New Course Proposal

Date Submitted: 04/17/25 9:57 pm

Viewing: CSI 671 : Natural Language Processing for Complex Systems in Science

Last edit: 04/17/25 9:57 pm

Changes proposed by: blaisten

Are you completing this form on someone else's behalf?

Yes

Requestor:

Name	Extension	Email
Estela Blaisten	31988	blaisten@gmu.edu

In Workflow

1. **CDS Chair**
2. SC Curriculum Committee
3. SC Assistant Dean
4. Assoc Provost-Graduate
5. Registrar-Courses
6. Banner

Approval Path

1. 04/17/24 3:16 pm
   Jason Kinser (jkinser): Approved for CDS Chair
2. 04/29/24 11:26 am
   Jennifer Bazaz Gettys (jbazaz): Rollback to Initiator

Effective Term: Spring 2026

Subject Code: CSI - Computational Science & Informatics

Course Number: 671

Bundled Courses:

Is this course replacing another course? No

Equivalent Courses: CSS 671 - Natural Language Processing for Complex Systems

Catalog Title: Natural Language Processing for Complex Systems in Science

Banner Title: Natural Language Processing

Will section titles vary by semester? No

Credits: 3

Schedule Type: Lecture

Hours of Lecture or Seminar per week: 2.5

Repeatable: May only be taken once for credit, limited to 2 attempts (N2)

Max Allowable Credits: 6

Default Grade Mode: Graduate Regular

Recommended Prerequisite(s):

CSI 500 and CSI 690 or equivalent or permission of the instructor

Recommended Corequisite(s):

Required Prerequisite(s) / Corequisite(s) (Updates only):

Registrar's Office Use Only - Required Prerequisite(s)/Corequisite(s):

And/Or	(	Course/Test Code	Min Grade/Score	Academic Level	)	Concurrency?

Registration Restrictions (Updates only):

NA

Registrar's Office Use Only - Registration Restrictions:

Field(s) of Study:

Class(es):

Level(s):

Degree(s):

School(s):

Catalog Description:

This course focuses on the fundamentals of Natural Language Processing (NLP), Natural Language Understanding (NLU) and Large Language Models (LLMs), comparatively or coupled with other computational sciences methods used in modeling, such as agent-based modeling, audio and image processing, network and biological sequence analyses, property prediction of chemical compounds, and knowledge representation in science. It teaches topic modeling, topic2vec, speech recognition,

grouping and categorization strategies, and more as applied to complex systems inn science.

Justification:

What: This course adds a component to the Computational Sciences and Informatics Ph.D instruction by teaching doctoral students specific applications of NLP/NLU and large language models (LLMs) in complex systems applications in science. Particularly, the CSI, PhD students are researching various advanced computational sciences approaches in modeling and simulation without having formal training or access to studing NLP/NLU/LLMs methods and their cross-methodological use with other data-based methods.

Why: The Computational Sciences and Information doctoral students need to become proficient in a range of computational methodologies in order to successfully graduate and competently represent CMU in the workplace. Currently, the graduate curriculum lacks NLP/NLU/LLM courses that focus on complex systems and on the use of these methods aligned with other computational science methods, e.g. machine learning and deep neural learning. Our department has assessed that the line of expertise that CSI 671 students will achieve enhances the quality of their CSI portfolios.

Does this course cover material which crosses into another department?	No

Learning Outcomes:

1. Students will acquire competency with state-of-the art NLP/NLU/LLM methods and applications in current complex systems phenomena in science, such as information and opinion dynamics, speech recognition problems, system environment analysis, science news in different foreign languages.
2. Students will reach proficiency in the use of NLP/NLU/LLM methods comparatively or coupled with other methodologies, such as agent-based models, network analysis, image processing, audio processing, from humans, AI bots, and cetacean (dolphins and whales) recordings data.
3. Students will be competent in the analysis of NLP/NLU/LLM applications of complex systems in various areas of science as well as science-based theories with respect to language convergence, origins of language theories, system complexity, communication patterns and scalability in living systems, and communication for space missions.

Will this course be scheduled as a cross-level cross listed section? No

Attach Syllabus

CSI671_Syllabus_04-17-2025.pdf

Additional Attachments

Staffing:

Dr. Anamaria Berea, Associate Professor, Department of Computational and Data Sciences, College of Science

Relationship to Existing Programs:

The CSI, PhD and the CSS, PhD are two doctoral progrrams administered by the Department of Computational Sciences. It is a strong interest of the department to have the new CSI 671 course "equivalent to CSS 671," which will eliminate unnecessary course substitutions currently unavoidable for the doctoral students to maintain the number of CSI credits required by the CSI, PhD program if they register in other than CSI courses.

Relationship to Existing Courses:

Provides intermediate and advanced coding and logical elements that can be further used in other courses within the CSI, PhD that emphasize on data and statistical analyses such as CSI 674, CSI 695, CSI 703, CSI 772, CSI 777.

This new course parallels, but does not overlap, the content and methods taught in the following courses: CS 678; AIT 626 and AIT 726.

