BIOF 395: Introduction to Text Mining

Between Electronic Medical Records and Electronic Health Records, PubMed, and collections of biomedical grant applications, there exist large quantities of medical information stored in databases waiting to be explored. Besides tables of numbers, medical records also contain a great amount of free-text paragraphs that are comprehensible to human readers but challenging to computers. Text mining is an interdisciplinary area that primarily combines advances in Natural Language Processing (NLP), Information Retrieval (IR), and Machine Learning (ML) to help the computers understand human written language and thus extract medical and clinical information from free-text records. This class aims to introduce fundamental subjects in text mining such as tokenization, named entity recognition (NER), grammars, parsing, relation extraction, and document classification. The class is oriented towards hands-on experience with Python and Natural Language Toolkit (NLTK).

Learning Objectives

- Learn basic programming in Python
- Master fundamental building blocks of Natural Language Processing
- Acquire hands-on experience with NLTK, a Python toolkit for NLP
- Gain an introduction to statistical models of Machine Learning applied to NLP and IR

Credits: 2
Class Type: Graduate Course
Prerequisites:
Prior exposure to programming and Python is encouraged but not required to attend this class
Program: Bioinformatics and Data Science