MIlestone 1 Proposal and Data Selection

Syllabus and Legal Document Keyword Extraction

Muduo Wang

Fall 2021

https://github.com/Mwang413/Mwang413

Which Domain?

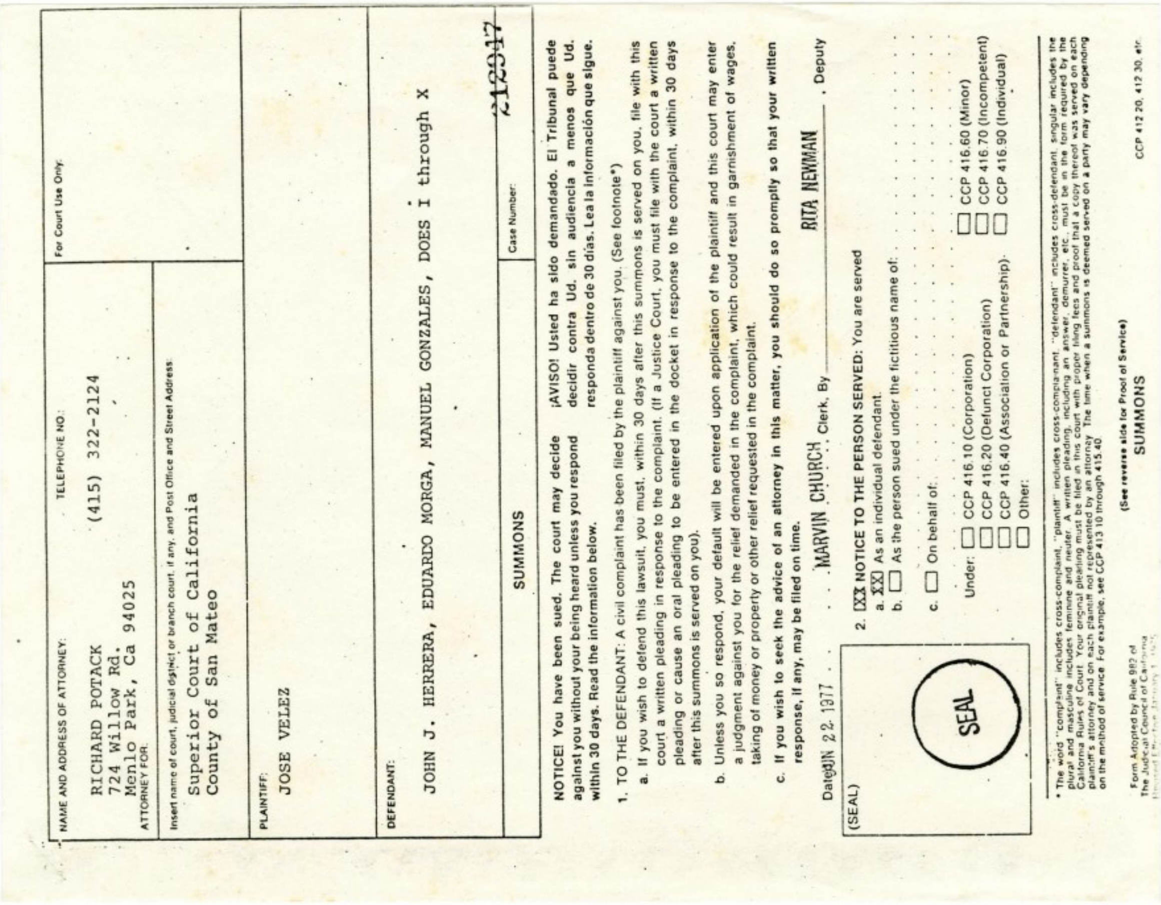
The data will come from two domains, legal and education. I will be working with NLP models and processes to extract important information from given documents.

Type of Legal Data:
https://texashistory.unt.edu/ark:/67531/metapth251296/
Amazon Textract:
https://aws.amazon.com/textract/
spaCy (spelled with lowercase "s"):
https://spacy.io/
spaCy Tutorials:
https://www.youtube.com/watch?v=WnGPv6HnBok
spaCy Models:
https://spacy.io/usage/models
Training models with spaCy:
https://course.spacy.io/en/chapter4
spaCy + Fuzzy Matching:
https://github.com/gandersen101/spaczz
Fuzzy Matching:
https://www.datacamp.com/community/tutorials/fuzzy-string-python
RegEx:
https://docs.python.org/3/howto/regex.html
NLP with Neural Networks:
https://www.youtube.com/watch?v=X2vAabgKiuM
Uvicorn App Deployment:
https://www.uvicorn.org/deployment/

Which Data?

