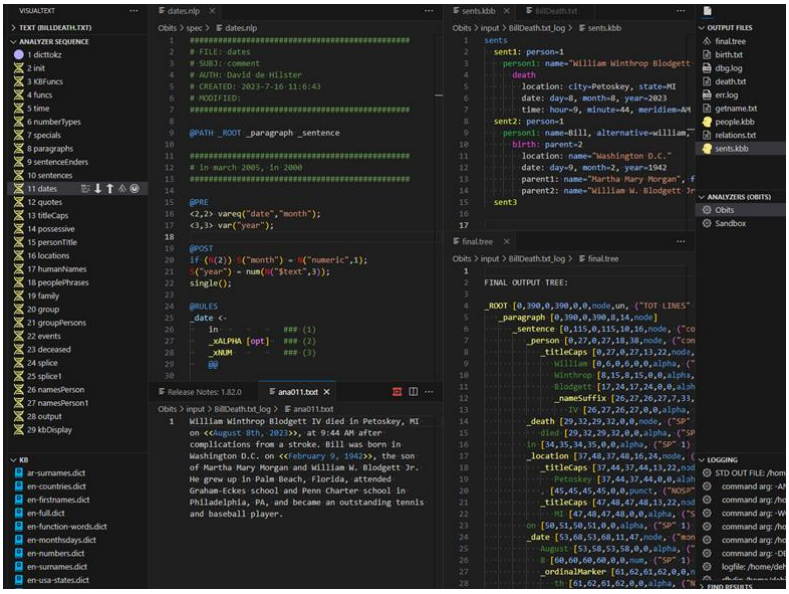
Rule-Based NLP Programming Course

With David de Hilster



STEM COURSE

This course uses human language as a gateway into STEM using a visual programming language that can be taught to non-programmers.

David de Hilster

(310) 991-5744
nluglob.org
david@dehilster.com

Course Page QR

Hands-On Intelligent Programming using the NLP++ VSCode Extension

Students will be exposed to computer programming using VisualText with the following three goals:

1. To teach students how to examine their thought processes while reading and understanding text and translate that into NLP++. This is what we call “Intelligent Programming”.
2. Introduce students to trustworthy rule-based natural language processing.
3. To learn a computer programming language that will serve students well into their futures.

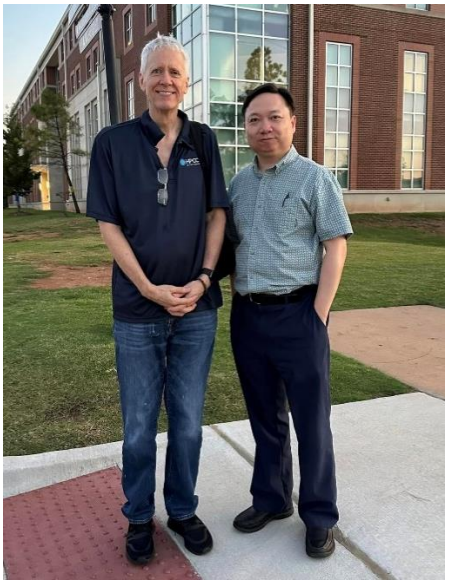


Photo: 1 David with Dr. Gang Qian, the Dean of Computer Science at the University of Central Oklahoma in September 2023

ABOUT THE TEACHER

David de Hilster has been in the Natural Language Processing industry for over 40 years and is co-author of the intelligent computer programming language, NLP++.

Recently, he has been mentoring college and high school students in AI and NLP from universities around the world and has given several in-person and remote talks.

The projects included sentiment analyzers for cricket in English and for soccer games in Portuguese, resume processing, and creating dictionaries for English, Nepali, and Chinese.

David has a Masters in Linguistics and a Bachelor of Science in mathematics from Ohio State University where he taught undergrad linguistics.

Experience in Teaching

David de Hilster is popular with his students and interns. He received high ratings for student mentorship at LexisNexis as well as a student teacher at Ohio State University.

Photo: 2 David with students and Faculty at Clemson University in August of 2022.

About the Courses

Here are some of the suggested NLP courses using NLP++:

- **Course #1:** NLP++: Encoding What Humans Think
- **Course #2:** NLP++ Knowledge Building and LLMs
- **Course #3:** Ethical and Environmental Implications of NLP++ Versus Statistical Methods
- **Course #4:** NLP++ for Programmers

Course #1

NLP++: Encoding What Humans Think

This course introduces students to a new way of programming. It requires them to examine their own thought processes and encode then into the NLP++ programming language.

