## CSC733 Fall 2024 Natural Language Processing Syllabus revision 12/11/2024

## CSC733 is online and face-to-face course meeting in 1N111 on Wednesdays at 6:30pm. You are required to attend all classes. If possible, the classes in 1N111 will be zoomed online for those students who have difficulty getting to Staten Island. Expect to spend at least 6 hours per week on the assigned videos, readings and the term project.

The CSC733 class web site is [**CSC733 Fall 2024 Web site**](http://163.238.35.144/~chi/CSC733/CSC733%20Fall%202024.html)

## Course Prerequisites:

## Good programming skills, knowledge of basic algorithms and data structures (including sorting & trees). You should have taken a course on data structures and algorithms at the graduate or undergraduate level, or have permission from the instructor.

## Course Description:

Natural language processing (NLP) or computational linguistics is one of the most important technologies of the information age. NLP is needed because people communicate almost everything in language. Starting in the 1960s Zelig Harris started the study of English grammar as a mathematical structure. His student Naomi Sager continued his work with considerable success in developing queryable natural language databases. With increased computing power developed in the 2000s, brute force techniques, as used in chess, became more successful than grammatical analysis. The invention of the Transformer by Google in 2017 has led to the current state of the art.

In the 2010s, deep learning (or neural network) approaches obtained very high performance across many different NLP tasks. In the 2020s amazing further progress was made through the scaling of Large Language Models, such as ChatGPT. In this course, students will gain a thorough introduction to both the basics of Deep Learning for NLP and the latest cutting-edge research on Transformers and Large Language Models (LLMs). Through lectures, assignments and a final project, students will learn the necessary skills to design, implement, and understand their own neural network models. We will also cover generative AI which can create new content text, images, video. We will use it to summarize and query natural language texts.

This syllabus has been completely revised since CSC733 was last taught at CSI in 2022 to reflect the changes to NLP since the debut of [ChatGpt](https://chatgpt.com/).

## Course Prerequisites:

## Good programming skills, knowledge of basic algorithms and data structures (including sorting & trees). You should have taken a course on data structures and algorithms at the graduate or undergraduate level, or have permission from the instructor.

## Course Description:

* History of Natural Language Processing
* Programming in Python review if necessary
* NLTK the natural language toolkit
* Extracting information from text
* Parsing; syntax & semantics
* Linguistic Data – Word Vectors
* Neural Networks
* Deep Learning
* Transformers
* Large Language Models

## Course Objectives:

## Acquire the skill to program neural networks & large language models in Python

## Acquire the skill to automatically extract information from linguistic data, i.e. texts

## Learn how to manage linguistic data and used it in developing linguistic databases

## 

**Course Tools**

**Required Textbook (free):**

* **Natural Language Processing with Python** by Bird, Klein & Loper
  + 1st edition (2009) for Python 2 available [At this link](https://tjzhifei.github.io/resources/NLTK.pdf)
    - Free & easy to read
  + 1st edition available from Amazon for $52.51 new or rent $30.71
    - Easy to read but expensive
  + Updated edition for Python 3 available from [FREE Textbook](https://www.nltk.org/book/)
    - A little hard to read but FREE and up-to-date
* **Multiple web sites** detailed below

**Software:**

* Python 3.  This open source software is free at [Python.org](http://www.python.org)  Setup of Python is not difficult and you can easily do it on your own. Instructions for the Windows setup are in [Python](http://163.238.35.144/~chi/CSC733/Python%20Installation.docx).
* Natural Language Toolkit [NLTK](https://www.nltk.org/) and [NLTK startup](http://163.238.35.144/~chi/CSC733/NLTK%20start%20script.txt)

**Course Requirements:**

**Please note that in CSC733 *ALL* work is to be done by yourself or the source cited. Anything which you have not written yourself and not cited will receive an immediate 10 point penalty in your grade. No group work.**

* No problem sets
* No exams
* No final exam
* Project 100% of your grade. The final project is due on December 18 for presentation in class. Project deliverables are due on the dates given below. A penalty of 10 points per week if not delivered on time. The professor needs to approve your text(s) for the project.