The legal documents will come from a client that has given us a project to extract key information. Specifically, I will be working with documents that came from California.

Below is an image of one closely resembling the documents that I will be working with.



NAME AND ADDRESS OF ATTORNEY: TELEPHONE NO.:

RICHARD POTACK (415) 322-2124
724 Willow Rd.
Menlo Park, Ca 94025

ATTORNEY FOR:

For Court Use Only:

Insert name of court, judicial district or branch court, if any, and Post Office and Street Address:

Superior Court of California
County of San Mateo

PLAINTIFF:

JOSE VELEZ

DEFENDANT:

JOHN J. HERRERA, EDUARDO MORGA, MANUEL GONZALES, DOES I through X

SUMMONS

Case Number: 212947

NOTICE! You have been sued. The court may decide against you without your being heard unless you respond within 30 days. Read the information below.

¡AVISO! Usted ha sido demandado. El Tribunal puede decidir contra Ud. sin audiencia a menos que Ud. responda dentro de 30 días. Lea la información que sigue.

1. TO THE DEFENDANT: A civil complaint has been filed by the plaintiff against you. (See footnote*)

a. If you wish to defend this lawsuit, you must, within 30 days after this summons is served on you, file with this court a written pleading in response to the complaint. (If a Justice Court, you must file with the court a written pleading or cause an oral pleading to be entered in the docket in response to the complaint, within 30 days after this summons is served on you).

b. Unless you so respond, your default will be entered upon application of the plaintiff and this court may enter a judgment against you for the relief demanded in the complaint, which could result in garnishment of wages, taking of money or property or other relief requested in the complaint.

c. If you wish to seek the advice of an attorney in this matter, you should do so promptly so that your written response, if any, may be filed on time.

DateJUN 22 1977 MARVIN CHURCH Clerk, By RITA NEWMAN, Deputy

(SEAL)

SEAL

2. ☒ NOTICE TO THE PERSON SERVED: You are served

a. ☒ As an individual defendant.

b. □ As the person sued under the fictitious name of:

c. □ On behalf of:

Under: □ CCP 416.10 (Corporation) □ CCP 416.60 (Minor)
□ CCP 416.20 (Defunct Corporation) □ CCP 416.70 (Incompetent)
□ CCP 416.40 (Association or Partnership) □ CCP 416.90 (Individual)
□ Other:

* The word "complaint" includes cross-complaint, "plaintiff" includes cross-complainant, "defendant" includes cross-defendant, singular includes the plural and masculine includes feminine and neuter. A written pleading, including an answer, demurrer, etc., must be in the form required by the California Rules of Court. Your original pleading must be filed in this court with proper filing fees and proof that a copy thereof was served on each plaintiff's attorney and on each plaintiff not represented by an attorney. The time when a summons is deemed served on a party may vary depending on the method of service. For example, see CCP 413.10 through 415.40.

Form Adopted by Rule 982 of
The Judicial Council of California
Revised Effective January 1, 1975

(See reverse side for Proof of Service)

SUMMONS

CCP 412.20, 412.30, etc.

[1]

I will be extracting the name, address of the Plaintiff, the Defendant, and the Attorney, as well as the address of the courthouse.

The second type of documents are course syllabi. For these documents, I will be extracting keywords with matching tools, such as spaCy, RegEx and Fuzzy Matching, and Neural Network training. The client does not provide rules regarding what constitutes a keyword; it is left to my discretion what the keywords are in the document.

Research Questions? Benefits? Why analyze these data?

The research questions are:

What methods can we use to best extract keywords from text?

Which language rules are best to extract keywords accurately?

The approach is doing research about NLP rules that we can use, continually finding new ones, and implementing them exhaustively and efficiently. The benefits of this research is to automate the process of reading and extracting key information from

[1] PC: https://texashistory.unt.edu/ark:/67531/metapth251296/

texts, saving cost of labor to read the text by human eyes, as well as increasing the accuracies of current text-extracting methods and models.

