An Online Service for Topics and Trends Analysis in Medical Literature

Spyridon Kavvadias, George Drosatos and Eleni Kaldoudi

School of Medicine, Democritus University of Thrace, Alexandroupoli, Greece

Presenter: George Drosatos
The aim of this work

- Propose and develop an online and platform independent service to support topic modeling and trends analysis for the biomedical experts

- Why?
  - Biomedical sciences constitute from large amount of scientific publications each year
  - Literature topics and trends analysis gives insights on past and future research directions
  - These methods require considerable mathematical and programming background from the researchers

- Our platform integrates a variation of topic modeling algorithms and provides in an easy way the complete workflow process of topics and trends analysis
Background
Topic modeling

- Topic modeling refers to a suite of algorithms that provides methods for automatically organizing, understanding, searching and summarizing large electronic archives:
  - Discover the hidden themes in the collection
  - Annotate the documents according to these themes
  - Use annotations to organize, summarize, search and form predictions
- The large collection of data requires unsupervised probabilistic models
- The most popular of them is the Latent Dirichlet Allocation (LDA) and there are many other variants of this model
- Applications of topic modelling: genomic sequence, image classification, discussion themes in social networks and source code analysis
Latent Dirichlet Allocation (LDA)

- Each document is a mixture of corpus-wide topics
- Each topic is a distribution over words
- Each word is drawn from one of these topics

Topic modeling and trends analysis service

https://trends.duth.carre-project.eu/
Workflow of our service

- Generate the initial literature corpus, as a collection of papers
- Apply text preprocessing by performing transformations in the text
- Parametrize and apply the topic modeling algorithm to identify topics
- Topics labeling by the user to make them meaningful
- Trend analysis by plotting topics over time
Login page

- The researcher registers or signs-in in the service
Dashboard

- The researcher could have an overview of their experiments
Create Corpus

- The researcher upload a XML file with papers from the PubMed
List of corpora

- The researcher browses the available corpora
Preprocessing

The researcher applies Krovetz stemmer and uses title, abstract and keywords from corpus' papers.
List of preprocessed corpora

- The researcher browses the available preprocessed corpora
Topic modeling

- The researcher applies Mallet implementation of LDA for 40 topics and 4000 iterations.
List of topic modeling experiments

- The researcher browses the available topic modeling experiments

05/06/2018
Labeling process

- The researcher assigns labels to topics to make them meaningful for human interpretation
Visualization of results

- The researcher performs trend analysis and visualizes the topics per year
Implementation details
Implementation details

- **Backend**
  - Server: NodeJS
  - API: LoopBack framework
  - Database: MongoDB

- **Frontend**
  - Visual Interface: AngularJS
  - Graph visualizations: Chart.js and Vis.js library

- Integration in the backend with
  - Mallet ParallelTopicModel Java library
  - jLDADMM Java library
Open source

- The source code is available as a git repository

http://gogs.duth.carre-project.eu:3000/spiroskvd/tm-toolkit
Conclusions & Work in progress
Conclusions & Work in progress

We proposed a web-based service that allows biomedical researchers with no experience in data modeling and programming to execute topic modeling and trends analysis experiments.

Work in progress includes:
- Make our web service free of bugs
- Support more topic modeling algorithms with an easy mechanism to add new implementations of them
- Develop a mechanism that would add a batch of processes with different parameters with goal to select the appropriate one
- Perform an evaluation of our system regarding the system’s performance and the users’ satisfaction
Thank you!
This work was supported by the FP7-ICT project CARRE (No. 611140), funded in part by the European Commission and Greek National Matching funds (DUTH Grant No. 81442)

CARRE Project: Personalized patient empowerment and shared decision support for cardiorenal disease and comorbidities

www.carre-project.eu