



Delivering Successful Data Annotation Projects

Dia Trambitas, PhD, Head of Product

Andrei Feier, MD, PhD, Lead Clinical Annotator

Radu Bisca, NLP Lab Product Manager

Outline

- 1. Introduction to Text Annotation**
- 2. Annotation Projects Setup in NLP Lab**
- 3. Annotation Guidelines**
- 4. Preannotation Resources**
- 5. Model Training**
- 6. Conclusions and further resources**

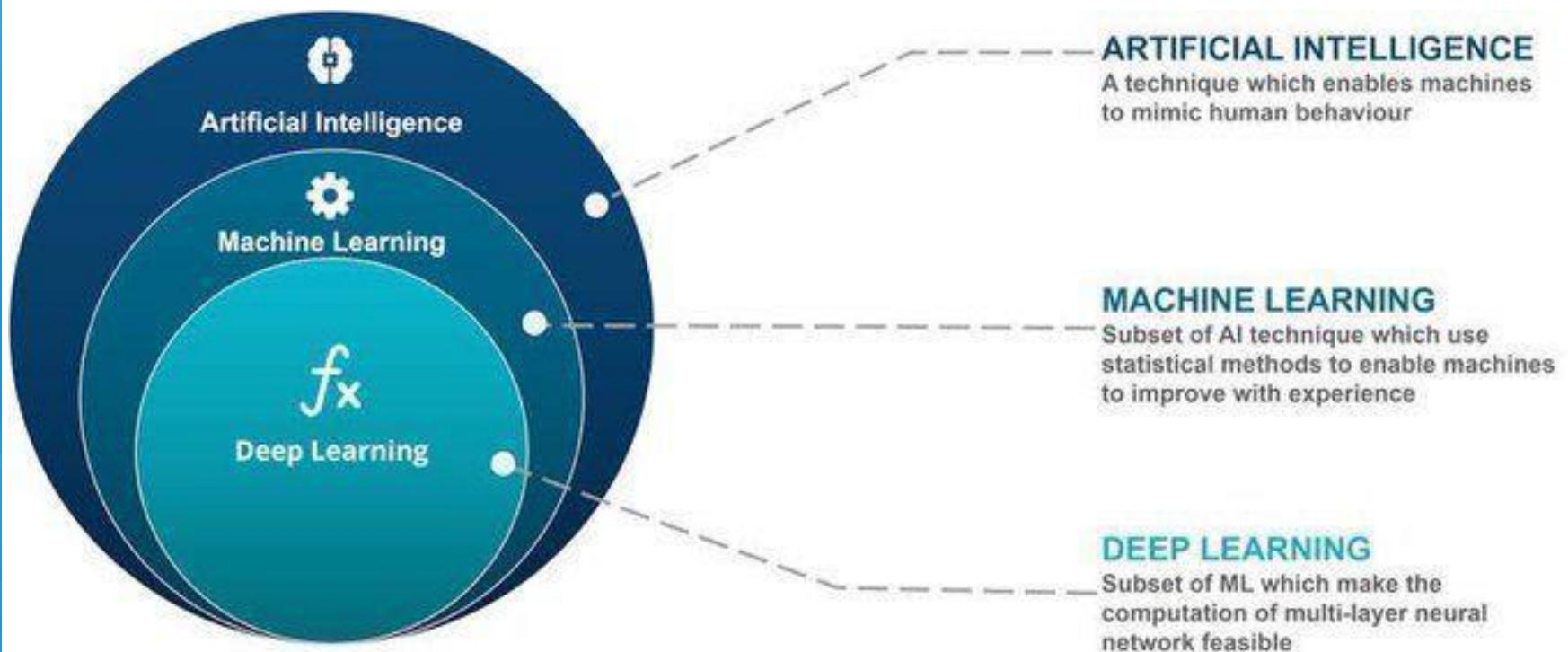
Outline

- 1. Introduction to Text Annotation**
- 2. Manual Annotation**
- 3. Annotation Guidelines**
- 4. Project Setup and Management**
- 5. Preannotations**
- 6. Conclusions and further resources**