What Method?

The method is to learn from spaCy tutorials, Google a lot of questions and read on NLP, implement models, trial and error, adjust, and produce models and rules to extract the keywords with high accuracy.

spaCy helps with the development of language rules, which processes the document based on patterns. There are also a lot of articles and YouTube tutorials on NLP work.

Potential Issues?

The primary challenge that this project contains is to come up with rules. NLP engineers must study the language, think outside the box, and analyze the language in ways like a grammarian as well as an AI developer at the same time. This can be very hard to do, especially when sometimes rule-based pattern-matching requires one to be both exhaustive and efficient at the same time. It combines very abstract work (NLP) with very arithmetic work (AI).

Another difficulty that I anticipate having is trouble adopting other codes that are supposed to help with my project. In a lot of machine learning tasks, borrowing shared code can accelerate tasks, however, I have a hard time reading others' codes and logic.

Concluding Remarks

In conclusion, this project will present some of the ways NLP works, specifically, rule-based matching is done on different sets of data. The project uses spaCy, a NLP API service, to create rules and train models to recognize keywords from textual data from legal documents and course syllabi.

Milestone 2 **Check Point**

Any surprises from your domain from these data?
For the legal documents, there was some difficulty extracting the text itself. The text extraction process that we started with, which reads texts from PDF (the format of the legal document), is not doing a great job at extracting the text.
For the course syllabi, it was hard to know which words are important for the description of the course. Nothing else proved difficult.

The dataset is what you thought it was?
The legal document also uses a stamp for one of the key information, the case number, which became very difficult also for our text extraction model to recognize.
From the course syllabus, the surprise was how difficult it was to determine what is important and what isn't to the course. The course contains around 10-20 sentences and phrases, along with many other bullet-point-like objects, such as dates, assignments, and regulations.

Have you had to adjust your approach or research questions?
No, my approach is showing itself to be working. There are tons of great resources out there, and I am able to learn a lot of them. However, sometimes it's difficult to see whether a tutorial or an article is relevant and useful until some time is spent understanding it. This cost some time that would be avoidable if I was more experienced in NLP.

Is your method working?
My method of thinking outside the box and finding the maximally efficient as well as maximally exhaustive has been working. However, it is difficult, and requires a lot of abstract thinking. And that's what I spent a lot of time doing: thinking. This is an interesting type of work, because it sometimes feels like I'm not doing anything productive, but I always feel that I have no time to do anything.

What challenges are you having?
The challenges, like I've mentioned above, include extracting text correctly from a PDF, instating principles for determining the keywords to course syllabi, finding relevant resources quickly, as well as balancing between efficiency and exhaustivity.

