Computational Linguistics

Spring 2022
"Honestly? I preferred when we didn’t talk about the elephant."

Shannon Wheeler, 2011
Computers and language
This is a class about computational linguistics.
This is a class about computational linguistics. …or is it?
Computational linguistics is the study of language, using computers.

Natural language processing (NLP) is the study of how to make computers use language.

In this course, we’ll do a bit of each, but the emphasis will really be on NLP.
(natural language) processing,
not

natural (language processing)
NLP can do amazing things.
Alexander joins forces with James Madison
And John Jay to write a series of essays
Defending the new United States Constitution
Entitled The Federalist Papers
The plan was to write a total of 25 essays
The work divided evenly among the three men
In the end, they wrote 85 essays
In the span of six months
John Jay got sick after writing five
James Madison wrote 29
Hamilton wrote the other 51

Lin-Manuel Miranda, “Non-Stop”
The Federalist; A Collection of Essays, Written in Favour of the New Constitution, as Agreed Upon by the Federal Convention, September 17, 1787. In Two Volumes.

James Madison

Alexander Hamilton
February 2011
Word Lens
NLP also messes up a lot.
Hey @Google, this is...real weird.

The first colony was founded at Jamestown, Virginia, in 1607. Many of the people who settled in the New World came to escape religious persecution. The Pilgrims, founders of Plymouth, Massachusetts, arrived in 1620. In both Virginia and Massachusetts, the colonists flourished with some assistance from Native Americans.

www.americaslibrary.gov > colonial > jb_colonial_subj
Colony of Virginia (1492-1763)

33,000 years ago
The 'Clovis first theory' refers to the 1950s hypothesis that the Clovis culture represents the earliest human presence in the Americas, beginning about 13,000 years ago; evidence of pre-Clovis cultures has accumulated since 2000, pushing back the possible date of the first peopling of the Americas to 33,000 years ago.

en.wikipedia.org > wiki > Settlement_of_the_Americas
Settlement of the Americas - Wikipedia
Female doctors don't exist, says Google Translate

Correct translations for 20 translation pairs to and from French, German, Spanish, Italian and Polish.

<table>
<thead>
<tr>
<th>Correct translations (female form)</th>
<th>Correct translations (male form)</th>
</tr>
</thead>
<tbody>
<tr>
<td>historian</td>
<td>0</td>
</tr>
<tr>
<td>doctor</td>
<td>0</td>
</tr>
<tr>
<td>soldier</td>
<td>1</td>
</tr>
<tr>
<td>president</td>
<td>1</td>
</tr>
<tr>
<td>student</td>
<td>3</td>
</tr>
<tr>
<td>pilot</td>
<td>3</td>
</tr>
<tr>
<td>boss</td>
<td>5</td>
</tr>
<tr>
<td>driver</td>
<td>9</td>
</tr>
<tr>
<td>teacher</td>
<td>10</td>
</tr>
<tr>
<td>shop assistant</td>
<td>13</td>
</tr>
<tr>
<td>nurse</td>
<td>20</td>
</tr>
</tbody>
</table>

How to read the chart: Out of 20 translations of a female doctor, none were correct (e.g. "die Doktorin" becomes "le docteur", "la dottoressa" becomes "der Doktor" etc.)

Source: AlgorithmWatch • Get the data • Created with Datawrapper
Kuczmarski, 2018

Before

Google Translate

TURKISH

o bir doktor

ENGLISH

he is a doctor

After

Google Translate

TURKISH

o bir doktor

ENGLISH

Translations are gender-specific. LEARN MORE

she is a doctor (feminine)

he is a doctor (masculine)
but

Kayser-Bril, 2020
Good NLP work relies on:

- Acknowledgment of limitations of methods
- Well-defined strategies of evaluation
- Engagement with the text itself
What will we do in this course?
Tuesdays: Discussion

Thursdays: Hands-on
CLASS GOALS

Implement classic algorithms
Use appropriate metrics of evaluation
Read primary literature
Critique assumptions
Practice scholarly skills
CLASS GOALS