Have you reached out to the Libraries to determine whether there are adequate resources to support your course? If not, please email Meg Meiman, Associate University Librarian for Learning, Research, and Engagement at mmeiman2@gmu.edu.

Yes

Additional Comments:

This course will be added to the Elective Courses category of the CSI, PhD program requisites.

Reviewer Comments

Jennifer Bazaz Gettys (jbazaz) (04/29/24 11:26 am): Rollback: Rolling back for further discussion.

Key: 18563

CSI 671

NATURAL LANGUAGE PROCESSING FOR COMPLEX SYSTEMS IN SCIENCE

1. General Information

Instructor:	Dr. Anamaria Berea, aberea@gmu.edu
Location:	Department of Computational and Data Sciences, College of Science
Course website:	Blackboard/Canvas website
Code repository:	https://gitlab.orc.com/aberea/css_csi_nlp
Credits:	3
Recommended prereq:	**CSI 500, CSI 690 or equivalent**
Office Hours:	By appointment

2. Course Description

This course focuses on the fundamentals of Natural Language Processing (NLP), Natural Language Understanding (NLU) and Large Language Models (LLMs), comparatively or coupled with other computational sciences methods used in modeling such as agent-based modeling, audio and image processing, network and biological sequence analyses, property prediction of chemical compounds, and knowledge representation in science. It teaches topic modeling, topic2vec, speech recognition, grouping and categorization strategies, and more as applied to complex systems in science.

3. Learning Outcomes

1. Students will acquire competency with state-of-the art NLP/NLU/LLM methods and applications in current complex systems phenomena in science, such as information and opinion dynamics, speech recognition problems, system environment analysis, science news in different foreign languages.
2. Students will reach proficiency in the use of NLP/NLU/LLM methods comparatively or coupled with other methods, such as agent-based models, network analysis, image processing, audio processing, from human, AI bots and cetacean (dolphins and whales) recordings data.
3. Students will be competent in the analysis of NLP/NLU/LLM applications of complex systems in various areas of science as well as science-based theories with respect to language convergence, origins of language theories, system's complexity, communication patterns, scalability in living systems, and communication for space missions.

4. Textbooks

Required textbook:

1. Computational Analysis of Communication: Van Atteveldt, Wouter, Damian Trilling, and Carlos Arcíla Calderón. Computational analysis of communication. John Wiley & Sons, 2022. https://cssbook.net

Recommended books:

1. Bird, Steven, Ewan Klein, and Edward Loper. Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc., 2009. https://tjzhifei.github.io/resources/NLTK.pdf
2. Adger, David. Language unlimited: The science behind our most creative power. Oxford University Press, USA, 2019.
3. Jockers, Matthew L., and Rosamond Thalken. Text analysis with R. Springer International Publishing, 2020.

Additional useful online resources:
https://www.complexityexplorer.org/courses/135-foundations-applications-of-humanities-analytics.

5. Technology Requirements

Hardware: Students need access to a Windows, Mac, or Linux computer with at least 8 GB of RAM and a fast, reliable broadband Internet connection (e.g., cable, DSL). Most programming exercises can be performed on a regular, average, laptop. For the programming exercises that are computationally intensive, students will be instructed how to access open source cloud systems (i.e., Google Colab) and/or the 20 workstations available in our department computing lab.

Software: This course will be using primarily Python, and/or secondarily R as a programming language. Prior intermediate level knowledge of these languages is required.

6. Course outline (Tentative)

Week	Topic	What is Due
1	Natural Language Processing and its relationship with information theory	
2	Fundamentals of text mining and topic modeling	SP
3	Categorizing and tagging words; text classification and n-gram structures	SP
4	Extracting information from text, audio, and image files	SP
5	Syntax and Context-Free Grammars with applications to non-human language data	SP
6	Feature-based grammars and neural network models of language inference	SP
7	Semantic analysis, SemanticML and Natural Language Understanding (NLU) in foreign news and scientific media	SP
8	Applications of NLP/NLU to underlying complex informational structures	SP
9	Speech recognition, communication patterns and scalability in living systems	SP
10	Fundamentals of Large Language Models (Bert, PaLM and GPT systems)	SP
11	Pattern recognition in complex datasets; language design for space missions	SP
12	Universal patterns of communication in complex systems within science	SP
13	Summary of the semester	SP
14	Final project presentation	FP

7. Grades

Final Project (FP)	**60%**
Student presentations (SP)	**40%**

Final Mark	Corresponding Grade
> 99	A+
95.1 – 99	A
90.1 – 95	A-
87.1 – 90	B+
83.1 – 87	B
80.1 – 83	B-
70.1 – 80	C
<=70	F

8. Student Presentations and Final Project

Each student will present once during the semester, at the end of the class, a published article, a model, or an algorithm of their choice, but related to the course topic of that week.
Final project consists of an original computational project in NLP/NLU/LLM of student's choice, that students will be encourage of working on throughout the semester.

9. **General University Policies:** https://stearnscenter.gmu.edu/knowledge-center/designing-your-syllabus/