Milestone 3 **Whitepaper**

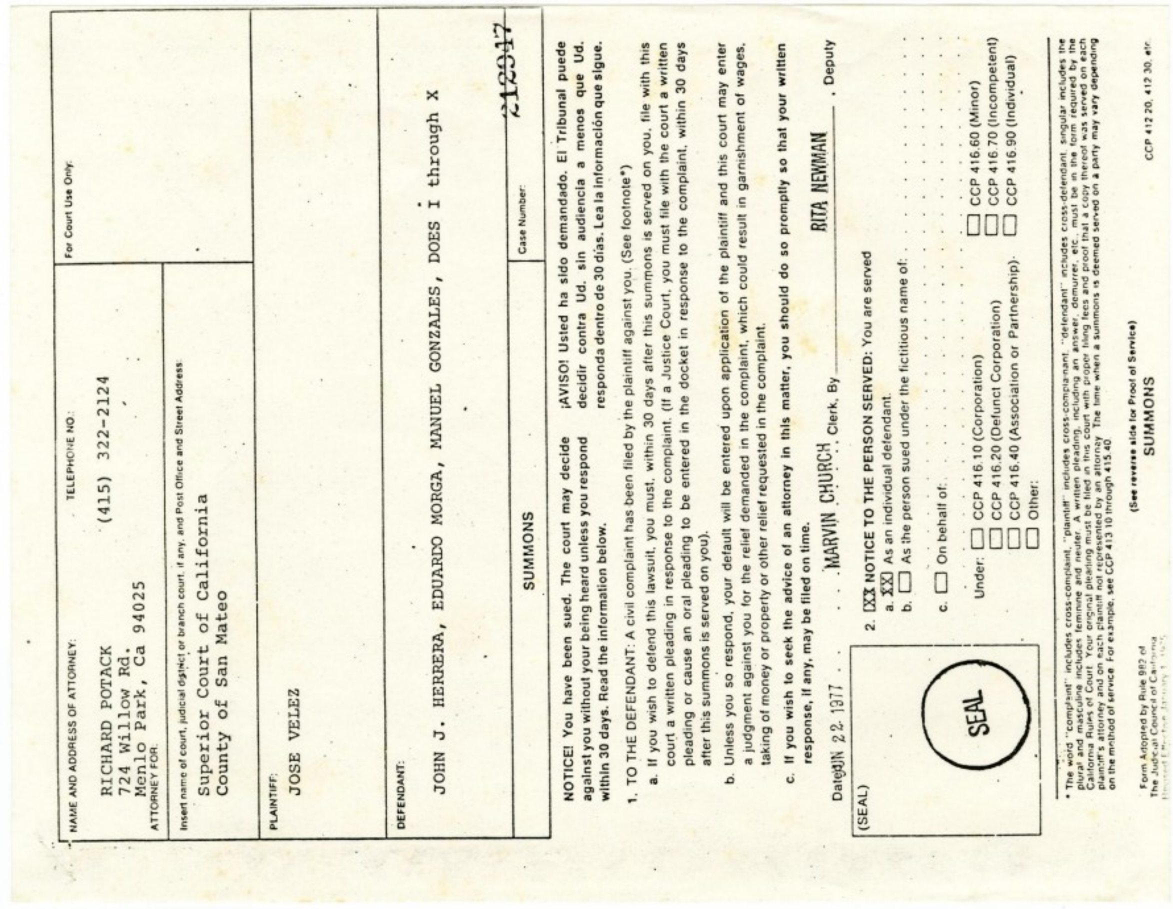
Introduction

There are three different parts to this paper: the first is to present the data itself, the second is to exhibit work done for the data, the third is to show research done and tools acquired for the data but not yet implemented. This is an on-going project, and will continue to be worked on after the submission of this project, since many clients to Width.ai (my workplace) continually request Natural Language Processing (NLP) services for their data (2021).

Data

For NLP projects, the data obviously falls into some sort of language. Most of the NLP projects done at Width.ai are in English, and occasionally, there are also requests for Spanish textual data. The specific two types of data that will be presented in this paper are legal documents and course syllabi. The legal document comes in the form of PDFs, and number in thousands. They contain information like the names of the parties involved, including the Plaintiff, Defendant, Judge, courthouse, type of case, case number, and other kinds of legal information which are mostly ignored because they are there for legality (mostly trivially). These documents mostly fall into the type of a subpoena or other types of legal notices, rather than thousands of paragraphs of legalese. The documents are usually one page long, but can sometimes span the length two pages.

This is an example of a legal document similar to ones found in this dataset:



NAME AND ADDRESS OF ATTORNEY: TELEPHONE NO.: For Court Use Only:

RICHARD POTACK (415) 322-2124
724 Willow Rd.
Menlo Park, Ca 94025

ATTORNEY FOR:

Insert name of court, judicial district or branch court, if any, and Post Office and Street Address:

Superior Court of California
County of San Mateo

PLAINTIFF:

JOSE VELEZ

DEFENDANT:

JOHN J. HERRERA, EDUARDO MORGA, MANUEL GONZALES, DOES I through X

SUMMONS

Case Number: 212947

NOTICE! You have been sued. The court may decide against you without your being heard unless you respond within 30 days. Read the information below.

¡AVISO! Usted ha sido demandado. El Tribunal puede decidir contra Ud. sin audiencia a menos que Ud. responda dentro de 30 días. Lea la información que sigue.

1. TO THE DEFENDANT: A civil complaint has been filed by the plaintiff against you. (See footnote*)

a. If you wish to defend this lawsuit, you must, within 30 days after this summons is served on you, file with this court a written pleading in response to the complaint. (If a Justice Court, you must file with the court a written pleading or cause an oral pleading to be entered in the docket in response to the complaint, within 30 days after this summons is served on you).

b. Unless you so respond, your default will be entered upon application of the plaintiff and this court may enter a judgment against you for the relief demanded in the complaint, which could result in garnishment of wages, taking of money or property or other relief requested in the complaint.

c. If you wish to seek the advice of an attorney in this matter, you should do so promptly so that your written response, if any, may be filed on time.

