Foundations of Text Mining for Biologists

Python for Scientific Concordians Workshop Series

Facilitator:	Bahar Sateli
Date:	June 20^{th} 2016
Location:	Concordia University – Loyola Campus

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!

Recommended Background Knowledge

Although this workshop does not require any strong programming background, a fair knowledge of writing scripts and regular expressions is recommended.

Learning Outcomes

At the end of this workshop, you will accomplish the following goals:

- Understand the application of text mining techniques in biomedical literature
- Develop a lightweight text mining pipeline
- Examine and explain the performance of the developed pipeline

Outline

A (tentative) list of topics to be covered is as follows:

- Introduction to text mining techniques
- Syntactic and semantic analysis of text
- Introduction to the GATE framework
- Information Extraction
- Metrics in evaluation of text mining solutions
- Showcase of various BioNLP tools

Materials

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

The hands-on material will be made available online on the day of the workshop. Please make sure your device is properly connected to the campus Wi-Fi network.

Facilitator's Biography

Bahar Sateli is a doctoral researcher at Concordia University's Semantic Software Lab. She obtained her MSc degree in Software Engineering from Concordia University in 2012 and her BSc in Information Technology from Islamic Azad University of Tehran in 2009. Her research focuses on applications of semantic technologies, in particular text mining and semantic web, in knowledge-intensive domains. She has several years of academic and industrial work experience, including startups in Montréal, collaborations with the University of Jena, Germany, as well as Concordia's Centre for Structural and Functional Genomics. Her PhD topic is Semantic Publishing, which examines automatic knowledge extraction and formal modeling of scientific literature, with the ultimate goal of creating a semantically-rich knowledge base of queryable scholarly data. Most recently, her work won the Semantic Publishing Challenge at ESWC 2015 (Task 2 - Most Innovative Approach) and a Best Paper Award at the WWW 2015 SAVE-SD Workshop on enhancing scholarly data.