**Schedule**

August 28, 2024  **1 Introduction to NLP; History of NLP** [**Stanford course**](https://web.stanford.edu/class/cs224n/index.html#schedule)

September 4, 2024 **2 The Linguistic String Project**

**The Natural Language Toolkit 1** [**NLTK**](https://www.nltk.org/)

Read the textbook Chapters 1 & 2

September 11, 2024 **3 The Natural Language Toolkit 2**

Read the textbook Chapters 3 & 4

[**Project deliverable 1**](http://163.238.35.144/~chi/CSC733/Fall%202024/Project/2024%20Deliverables/project1a.txt)due September 25

September 18, 2024 **4 The Natural Language Toolkit 3**

Read the textbook Chapters 5 & 6

Project deliverable #1 due free 1 week extension

[**Project deliverable 2**](http://163.238.35.144/~chi/CSC733/Fall%202024/Project/2024%20Deliverables/project%202.txt)

[**Chi's deliverable 2**](http://163.238.35.144/~chi/CSC733/Fall%202024/Project/2024%20Deliverables/Chi%20project%202.txt)due October 9

September 25, 2024 **5 The Natural Language Toolkit 4**

Read the textbook Chapters 7 & 8

[**Project deliverable 3**](http://163.238.35.144/~chi/CSC733/Fall%202024/Project/2024%20Deliverables/deliverable3.docx)due October 23

[**Chi's deliverable 3**](http://163.238.35.144/~chi/CSC733/Fall%202024/Project/2024%20Deliverables/Chi%20deliverable3.docx)

October 2, 2024 **CSI holiday**

October 9, 2024 **6 Word Vectors** [**Stanford lecture**](https://www.youtube.com/watch?v=rmVRLeJRkl4&list=PLoROMvodv4rMFqRtEuo6SGjY4XbRIVRd4&index=3)if lecture doesn’t open click **Browse**

[**course in machine learning**](http://ciml.info/)

October 16 2024 **7 Word Vectors** [**Stanford lecture**](https://www.youtube.com/watch?v=rmVRLeJRkl4&list=PLoROMvodv4rMFqRtEuo6SGjY4XbRIVRd4&index=3)continued

[**Project deliverable 4**](http://163.238.35.144/~chi/CSC733/Fall%202024/Project/2024%20Deliverables/deliverable4.docx)due October 30

[**Project deliverable 4 script**](http://163.238.35.144/~chi/CSC733/Fall%202024/Project/2024%20Deliverables/Project%20deliverable%204%20Gensim.txt)

October 23, 2024 **8 Word Vectors** [**Stanford lecture #2**](https://www.youtube.com/watch?v=gqaHkPEZAew&list=PLoROMvodv4rMFqRtEuo6SGjY4XbRIVRd4&index=2)continued

October 30, 2024 **9 Machine Learning & Neural Networks**

[**1st neural network for beginners**](https://towardsdatascience.com/first-neural-network-for-beginners-explained-with-code-4cfd37e06eaf)

[**Stanford lecture #3**](https://www.youtube.com/watch?v=X0Jw4kgaFlg&list=PLoROMvodv4rMFqRtEuo6SGjY4XbRIVRd4&index=3)

[**Speech & Language Processing - Jurafsky Stanford**](https://web.stanford.edu/~jurafsky/slp3/)

November 6, 2024 **10 Neural Networks**

[**Build a Neural Network & Make Predictions – Real Python**](https://realpython.com/python-ai-neural-network/#artificial-intelligence-overview)

[**Create a simple neural network in python**](https://www.kdnuggets.com/2018/10/simple-neural-network-python.html)

[**RealPython neural network - Mesquita**](https://realpython.com/python-ai-neural-network/)

[**project deliverable 5**](Project/2024%20Deliverables/deliverable5.docx) due December 4

November 13, 2024 **11 Neural Networks**

[**Deep neural networks**](https://www.datacamp.com/tutorial/introduction-to-deep-neural-networks)