Dated JUN 22 1977 . . . MARVIN CHURCH Clerk, By RITA NEWMAN , Deputy

(SEAL)

SEAL

2. ☒ NOTICE TO THE PERSON SERVED: You are served

a. ☒ As an individual defendant.

b. □ As the person sued under the fictitious name of:

c. □ On behalf of:

Under: □ CCP 416.10 (Corporation) □ CCP 416.60 (Minor)
□ CCP 416.20 (Defunct Corporation) □ CCP 416.70 (Incompetent)
□ CCP 416.40 (Association or Partnership) □ CCP 416.90 (Individual)
□ Other:

* The word "complaint" includes cross-complaint, "plaintiff" includes cross-complainant, "defendant" includes cross-defendant, singular includes the plural and masculine includes feminine and neuter. A written pleading, including an answer, demurrer, etc., must be in the form required by the California Rules of Court. Your original pleading must be filed in this court with proper filing fees and proof that a copy thereof was served on each plaintiff's attorney and on each plaintiff not represented by an attorney. The time when a summons is deemed served on a party may vary depending on the method of service. For example, see CCP 413.10 through 415.40.

Form Adopted by Rule 982 of
The Judicial Council of California
Revised Effective January 1, 1975

(See reverse side for Proof of Service)

SUMMONS

CCP 412.20, 412.30, etc.

[2]

The second type of documents is the course syllabus. The goal is to extract keywords to courses, which are pertinent to describing the course in the concise and quick to read way. The client requests this service in order to extract insights and integrate the results of this project to assist in their work to browse large amounts of courses for the purpose of approving them as viable courses for transfer credits. Having a language and a keyword extraction model can save lots of time spent viewing individual courses in hopes of understanding their sufficiency for substitution at another institution. Width.ai hopes to process through up to one thousand diverse types of courses in order to create a generalizable pipeline in which the model can process

[2] PC: https://texashistory.unt.edu/ark:/67531/metapth251296/

many types of different courses well. The task is then to browse free and open-access course descriptions and syllabi found in websites belonging to college and other educational institutions.

Process

For the legal documents, the first part of the work is to use a computer-vision model for textual recognition, also called Optical Character Recognition. In this article, Sable demonstrates ways to build custom deep learning OCR models (2021). However, the OCR process continually experiences challenges and err frequently on extracting the text from the PDFs, partially because of the lack of clarity of faxed scans of the documents where the ink often goes missing, and partially due to the general lack of up-to-date systems of all sorts of governmental processes, thus create challenges for the recognition of the documents' content. But once this step is reached, the output of the textual data becomes available to work with and tokenized.

For the course syllabi, there is no need for OCR. However, a web-crawling codeset must be designed for scraping through all the sites that are open to browsing for syllabi, and mine only the useful textual data from these sites. This part of the process was left to Width.ai and was not a part of the current project.

