Foundations of Text Mining for Biologists Workshop

Introduction

Text Mining (TM) is the automatic extraction of structured information from mainly free-form written content. TM uses various techniques from the computational linguistics and artificial intelligence domains to discover previously unknown information from text, as opposed to (web) search, where users' information needs are known beforehand. Biomedical text mining, sometimes referred to as BioNLP, is the application of text mining tools on biomedical literature to extract information regarding biological entities, processes and diseases. The ever-increasing growth of biomedical scientific literature has prompted the need for research and development of automatic text mining solutions and several international academic venues have been established for scientists and domain practitioners to present their work and exchange ideas.

In this 3-hour workshop we will cover the foundations of text mining systems and how they pertain to the biomedical domain. You will be provided with hands-on material to develop a lightweight text mining pipeline and learn how to evaluate and compare the performance of text mining applications. A number of BioNLP tools will be showcased to inspire you in developing your next big BioNLP tool!

Required Software

You are expected to bring your own laptop to the workshop, as the nature of the session will be interactive and hands-on. You will need the following tools and libraries installed on your device:

- GATE v8.2 or better (available for download from the GATE website, ?570 MB)
- Java (JDK) v7 or better (available for download from Oracle, ?200 MB)
- Your favourite text editor

Hands-on Material

Corpus

Create a corpus from one or more of the following documents' Abstract sections:

1. Characterisation of new intracellular membranes in Escherichia coli accompanying large scale over-production of the b subunit of F1Fo ATP synthase
   [plain full-text] [Gold Standard]
2. Construction of a novel hydroxyproline-producing recombinant Escherichia coli by introducing a proline 4-hydroxylase gene.
   [plain full-text] [Gold Standard]
3. Cloning of a Corynebacterium diphtheriae iron-repressible gene that shares sequence homology with the AhpC subunit of alkyl hydroperoxide reductase of Salmonella typhimurium.
   [plain full-text] [Gold Standard]
4. Tuberculosis vaccine strain Mycobacterium bovis BCG Russia is a natural recA mutant
   [plain full-text] [Gold Standard]
5. MutationFinder example

Pipeline Resources

You can download the pipeline resources as one ZIP file, or download individual files as below.

Gazetteers

Download and modify the provided gazetteers with the appropriate entities for your corpus.

- lists.def
- organisms.lst

Grammars

Download and modify the provided grammars with the appropriate JAPE rules for your corpus.

- main.jape
- organisms.jape

Recommended Readings

- H. Cunningham, V. Tablan, A. Roberts, K. Bontcheva (2013) Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. PLoS Comput Biol 9(2): e1002854. doi:10.1371/journal.pcbi.1002854 — http://tinyurl.com/gate-life-sci/
- Developing Language Processing Components with GATE Version 8 (a User Guide)
- Driving the GATE framework from Python