Course Topics

- **Thinking like a human**: traditional programming requires students to think like a computer whereas programming in NLP++ requires students to think like humans. (see video tutorial)
- **Rule based versus statistics:** learn the advantages of rule-based systems which do not carry the problems of needing large data sets to train, do not have problems with copyright, and are trustworthy given they are 100% traceable in their decision-making processes. (see article by David de Hilster)
- **Bootstrapping and Open Source:** discuss with students how NLP++ is open-source and how dictionaries and knowledge can be constructed and shared by all. Talk about what David de Hilster calls the 5th revolution in Linguistics.
- **Using VisualText:** explain the directories structure for analyzers. Talk about the connectability between trees, text, and rules and discuss the dictionaries and knowledge bases is available in the KB view as well as blocks that are available when creating a new text analyzer.
- **Tokenization:** NLP++ have various tokenization types available including immediate dictionary lookup to character-based tokenization. Talk about tokenization for emojis and special languages like Tamil and Nepali that have logic in the c++ NLP Engine code for these special cases. Emojis are particularly fun. (see video tutorial)
- **Dictionaries:** learn about the NLP++ dictionaries files and what is current available with the VSCode NLP++ language extension. Who dictionaries can hold ambiguity information and how students can quickly create dictionaries using NLP++ itself in a bootstrapping method. Instructors and students can take a look at NLP++ co-author David de Hilster's analyzer repo to see how he used NLP++ to generate many of the dictionaries available today.
- **Knowledge base**: learn the format of the KBB files and look at the existing KBB files currently available to users. Discuss with students how knowledge can be used to help understand text and even store and use short-term memory while parsing the text.
- **Island-Driven, Sequential Processing:** learn about how the sequence file works in NLP++ and how that solves many of the combinatoric and ambiguity problems of older traditional rule-based systems. Talk about the different types of NLP files that can be executed and how they work. For example, all functions are parsed from the entire sequence before the analyzer begins. (see video tutorial)
- **Trees Versus Knowledge Base:** NLP++ has a built-in tree structure and knowledge base. Discuss when one would keep parsed information on a tree and when to include it in a knowledge base. Discuss why being able for tree nodes to be pointing into the knowledge base is crucial for understanding text.
- **NLP++ Rules:** examine and learn the NLP++ rule system and talk about the power of being able to do a myriad of actions after the matching of rules as well as using preconditions. Discuss the superiority of the

NLP++ rules over Regex.

- **Anaphora:** discuss how NLP++ can solve anaphora in a very human-like way and how it is essential to use the knowledge base for this task. (see video tutorial)
- **Ambiguity:** teach the built-in mechanisms in NLP++ that help in resolving ambiguity. (see video tutorial)
- **Recursion:** Learn how to use recursion using NLP++ rules. (see video tutorial)

Exercises and Tests

- **Addresses:** This is a great exercise for students to build their first analyzer and learn NLP++'s rule system. It makes them learn how to use existing dictionaries that come with NLP++
- **Sentiment analysis:** These are very specific and demonstrate to students that sentiment analysis is extremely specific to a subject matter and task and how NLP++ allows students to create analyzers that are finely tuned to that specific task. This can be done by everyone in class given each sentiment analyzer can easily be built from scratch and the variety of these analyzers are limitless (see sentiment by past student interns).
- **Resume processing:** This is a great exercise for students in that they have to not only deal with text content, but the variation in text format. Processing the formatting of resumes alone is a formidable project in itself and along with processing education, skills, job history etc, this easily can turn into a group project.
- **Wiki to Dictionary and KBB Files:** Students will deal with using NLP++ to parse the wiki text format to get at content from Wiktionary and Wikipedia pages. (see article and papers)
- **Name Entity Recognition:** have students tackle this problem as a whole or in parts in a group. This is a harder problem and could be one-project for the entire group.
- **Anaphora:** write an NLP++ analyzer to resolve anaphora in text. This is a great exercise for students to learn how humans resolve this.

Course #2

NLP++ Knowledge Building and LLMs

This course introduces students to a new way of programming. It requires them to examine their own thought processes and encode then into the NLP++ programming language.