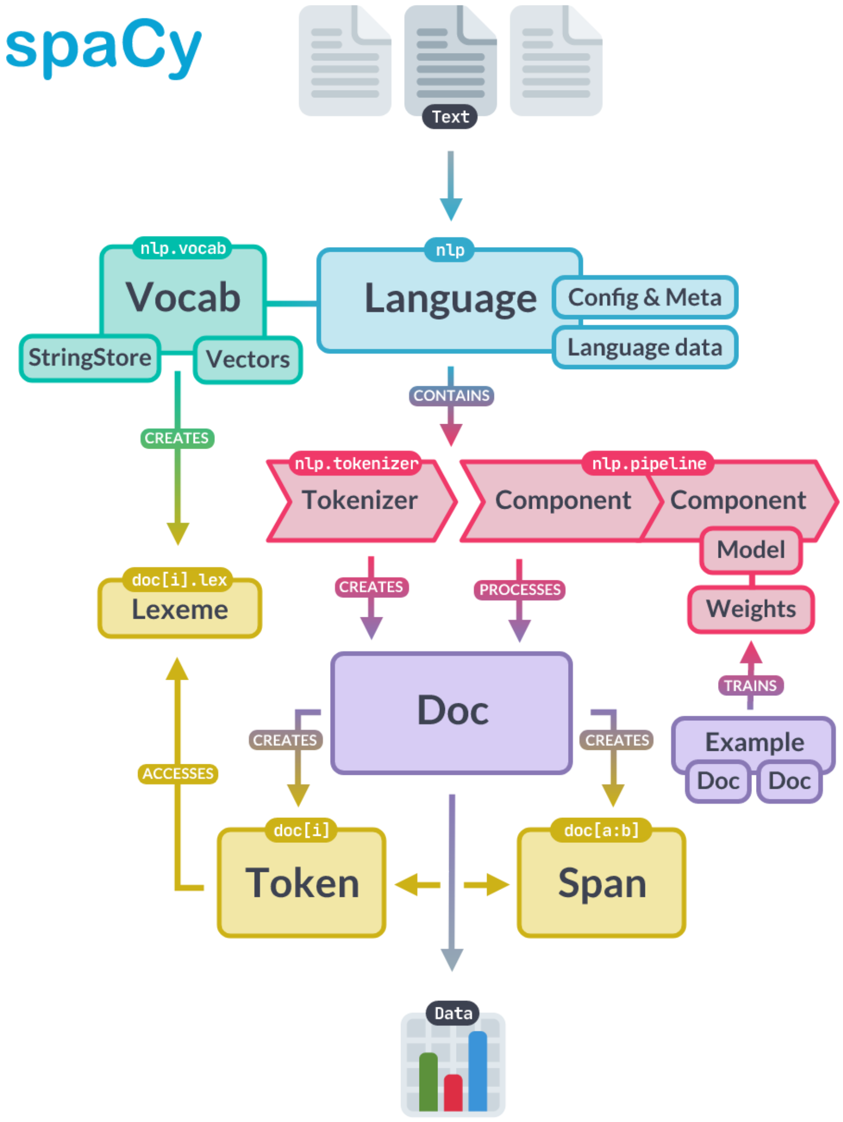
spaCy

At this step, both the legal documents and the course syllabi are turned into clean textual data, and become tokenizable. This is when spaCy comes to play. spaCy provides a large NLP API services, and assists with many different aspects of NLP work.

> spaCy is an open-source software library for advanced natural language processing… [and] focuses on providing software for production usage. spaCy

> also supports deep learning workflows that allow connecting statistical models trained by popular machine learning libraries like TensorFlow, PyTorch or MXNet through its own machine learning library Thinc. Using Thinc as its backend, spaCy features convolutional neural network models for part-of-speech tagging, dependency parsing, text categorization and named entity recognition (NER). Prebuilt statistical neural network models to perform these tasks are available for 17 languages, including English, Portuguese, Spanish, Russian and Chinese, and there is also a multi-language NER model…

