جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Course Syllabus: Application of AI in Bioinfomatics - CS 321

Division	Computer, Electrical and Mathematical Sciences & Engineering
Course Number	CS 321
Course Title	Application of AI in Bioinfomatics
Academic Semester	Spring
Academic Year	2018/2019
Semester Start Date	01/27/2019
Semester End Date	05/23/2019
Class Schedule (Days & Time)	09:00 AM - 12:00 PM | Sun

Instructor(s)

Name	Email	Phone	Office Location	Office Hours
Vladimir Bajic	vladimir.bajic@kaust.edu.sa	+966128082386	4219, 3, Ibn Sina (bldg. 3)	Wednesday 10.00 AM - 12:00 PM, Building 3, Floor 4, Room 4219 Phone: +966 (0)12 8087318

Teaching Assistant(s)

Name	Email
Dr. Christophe Van Neste	christophe.vanneste@kaust.edu.sa

Course Information

Comprehensive Course Description	**Summary** The course consists of selected projects. The projects may change each year. These projects cover the application of artificial intelligence (AI) to some of the relevant problems of analysis of large biological data and generally deal with complex information. Each year, the targeted problems change. Students get assigned one project and they work either alone or in groups of two. Students, in the interactive discussions with the whole class and the instructor, solve the project problems. Students regularly present their progress and defend their approach and result in front of the whole class. During one semester, several types of topics are dealt with (e.g., information integration; knowledge-, text- and data-mining of big biomedical data; sequence analysis related problems). Students get direct experience in research methodology, report writing, presentations, and, most importantly, different ways of approaching solving AI applications for different bioinformatics problems. **Students' prerequisites** Efficient programming in Python, preferable additional programming skills in C, C++, R, or Matlab. Preferable background: Computer Science/Engineering, Electrical Engineering, Mathematics. **Projects for the 2019 year** **Topic: Applications of word2vec-type methods for molecular function prediction** Data in many types of problems are described by a very large number of descriptors, so called features. This causes the problem in the downstream analysis of the data. There are many ways how to compress information that encodes the data. One of the latest approaches is based on the word2vec and similar types of methods. These methods are capable of compressing information coming from several tens of thousands of descriptors into several hundred new descriptors captured in the new feature vectors that describe original data items. This allows for significantly more efficient analysis of the original data that can be subjected further to various machine learning and AI processing. **Project 1. Apply above to analysis of the function of different transcripts.** **Project 2. Apply above to the analysis of different data describing cancers.** **Project 3. Apply above to the prediction of viruses in metagenomic sequences.** **Project 4. Natural Language Processing (NLP) for relationship/event extraction/modeling based on biomedical text** When analyzing text, it is much easier to assert that two concepts are associated (based on their co-occurrence frequencies), than to assert what type of relationship/event they are participating in. This is because the former is mainly based on named entity recognition (NER), however, the latter requires a deeper type of analysis. Even if relationship terms are identified within the text, it is much more challenging to assert which concepts they involve. e.g., consider the following sentence: “A lexer is generally combined with a parser, which together analyzes the syntax of text.” One can extract: [lexer] <- is generally combined with -> [parser], but how do these relate to [the syntax of text], the relationship is [analyze]. This needs the model to figure out that “which together” refers to [lexer] and [parser]. This is a very simple example, and sentences can get much more complicated than this. To complicate things further, two concepts can co-occur multiple times, in different contexts, describing different relationships. These can be used collectively to build some type of model for the association. The students will be advised to use NLP techniques to parse the sentences and based on the sentence structure explore different aspects of this challenging problem. **Project 5. Recognition of genomic signals** There are numerous types of signals in DNA and RNA sequences utilized by the biochemistry of living cells. None of these signals is simple to recognize. The current recognition models have different efficiencies and accuracies., but apparently none of the models appear to be sufficiently accurate. The problem of building efficient recognition models can be attached by different machine learning/deep learning approaches. In this project, students can use various types of deep learning networks, as well as the transformation of original sequences in order to get more accurate recognition models. **Project 6. Automated cleaning of topic-specific dictionaries** In many of the topic-specific dictionaries and ontologies, numerous concepts are described in addition to their official name/symbol, with alternative names/symbols, so-called synonyms. Frequently it happens that some of these synonyms also are used as synonyms for other concepts. during the automated analysis of biomedical text, it is necessary to minimize the effect of the potentially wrong association of such terms to concepts. The goal of this project is to develop an automated system that can with high accuracy pinpoint wrong associations. This can be achieved by using the word2vec type strategies, based on which the AI recognition model could be built.
Course Description from Program Guide	These projects cover application of AI to some of the relevant problems of analysis of large biological data and generally deal with complex information. Each year problems change. Students get assigned one (1) project and they work either alone or in groups of 2. Students in the interactive discussions with the whole class and the instructor solve the project problems. Students regularly present their progress and defend their approach and results in front of the whole class. During one (1) semester several types of topics are dealt with. Students get direct experience in research methodology, report writing, presentations and, most importantly, different ways of approaching solving AI problems