Implement classic algorithms

Use appropriate metrics of evaluation

*Read primary literature*

Critique assumptions

*Practice scholarly skills*

CLASS WORK

Worksheets
CLASS GOALS

Implement classic algorithms
Use appropriate metrics of evaluation

Read primary literature

Critique assumptions

Practice scholarly skills

CLASS WORK

Worksheets

Discussions & presentations
CLASS GOALS

Implement classic algorithms

Use appropriate metrics of evaluation

Read primary literature

Critique assumptions

Practice scholarly skills

CLASS WORK

Worksheets

Discussions & presentations

Assignments
CLASS GOALS

Implement classic algorithms

Use appropriate metrics of evaluation

Read primary literature

Critique assumptions

Practice scholarly skills

CLASS WORK

Worksheets

Discussions & presentations

Assignments

Final project
CLASS GOALS

Implement classic algorithms
Use appropriate metrics of evaluation
Read primary literature
Critique assumptions
Practice scholarly skills

CLASS WORK

Worksheets
Discussions & presentations
Assignments
Final project
Course information
CMPU 366

Computational Linguistics
Spring 2022

Tuesday 12:00-1:15 p.m.
Thursday 12:00-1:15 p.m.
New England 206

Prof. Jonathan Gordon

This schedule is tentative. Links will be updated here for readings and labs as we progress through the semester. Readings should be posted the Friday before the relevant Tuesday class, while assignments will be posted Thursday just before class.
CMPU 366
Computational Linguistics
Spring 2022

Tuesday & Thursday, 12:00–1:15 p.m.
New England 206
Prof. Jonathan Gordon
cs.vassar.edu/~cs366

Overview

Computational linguistics is broadly the study of how to enable computers to analyze language produced by people or even generate such language themselves. While the idea of machines that use language is quite old, today this technology is all around us, in our phone assistants, search engines, social media recommendations, YouTube captions, and more. In this class, we focus on core concepts and problems in computational linguistics, centered on how text can be represented, predicted, categorized, dismantled, and interpreted. We take a hands-on approach to exploring parts of this
Prerequisites

Two 200-level CS courses

Recommended:

cmpu 203: Software Design and Implementation

cmpu 240: Theory of Computation
Speech and Language Processing

An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition

Third Edition draft

Daniel Jurafsky
Stanford University

James H. Martin
University of Colorado at Boulder
<table>
<thead>
<tr>
<th>ACTIVE ASSIGNMENTS</th>
<th>RELEASED</th>
<th>DUE (EST)</th>
<th>SUBMISSIONS</th>
<th>% GRADED</th>
<th>PUBLISHED</th>
<th>REGRADES</th>
</tr>
</thead>
</table>

You currently have no assignments.

Create an assignment to get started.

Create Assignment
Vassar CS Student Integrity Guide

This guide is designed to clarify Vassar College’s academic integrity policy as it applies to the Computer Science Department. Furthermore, it provides advice on how to best navigate integrity issues in the context of the field, where source code authorship is a central issue.

The goal of our computer science courses is to promote understanding of the field, not competition among students. As such, students are encouraged to discuss class material, ideas, sample exercises, etc., with other students.

However, when it comes to graded work (e.g., programming assignments, programming labs, take-home exams), it is important to know when to collaborate and when to work individually. Taking shortcuts, while seemingly beneficial in the short term, will inevitably backfire later on. Conversely, the challenges of working through a problem will pay off greatly in future courses and postgraduate life, as they will enable students to be more independent in their work.

1. Policy

1.1. Guidelines for individual work

individual work is to assess the learning of each person in isolation. The guidelines are the following:
Grading

- 40% for Final project
- 40% for Assignments
- 10% for Special topic presentation
- 10% for Worksheets
Pair / group work
You created this private channel today. This is the very beginning of the cmpu366-general channel.

Add description  Add people  Send emails to channel

Jonathan 23:38
joined cmpu366-general.

Message @cmpu366-general
“All through our education, we are being taught a kind of reverse mindfulness. A kind of Future Studies where – via the guise of mathematics, or literature, or history, or computer programming, or French – we are being taught to think of a time different to the time we are in. Exam time. Job time. When-we-are-grown-up time.

To see the act of learning as something not for its own sake but because of what it will get you reduces the wonder of humanity. We are thinking, feeling, art-making, knowledge-hungry, marvelous animals, who understand ourselves and our world through the act of learning. It is an end in itself. It has far more to offer than the things it lets us write on application forms. It is a way to love living right now.”

Matt Haig, Notes on a Nervous Planet
Getting help

Extension policy
Question 1: What are words?
Question 1: What are words?