Course Topics

- **Knowledge Migration to Computers**: discuss what David de Hilster calls the Fifth Linguistic Revolution. Why computers must be "given" certain knowledge and why it can't learn this simply from statistics and reading text. Discuss how we erroneously assign meaning to statistical relationships and responses from LLMs.
- **NLP++ Output to Python:** how do you write and NLP++ analyzer to take input from Python and return output. Talk about generating JSON strings and JSON objects using the NLPPlus Python package.
- **NLPPlus Python Package:** how to use the NLP++ python package. Talk about how the package loads the dictionaries and knowledge bases on initialization and then can call the analyzer multiple times.
- **Untrustworthiness of Statistical NLP Methods**: talk about statistical methods such ML, NN, and LLMs and how they limit their usefulness in industry. Contrast that to the rule-based system of NLP++.
- **NLPPlus Python Package to Create Self-Learning Systems**: learn how NLP++ can take untrustworthy data and turn it into trustworthy data that can be used in trustworthy NLP. Discuss why NLP++ cannot fix untrustworthy data that comes out of statistical systems given that statistical errors are not logical. Discuss how using the NLPPlus Python package could be used to create automatic learning systems instead of programmers having to create new dictionary entries or knowledge by hand. (see video tutorial)
- **Visual Versus Hidden Problems:** discuss how problems in NLP++ are visual and can be easily corrected over statistical methods whose problems are impossible to fix logically. Talk about when NLP++ uses LLMs or a webpage that if the information is incorrect, humans will spot it and correct it – something not possible with statistical methods.

Exercises and Tests

- **Building Dictionaries and KBs:** have students use NLP++ to build new dictionaries and KB files for NLP++ and the open-source project. Have students look at David de Hilster' analyzer repo where he stores his NLP++ analyzers that build many of the dictionaries and KBs that are in the NLP++ library.
- **Self-Learning Systems with NLP++ and NLPPlus**: create an NLP++ analyzer to do a specific task. Identify unknown words or concepts. Use the NLPPlus Python package to call a webpage or LLM find the unknown word or knowledge. Call a second NLP++ analyzer to read the answer and create new dictionary or world knowledge.

Course #3

Ethical Implications of NLP++ Versus Statistical Methods

With statistical NLP, there are ethical and environmental issues that need to be addressed. This course will introduce students to these topics and how the compare with NLP++.

Course Topics

- **Ethical Implications of Statistical NLP Methods:** Copyright, access to private data are legal problems that plague statistical methods. The huge costs of large language models make it only accessible to huge tech companies leaving others behind.
- **Environmental Implications of Statistical Methods:** In order to increase accuracy and coverage, more and more tokens are needed and bigger and bigger systems., The monetary and Enviromental costs are substantial and is it all worth it given statistical models will always be wrong.
- **Ethical NLP Development:** NLP++ was developed with ethical concerns—such as privacy, bias, and AI ethics—in mind. Its transparent and prescriptive nature leads to responsible and fair NLP applications.
- **Practical Applications:** NLP++ is not just a research tool; it is a framework built for real-world conditions. It is scalable across various industries, including healthcare, legal, and finance, where accuracy and traceability are paramount.
- **Educational Impact:** NLP++ serves as an educational tool, helping students and developers understand the intricacies of human language processing. It bridges the gap between theoretical linguistics and practical NLP development, making it an invaluable resource for learning and innovation.
- **Singularity**: will statistical methods eventually lead to uncontrolled AI and conversely, will NLP++ allow us to avoid the singularity?

Exercises and Tests

- Pick an NLP application and write about the ethical and environmental implications if using statistical methods versus rule and knowledge based.
- Come up with a course for high school students that teaches them the ethics and environmental impacts of statistical NLP methods and include NLP++ as a contrasting technology.

Course #4

NLP++ for Programmers

NLP++ is a programming language that all programmers should know and learn. All programmers come across files that they need to process that neither Regex nor NLP Toolkits can solve. It could be html files, xml files, or text documents which need to be skimmed to perform a particular task but is way too complex for Regex and which are too specific to use generalized NLP toolkits.

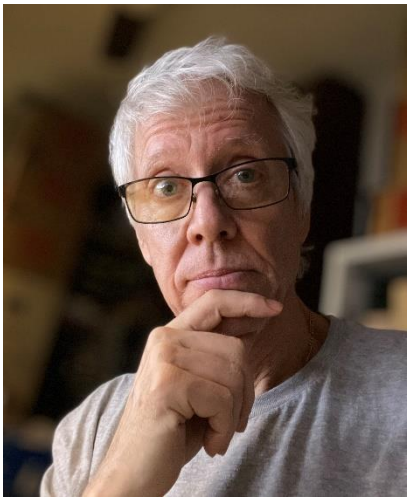
Programmers can use NLP++ to perform tasks on text that before were simply not practical or deemed impossible.