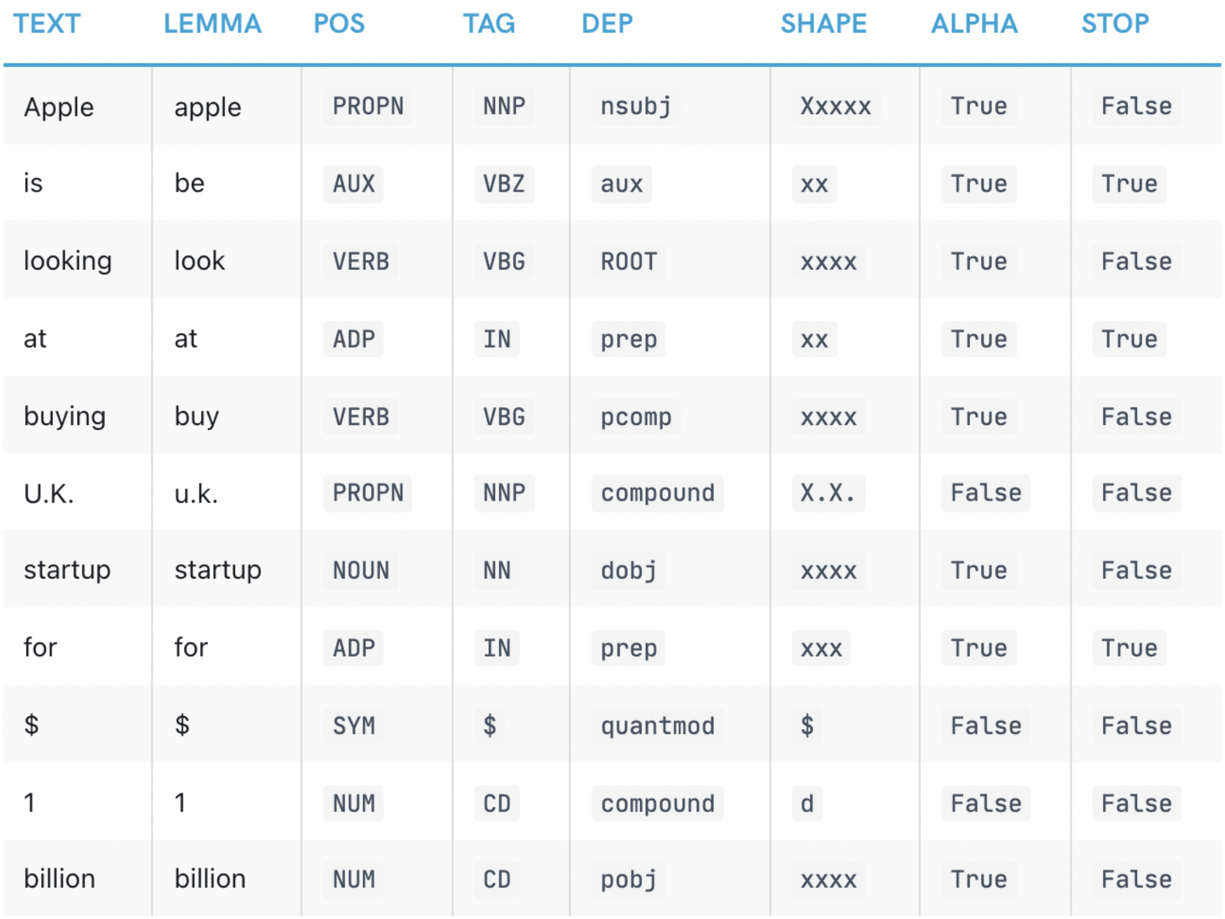
The below illustration graphically presents all the different areas that spaCy is used in the development of NLP pipelines.



[3]

[3] PC: https://spacy.io/api

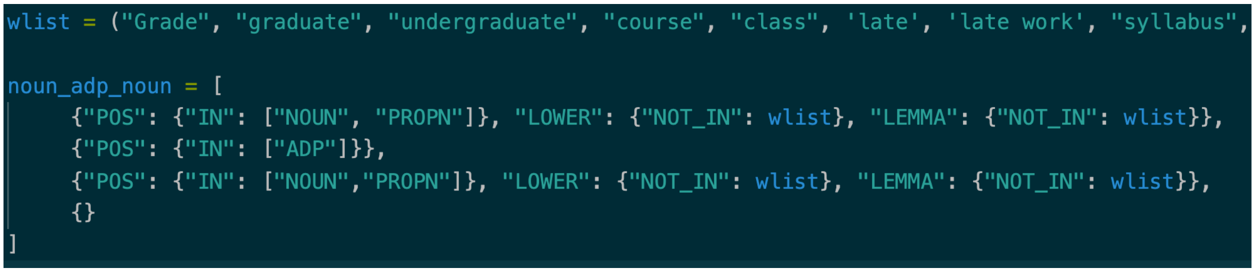
This project primarily uses spaCy for its rule-creation, used for pattern-matching. The first step is to create a pattern (rule) for a key phrase that would be important within a document. The rules are based on a specific word token's different grammatical properties. Here is an illustration of the different ways a word token can be matched by its "linguistic feature" (spaCy, 2021).[4]