Question 0: How do computers encode text?
An *encoding* is a mapping from $A \rightarrow B$.

In computers: bits $\rightarrow$ data

For text in computers: bits $\rightarrow$ characters in a string
ASCII
telnet towel.blinkenlights.nl
ASCII

telnet towel.blinkenlights.nl
**ASCII**  *American Standard Code for Information Interchange*

<table>
<thead>
<tr>
<th>Bits</th>
<th>Code point</th>
<th>Character</th>
</tr>
</thead>
<tbody>
<tr>
<td>01000001</td>
<td>65</td>
<td>&quot;A&quot;</td>
</tr>
<tr>
<td>01010011</td>
<td>83</td>
<td>&quot;S&quot;</td>
</tr>
<tr>
<td>01000011</td>
<td>67</td>
<td>&quot;C&quot;</td>
</tr>
<tr>
<td>01001001</td>
<td>73</td>
<td>&quot;I&quot;</td>
</tr>
<tr>
<td>01001001</td>
<td>73</td>
<td>&quot;I&quot;</td>
</tr>
<tr>
<td>Dec</td>
<td>Char</td>
<td>Dec</td>
</tr>
<tr>
<td>-----</td>
<td>---------------</td>
<td>-----</td>
</tr>
<tr>
<td>0</td>
<td>NUL (null)</td>
<td>32</td>
</tr>
<tr>
<td>1</td>
<td>SOH (start of heading)</td>
<td>33</td>
</tr>
<tr>
<td>2</td>
<td>STX (start of text)</td>
<td>34</td>
</tr>
<tr>
<td>3</td>
<td>ETX (end of text)</td>
<td>35</td>
</tr>
<tr>
<td>4</td>
<td>EOT (end of transmission)</td>
<td>36</td>
</tr>
<tr>
<td>5</td>
<td>ENQ (enquiry)</td>
<td>37</td>
</tr>
<tr>
<td>6</td>
<td>ACK (acknowledge)</td>
<td>38</td>
</tr>
<tr>
<td>7</td>
<td>BEL (bell)</td>
<td>39</td>
</tr>
<tr>
<td>8</td>
<td>BS (backspace)</td>
<td>40</td>
</tr>
<tr>
<td>9</td>
<td>TAB (horizontal tab)</td>
<td>41</td>
</tr>
<tr>
<td>10</td>
<td>LF (NL line feed, new line)</td>
<td>42</td>
</tr>
<tr>
<td>11</td>
<td>VT (vertical tab)</td>
<td>43</td>
</tr>
<tr>
<td>12</td>
<td>FF (NP form feed, new page)</td>
<td>44</td>
</tr>
<tr>
<td>13</td>
<td>CR (carriage return)</td>
<td>45</td>
</tr>
<tr>
<td>14</td>
<td>SO (shift out)</td>
<td>46</td>
</tr>
<tr>
<td>15</td>
<td>SI (shift in)</td>
<td>47</td>
</tr>
<tr>
<td>16</td>
<td>DLE (data link escape)</td>
<td>48</td>
</tr>
<tr>
<td>17</td>
<td>DC1 (device control 1)</td>
<td>49</td>
</tr>
<tr>
<td>18</td>
<td>DC2 (device control 2)</td>
<td>50</td>
</tr>
<tr>
<td>19</td>
<td>DC3 (device control 3)</td>
<td>51</td>
</tr>
<tr>
<td>20</td>
<td>DC4 (device control 4)</td>
<td>52</td>
</tr>
<tr>
<td>21</td>
<td>NAK (negative acknowledge)</td>
<td>53</td>
</tr>
<tr>
<td>22</td>
<td>SYN (synchronous idle)</td>
<td>54</td>
</tr>
<tr>
<td>23</td>
<td>ETB (end of trans. block)</td>
<td>55</td>
</tr>
<tr>
<td>24</td>
<td>CAN (cancel)</td>
<td>56</td>
</tr>
<tr>
<td>25</td>
<td>EM (end of medium)</td>
<td>57</td>
</tr>
<tr>
<td>26</td>
<td>SUB (substitute)</td>
<td>58</td>
</tr>
<tr>
<td>27</td>
<td>ESC (escape)</td>
<td>59</td>
</tr>
<tr>
<td>28</td>
<td>FS (file separator)</td>
<td>60</td>
</tr>
<tr>
<td>29</td>
<td>GS (group separator)</td>
<td>61</td>
</tr>
<tr>
<td>30</td>
<td>RS (record separator)</td>
<td>62</td>
</tr>
<tr>
<td>31</td>
<td>US (unit separator)</td>
<td>63</td>
</tr>
</tbody>
</table>

These are all of the characters in ASCII.
1963 ASCII, a 7-bit standard


Examples: Latin-1, Latin/Cyrillic, Latin/Greek, Latin/Hebrew,

But CJKV languages need bigger character sets, e.g., Big5 (two bytes), GB 18030 and ISO 2022 (more complex).