Course Topics

- **History of NLP++**: talk about the journey of co-authors Amnon Meyers and David de Hilster and how they arrived at the conceptual grammar, NLP++, and VisualText. Talk about how the trees, rules, functions, and knowledge base work together to mimic how humans read and understand text.
- **Comparing NLP++ with Regex**: discuss readability and the limitations of Regex. Talk about how the tree structure, NLP++ rule system, the conceptual grammar combine to allow for mimicking human beings ability to skim text and perform useful tasks.
- **Comparing NLP++ with Traditional NLP Python packages:** talk about the problems with generic NLP as well as the inability of traditional packages to be modified.
- **NLP++ Language Extension for VSCode**: have students download the NLP++ language extension and talk about the Text view, Analyzer view, KB view, and bottom panel Output, Analyzer, Find, and Log views.
- **NLPPlus Python Package**: discuss the NLP++ Python package and how to use the standard NLP analyzers that come with it. Discuss how to build NLP++ analyzers that return usable data back to Python including the output.json file, and using the JSON functions in KBFuncs.nlp.

Exercises and Tests

- **Do not translate tags**: have students pick a document that has computer code embedded (perhaps a computer language manual) and build an NLP++ analyzer to add <do_not_translate> tags.
- **Database mining**: have students find webpages with information that can be put into structure data. Use NLP++ and compare it to using something like the Python package BeautifulSoup. Which one is simpler on which types of webpages? Which one is more readable and more maintainable?
- **Contribute to the NLPPlus library analyzers**: have students find a specific Python package that parses telephone numbers or other simpler patterns. Recreate that in NLP++ and submit it to GitHub to be included in the NLPPlus library of analyzers that can be used in edited by python programmers.



David de Hilster

Email: david@dehilster.com

Cell: 310-991-5744

Websites
http://nluglob.org
http://dehilster.com

Industry Experience

LexisNexis, Boca Raton, FL
NLP Engineer – NLP++ plugin for HPCC Supercomputing System

Text Analysis Internation, Mountainview, CA
NLP Engineer – Co-Author of NLP++

I-Search, Los Angeles, CA
NLP Engineer – Resume processor

Battelle Memorial Institute, Columbus, OH
NLP Engineer – NLQ Natural Language Interface to Databases

Teaching Experience

LexisNexis, Boca Raton, FL
Student Mentor – Mentored high school and universities students in NLP and AI in the United States, India, and Brazil

The Ohio State University, Columbus, OH
Student Teacher – Student teacher for undergraduate Linguistics

Testimonial *(from LinkedIn)*

Hugo Watanuki
Information Engineer, LexisNexis Risk

"I had a very positive experience working with David de Hilster in different educational settings around the globe. One of David's standout qualities is his dedication to mentoring and guiding the next generation of talent. A prime example of this is his recent work mentoring an undergraduate student from the University of Sao Paulo in Brazil, on an innovative project involving emotion detection in social media posts. Under David's mentorship during a yearlong project, the student was able to develop a sentiment analyzer specifically for tweets in Brazilian Portuguese related to soccer games, showcasing both technical sophistication and practical application.

David's expertise in NLP and his hands-on approach to mentoring were crucial to the project's success. He provided the student with invaluable insights into building a real-world NLP system, emphasizing the importance of creating practical software rather than relying on generic "toy systems." The result was a sentiment analyzer with impressive accuracy and real-world potential, demonstrating David's ability to foster both technical excellence and innovative thinking in his mentees.

In addition to his technical prowess and mentoring skills, David is also renowned for his global impact as an educator. He has delivered lectures and workshops at prestigious universities around the world, including institutions in the United States, Brazil, and India. His ability to convey complex concepts in accessible terms has made him a sought-after speaker and educator, further showcasing his commitment to advancing knowledge and fostering international collaboration in the field of NLP.

David's blend of deep technical knowledge, effective mentorship, and global teaching experience makes him an exceptional asset to any educational organization in the world. His impact on both the practical and educational aspects of NLP is truly commendable."