TEXT	LEMMA	POS	TAG	DEP	SHAPE	ALPHA	STOP
Apple	apple	PROPN	NNP	nsubj	Xxxxx	True	False
is	be	AUX	VBZ	aux	xx	True	True
looking	look	VERB	VBG	ROOT	xxxx	True	False
at	at	ADP	IN	prep	xx	True	True
buying	buy	VERB	VBG	pcomp	xxxx	True	False
U.K.	u.k.	PROPN	NNP	compound	X.X.	False	False
startup	startup	NOUN	NN	dobj	xxxx	True	False
for	for	ADP	IN	prep	xxx	True	True
$	$	SYM	$	quantmod	$	False	False
1	1	NUM	CD	compound	d	False	False
billion	billion	NUM	CD	pobj	xxxx	True	False

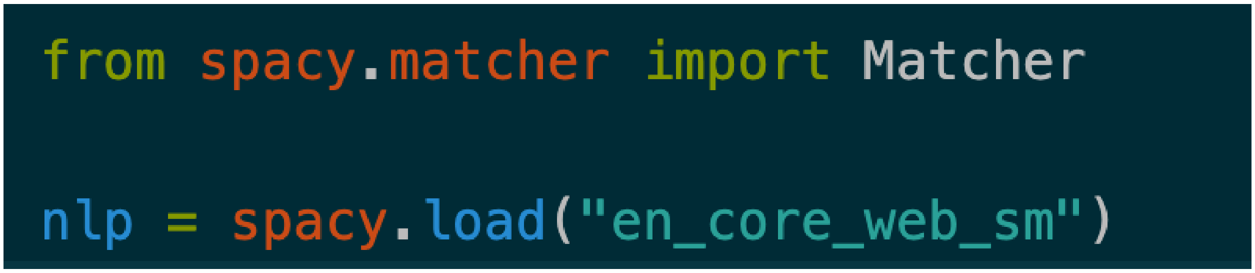
For example, a rule that would find phrases like this found in a resume:

[4] PC: https://spacy.io/usage/linguistic-features

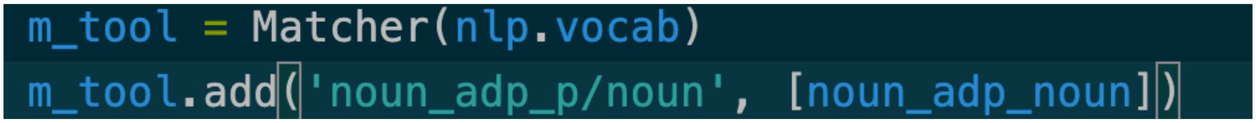


```
wlist = ("Grade", "graduate", "undergraduate", "course", "class", 'late', 'late work', "syllabus",

noun_adp_noun = [
    {"POS": {"IN": ["NOUN", "PROPN"]}, "LOWER": {"NOT_IN": wlist}, "LEMMA": {"NOT_IN": wlist}},
    {"POS": {"IN": ["ADP"]}},
    {"POS": {"IN": ["NOUN","PROPN"]}, "LOWER": {"NOT_IN": wlist}, "LEMMA": {"NOT_IN": wlist}},
    {}
]
```

This rule is then added to a set of rules within an NLP model, loaded by these lines:



```
from spacy.matcher import Matcher

nlp = spacy.load("en_core_web_sm")
```



```
m_tool = Matcher(nlp.vocab)
m_tool.add('noun_adp_p/noun', [noun_adp_noun])
```

The client does not provide rules regarding what constitutes a keyword, it is left to my discretion what the keywords are in the document. The output will be similar to a summary, but it need not be in complete sentences, rather phrases that would be important for understanding the key aspects of a course.

Conclusion

Width.ai will be completing the project over the span of the next month or so. Projects like this are often very big, and require multiple data scientists and machine learning engineering to each complete their parts of the process. Many parts of the project are in motion, and will soon see results and be delivered to its client. Hundreds of rules have been created to this date, and models have been trained to a 97% accuracy in detecting legal documents. While the course syllabi do not have a clear metric in determining its success besides the satisfaction of the clientele, many creative and efficient rules have been created, extracting nearly all the important data within course syllabi.

References

Sable, A. (2021). *Building Custom Deep Learning Based OCR models.* Nanonets.

https://nanonets.com/blog/attention-ocr-for-text-recogntion/

spaCy (2021). *Library Architecture.* spaCy.

https://spacy.io/api

spaCy (2021). *Linguistic Features.* spaCy.

https://spacy.io/usage/linguistic-features

Width.ai (2021). Width.ai LLC.

https://www.width.ai/