[**deep learning tutorial**](https://www.simplilearn.com/tutorials/deep-learning-tutorial)

November 20, 2024 **12 Transformers**

[**Huggingface**](https://huggingface.co/learn/nlp-course/chapter1/1)

[**Introduction to Transformers & Huggingface**](https://www.datacamp.com/tutorial/an-introduction-to-using-transformers-and-hugging-face)

[**Huggingface Transformers**](https://realpython.com/huggingface-transformers/)

[**Hugging Face video**](https://www.youtube.com/watch?v=00GKzGyWFEs)

[**project deliverable 6**](Project/2024%20Deliverables/deliverable6.docx)due December 11, no penalty for Dec 18

December 4, 2024 **13 Transformers & Large Language Models**

[**Transformers Explained Visually**](https://towardsdatascience.com/transformers-explained-visually-part-1-overview-of-functionality-95a6dd460452) 3 parts

[**How Transformers Work**](https://www.datacamp.com/tutorial/how-transformers-work)

[**Building a Transformer with PyTorch**](https://www.datacamp.com/tutorial/building-a-transformer-with-py-torch)

December 11, 2024 **14 Transformers & Large Language Models** continued

[**How Transformers Work**](https://www.datacamp.com/tutorial/how-transformers-work)

December 18, 2024 **15 Project presentations**

**Required in all CSI syllabi**

Suggested language for inclusion in department syllabi regarding mandatory reporting and pregnancy:

Support for Pregnant Students:

If you are a pregnant student in potential need of an accommodation and/or related resources, you are encouraged to contact the Director of CSI’s Office of Diversity, Equity & Inclusion as early as possible in the semester, who can coordinate reasonable modifications when necessary to ensure you maintain equal access to the College’s education programs and activities. Additionally, Title IX and CUNY policy prohibit the harassment of students based on sex, which includes pregnancy, childbirth and related conditions. If you believe you have been harassed on this basis, such conduct should be reported to Office of Diversity, Equity & Inclusion, as well. Contact information for the Director of CSI’s Office of Diversity, Equity & Inclusion is as follows:

Tara Mastrangelo, Director Office of Diversity, Equity & Inclusion

CDO/Title IX Coordinator/ADA s. 504 Coordinator Tara.Mastrangelo@csi.cuny.edu

(718) 982-2688

Building 1A, Room 205

Mandatory Reporting:

Please be on notice that I am a mandatory reporter pursuant to CUNY policy. As a result, I must notify the College’s Title IX Coordinator if I reasonably suspect, am informed of, or observe conduct that may constitute sex-based misconduct.

**For later** Read NLP:

* + Chapter 1: Language Processing and Python
* Do Exercises 1.8 #4

**Module 2: Accessing Text corpora & Lexical Resources**

* Read NLP:
  + Chapter 2: Accessing Text Corpora and Lexical Resources
    - Skip sections 4.2, 4.3, 4.4
* Do
  + Exercises 1.8 #6, #10, #13, #15, #17, #23

**Module 3: Processing Raw text**

* Read NLP:
  + Chapter 3: Processing Raw Text
* Do: Choose a book from [Project Gutenberg](https://www.gutenberg.org/)
  + Upload it to your NLTK
* Do
  + Exercises 2.8 #1, #4, #9, #17, #
* Read:
  + Chapter 3: Processing Raw Text

* Do: Go over the NLTK Scripts and run everything in them
  + We will discuss this in class on September 14

**Module 5: Tagging words**

* Read Chapter 5: Categorizing and Tagging Words
* Do: Load your own corpus for your term project

**Module 6: Classifying text**

* Read Chapter 5: Categorizing and Tagging Words
* Do: Find some interesting words in your project book(s) and run concordances on them

**Module 8: Extracting information from text 1**

**Module 9: Analyzing Sentence Structure 1**

**Module 11: Building feature-based grammars**

**Module 12: Analyzing the Meaning of Sentences 1**

**Module 14: Managing Linguistic Data 1**

## 