Goals and Objectives	The course consists of selected projects. The projects may change each year. These projects cover the application of artificial intelligence (AI) to some of the relevant problems of analysis of large biological data and generally deal with complex information. Each year, the targeted problems change. Students get assigned one project and they work either alone or in groups of two. Students, in the interactive discussions with the whole class and the instructor, solve the project problems. Students regularly present their progress and defend their approach and result in front of the whole class. During one semester, several types of topics are dealt with (e.g., information integration; knowledge-, text- and data-mining of big biomedical data; sequence analysis related problems). Students get direct experience in research methodology, report writing, presentations, and, most importantly, different ways of approaching solving AI applications for different bioinformatics problems.
Required Knowledge	Efficient programming in Python, preferable additional programming skills in C/C++, R, or Matlab. Preferable background: Computer Science/Engineering, Electrical Engineering, Mathematics.
Reference Texts	1. **Entity linking (entity normalization/disambiguation)** https://en.wikipedia.org/wiki/Entity_linking 2. **Information extraction** https://en.wikipedia.org/wiki/Information_extraction 3. **Controlled Vocabulary/Dictionary https://en.wikipedia.org/wiki/Controlled_vocabulary** 4. **Named Entity Recognition https://en.wikipedia.org/wiki/Named-entity_recognition** 5. **Relation Extraction http://stanford.edu/class/cs124/lec/rel.pdf** 6. **NCBI Text Mining Tools (including tmVar)** **https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools/** 7. **Approximate matching** **https://courses.cs.washington.edu/courses/cse427/16au/slides/approximate_matching.pdf** 8. **Word2Vec http://www-personal.umich.edu/~ronxin/pdf/w2vexp.pdf** 9. **HMMs (Hidden Markov models) https://en.wikipedia.org/wiki/Hidden_Markov_model** 10. **Chemical named entities recognition: a review on approaches and applications** **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4022577/**
Method of evaluation	**10.00**% - Written report **15.00**% - Presentation **50.00**% - Course Project(s) **10.00**% - Active participation **15.00**% - Attendance and Participation
Nature of the assignments	Research project as defined in the course description complemented by the presentations of results, discussions on the methods of solution, written mid-term and final report.
Course Policies	Student absence of more than three times without justifiable reason will lead to failing the course.
Additional Information	Assessment of students is continuous.

Tentative Course Schedule *(Time, topic/emphasis & resources)*		
Week	**Lectures**	**Topic**
1	Sun 01/27/2019	Introduction
2	Sun 02/03/2019	Students get assigned to the projects. Projects explanations.
3	Sun 02/10/2019	Presentations of progress on individual reports and discussions with the whole class
4	Sun 02/17/2019	Presentations of progress on individual reports and discussions with the whole class
5	Sun 02/24/2019	Presentations of progress on individual reports and discussions with the whole class
6	Sun 03/03/2019	Presentations of progress on individual reports and discussions with the whole class
7	Sun 03/10/2019	Presentations of progress on individual reports and discussions with the whole class
8	Sun 03/17/2019	Presentations of progress on individual reports and discussions with the whole class
9	Sun 03/24/2019	Mid-term report
10	Sun 03/31/2019	Presentations of progress on individual reports and discussions with the whole class
11	Sun 04/07/2019	Presentations of progress on individual reports and discussions with the whole class
12	Sun 04/14/2019	Presentations of progress on individual reports and discussions with the whole class
13	Sun 04/21/2019	Presentations of progress on individual reports and discussions with the whole class
14	Sun 04/28/2019	Presentations of progress on individual reports and discussions with the whole class
15	Sun 05/05/2019	Presentations of progress on individual reports and discussions with the whole class
16	Sun 05/12/2019	Presentations of progress on individual reports and discussions with the whole class
17	Sun 05/19/2019	Final report

Note

The instructor reserves the right to make changes to this syllabus as necessary.