant to most criminal hackers, like most cloud storage companies, it must provide the government with customer data when it is forced to. For the vast majority of Box customers, that isn't likely to change. However, the company is developing a system for the most...

Naiad: A Timely Dataflow System

Naiad is a distributed system for executing data parallel, cyclic dataflow programs. It offers the high throughput of batch processors, the low latency of stream processors, and the ability to perform iterative and incremental computations. Although existing systems offer some of these features, applications...
Topic 2: Multi-label document classification

• Goal: implement and compare several existing methods of multi-label document classification in the context of Estonian online media domain.

• Input: a corpus collected from major Estonian online news outlets together with the topic tags assigned by the editors.

• The student will need to:
  – implement/compare several methods for multi-label classification,
  – write a report summarizing experimental results.