How do you tell which one a file is using?
Today we mostly use *Unicode*. Unicode isn’t an encoding; it’s a listing of code points.

It can be encoded in different ways:

- UTF-32 is a 32-bit fixed-length encoding
- UTF-8 and UTF-16 are variable-length encodings
Announcing The Unicode® Standard, Version 14.0

Version 14.0 of the Unicode Standard is now available, including the core specification, annexes, and data files. This version adds 838 characters, for a total of 144,697 characters. These additions include five new scripts, for a total of 159 scripts, as well as 37 new emoji characters.

The new scripts and characters in Version 14.0 add support for modern language groups in Bosnia, India, Indonesia, Iran, Java, Malaysia, Mongolia, Myanmar, Pakistan, and the Philippines, plus other languages in Africa and North America, including:

- Arabic script additions that include honorifics and additions for Quranic use, and characters used to write languages across Africa, the Balkans, and South and Southeast Asia
- The Vithkuqi script historically used to write Albanian and currently undergoing a modern revival
- The Tangsa script used to write the Tangsa language, spoken in India and Myanmar
- The Toto
What's New in Unicode 14.0

Keith Broni  Sep 14, 2021 - 8 min read
Introduced by Unicode 12.1 2019
Computers have gotten better at typography, but they’ve made things more difficult for NLP.
Computers have gotten better at typography, but they’ve made things more difficult for NLP.
Things that can cause trouble:

- Contextual characters ("quotation marks" rather than "quotation marks")
- Newlines (LF, CR, or CRLF)
- Ligatures (sc**off**)
- Combining characters (internally)
Demo: *Reading and writing files in Python*
As an example of a file to read, we will use a relatively small, unannotated document from Project Gutenberg.

Project Gutenberg is an online collection of texts whose copyright has expired. It contains texts in many languages.

Download the First Project Gutenberg Collection of Edgar Allan Poe:

gutenberg.org/cache/epub/1062/pg1062.txt
with open("/Users/jgordon/Downloads/pg1062.txt") as f:
    for line in f:
        print(line)
Because we didn’t specify the encoding, Python will use its default.

What’s the default? 🤷

It depends on the system you’re running on!
with open("/Users/jgordon/Downloads/pg1062.txt", encoding="utf-8-sig") as f:
    for line in f:
        print(line)
EDGAR POE

LE

SCARABÉE D'OR

Traduction de Charles Baudelaire

ILLUSTRATIONS EN COULEURS ET EN NOIR

PAR

GEORGES ROCHEGROSSE

PARIS

LIBRAIRIE DES AMATEURS

A. FERROUD. — F. FERROUD, Successeur

127, Boulevard Saint-Germain, 127

1926
Download gutenberg.org/ebooks/67094.txt.utf-8
with open("/Users/jgordon/Downloads/pg67094.txt",
    encoding="utf-8-sig") as f:
    for line in f:
        print(line)

Compare with

with open("/Users/jgordon/Downloads/pg67094.txt",
    encoding="latin1") as f:
    for line in f:
        print(line)
If we want to write a file, change the "r" mode to "w":

```python
with open("myoutfile.txt", "w") as f:
    print("First line of output", file=f)
    print("Second line of output", file=f)
    print("Here's a number:", 5, file=f)
```
To do for Tuesday

Log in to Slack

Log in to Gradescope

Fill out course survey if you haven’t

Read Jurafsky & Martin §2.1 on regular expressions and fill out worksheet (posted tomorrow).

Check your email later today for invitations to Slack and Gradescope
To do for Wednesday night

Assignment 0: Python warm-up
We’ve got a big journey ahead of us. I hope you’re excited!
Acknowledgments

This lecture incorporates material from:

Katrin Erk, University of Texas at Austin
Xanda Schofield, Harvey Mudd College