Digital Humanities Applications of spaCy's Span Categorizer

Boyd, Adriane
adriane@explosion.ai
ExplosionAI

Kádár, Ákos
akos@explosion.ai
ExplosionAI

Janco, Andrew
apjanco@upenn.edu
University of Pennsylvania

Lassner, David
lassner@tu-berlin.de
TU Berlin

Budak, Nick
budak@stanford.edu
Stanford University

Tasovac, Toma
ttasovac@humanistika.org
DARIAH

Ermolaev, Natalia
nataliae@princeton.edu
Princeton University

Karajgikar, Jajwalya
jajk@upenn.edu
University of Pennsylvania

Digital Humanities Applications of spaCy's Span Categorizer

Brief Abstract

This 3-hour workshop will introduce span categorization as a method for the machine-annotation of text for various research tasks in the digital humanities. Participants will gain a conceptual understanding of how span categorization differs from entity recognition and complete practical exercises to train a spaCy span categorizer on the LitBank dataset.

Description

Annotation is fundamental to the analysis of text. Scholars mark significant sections of a text with key ideas and phrases. By marking the text, we add information, create connections, and form interpretations. This text markup can be captured as data through systems such as TEI and used as training data for statistical language models. Automating this process makes it possible to annotate a large corpus of texts for research tasks such as creating digital editions or text mining.

Named entity recognition (NER) is a common method for identifying consistent entities such as the names of people, places, and organizations in a text. In our DH2019 workshop "Introduction to Natural Language Processing for DH Research with SpaCy" (more), we demonstrated how digital humanists could train a NER model to identify early modern place names using training data from a TEI encoded text from the Perseus Collection.

Named entity recognition is very effective for finding clear syntactic units like proper names or place names. However, digital humanists frequently work with literary and historical materials whose names, places, and ideas have significant variations. Digitized documents contain OCR errors which adds further variation. The objects we mark in a text can also be abstract, such as allusion, metaphor, or affect. For example, the "Mapping Imagined Geographies" project (link) seeks to annotate place-related concepts and toponyms in poetic texts. Many of these are imagined or metaphorical places such as "the promised land" or "the new world." A place can play many roles in a text, such as the setting or the destination of a journey. Span categorization opens a larger vocabulary of overlapping labels and roles for the analysis of literary texts.

Additionally, recent work has demonstrated that the recognition of non-named entities can be a powerful tool for decolonizing historical texts. Enslaved people are often mentioned but not by name. There is key information in historical records about these people and their history. Span categorization provides the ability to find unnamed entities in a large corpus of texts. The LitBank dataset offers annotations from literary texts with annotations for a wide range of people that are mentioned but whose proper name is not given.

Learning Outcomes

By the end of the workshop, participants will have:

- A conceptual understanding of natural language processing and its uses in digital humanities research.
- Be able to assess when named entity recognition or span categorization is the better tool for a research task.
- practical experience training a spaCy component for ner and spancat.

Outline

1. Introduction to natural language processing and span categorization (15 minutes)
2. Comparison of named entity recognition and span categorization (15 minutes)
3. Exercise to manually annotate literary texts using Prodigy (20 minutes)
4. Introduction to the LitBank dataset (10 minutes)
5. Exercise to train an ner component on the Litbank data (20 minutes)
6. Break (10 minutes)

7. Training a span categorization model with the LitBank data (20 minutes)
8. Compare and interpret the differing results of the two models (20 minutes)
9. Designing your own suggester function to improve span categorization (20 minutes)
10. Use Streamlit to create a demonstration app to share your model and results (10 minutes)

[160 minutes total]

Our collaboration

This workshop builds on an existing collaboration between the Princeton Center for Digital Humanities and the developers of the spaCy open-source NLP library. At DH2019, the organizers offered a workshop on “Introduction to Natural Language Processing for DH Research with SpaCy” (more). The session was well attended and featured a valuable discussion between digital humanists and the spaCy developers. Academic research brings new problems and use cases that can inform the design of open-source tools. A tool that David Lassner created for the session was later awarded the 2021 Rahtz Prize for TEI Ingenuity. The organizers build on this experience for an NEH-funded Institute called “New Languages for NLP: Building Linguistic Diversity in the Digital Humanities.” The new languages project developed instructional materials to train digital humanists how to annotate texts and train spaCy models for the languages and domains of their research. Ines Montani, the co-founder of Explosion and core developer of spaCy participated in the workshop and gave a keynote lecture.

Target audience and expected number of participants

We expect to have 20-30 participants for the workshop based on past experience.

This is an intermediate-level workshop and participants should come with basic knowledge of Python. Participants will need to bring a laptop or tablet and keyboard with them to the session. A wifi connection is needed to connect to the cloud notebooks.

Technical Support

For the workshop, we will provide virtual machines that participants can access with a browser during the workshop using their laptops or tablets. The only requirement will be an internet connected device. All materials will be shared on GitHub for later reference.

Organizers

- **Adriane Boyd** is a computational linguist who has been engaged in research since 2005, completing her PhD in 2012. She has extensive experience in quality control for linguistic annotation, parsing, and NLP for non-standard language.
- **Ákos Kádár** is a Machine Learning Engineer at Explosion and developer in Natural Language Processing.
- **David Lassner** holds a Ph.D. from the machine learning group at TU Berlin.
- **Nick Budak** is a Digital Library Software Developer at Stanford University. Nick enjoys imagining and implementing accessible, dynamic interfaces for digital humanities projects. His current research is in the area of computational phonology tools for Old Chinese through the DIRECT project. At Princeton, he developed backend and frontend interfaces for flagship DH projects like the Shakespeare and Company Project and Princeton Prosody Archive.
- **Natalia Ermolaev** is the Associate Director of the Center for Digital Humanities at Princeton University and the Project Archivist for the Serge Prokofiev Archive at the Rare Books and Manuscript Library at Columbia University. She is a Slavist, digital humanist, and archivist.
- **Toma Tasovac** is President of the Board of Directors of the pan-European Digital Research Infrastructure for the Arts and Humanities (DARIAH). With an academic background in Comparative Literature and degrees from Harvard, Princeton and Trinity College Dublin, Toma's areas of scholarly expertise include historical and electronic lexicography, data modeling, digital editions, and research infrastructures.
- **Andrew Janco** is a research software engineer at the University of Pennsylvania Libraries. He has organized multiple workshops connecting natural language processing and digital humanities research, including DH2019, DH Budapest, and the “New Languages for NLP” Institute. Andy currently serves as co-Vice President of the Association for Computers and the Humanities (ACH).
- **Jajwalya Karajgikar** is the Applied Data Science Librarian at the University of Pennsylvania Libraries. Jaj engages with researchers across disciplines interested in employing techniques for data storytelling, natural language processing, computational social sciences, data visualization, network analysis, and text mining. She works with campus partners to establish foundational programming in research computing, data literacy, and data ethics.