Context



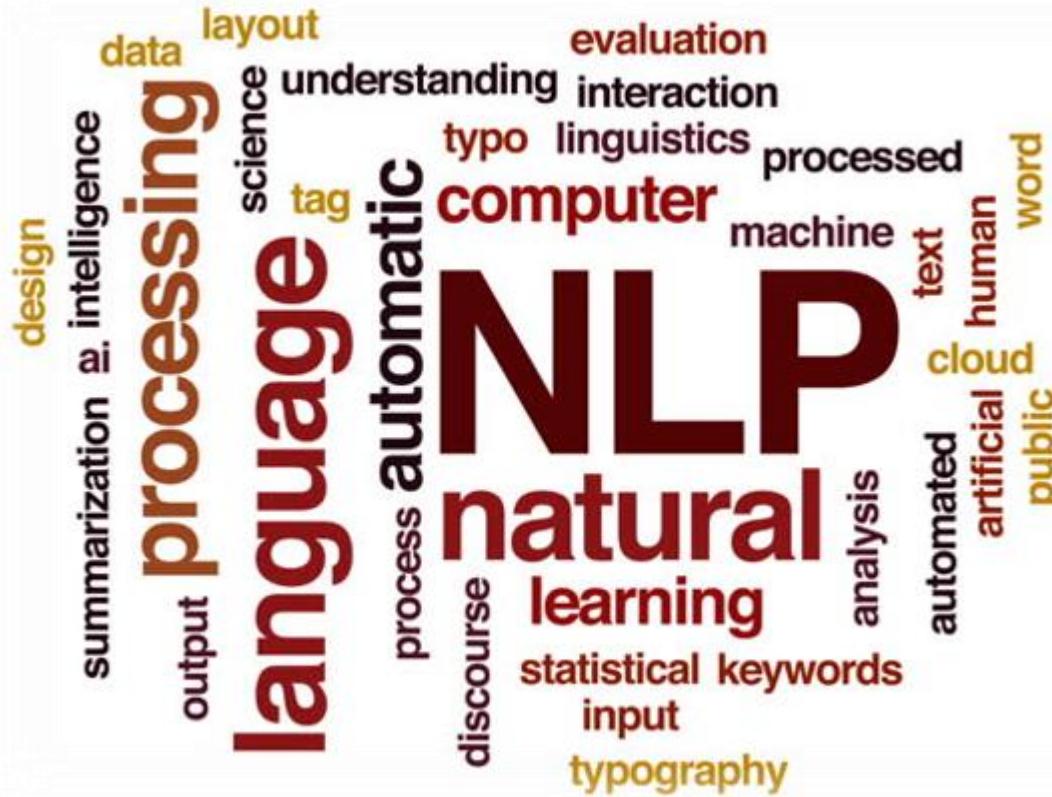
- ML is the study of computer algorithms that improve automatically through experience.
- ML models make decisions based on sample data (training data), without being explicitly programmed.



Natural Language Processing



- NLP is a subfield of linguistics and computer science concerned with the processing and analysis of natural language data.



What is Data Annotation?

- The process of labeling or tagging raw data with meaningful information, making it usable for training machine learning models.

Absent^[Z] Past^[X] Hypothetical^[C] Family^[M] SomeoneElse^[M] Possible^[N] Planned^[B] Allergy^[m]

PAST SURGICAL HISTORY: Section_Header Colon resection Procedure Past in 1990 and sinus surgeries Procedure Past in 1987, 1990 and 2005.

ALLERGIES: Section_Header PENICILLIN Drug Allergy.

SOCIAL HISTORY: Section_Header The patient is married.
She uses no ethanol Substance Absent, no tobacco Substance Absent and no illicit Substance Absent. She has a very support family unit.

FAMILY HISTORY: Section_Header Positive for diabetes mellitus type 2 Diabetes Family in both mother and her sister.

REVIEW OF SYSTEMS: Section_Header The patient currently denies any pain Symptom Absent, denies any headache Symptom Absent or blurred vision Symptom Absent.
Denies chest pain Symptom Absent or shortness of breath Symptom Absent.
She denies any nausea Symptom Absent or vomiting Symptom Absent.
Otherwise, systems are negative.

PLAN: Section_Header Left breast excisional biopsy Procedure Planned with preoperative guidewire localization and intraoperative specimen radiography Test Planned.

Annotation Tasks in NLP



Entity Recognition

40 units **DOSAGE** of
insulin glargine **DRUG**
at night **FREQUENCY**

Assertion Status

Fever and sore throat → PRESENT
No stomach pain → ABSENT
Father with Alzheimer → FAMILY

Text Classification



Emotion Detection



Relation Extraction



Why is Data Annotation Crucial for NLP?

- Transforming unstructured data into **structured** format.
 - Providing "ground truth" for model training and validation.
 - Enhancing the performance of NLP models.
 - Ensuring models understand context, sentiment, entities, etc.
 - A way of transferring annotators knowledge and experience to models.
-
- **The resulting models can only be as good as the training data.**

Role of Annotation Tools

- Facilitate **systematic and organized annotation**.
- Allow for **collaborative annotation** projects.
- Ensure standardization and **consistency**.
- Offer **quality control** and **review** functionalities.
- Provide **automation** and **AI-assisted features**.

Key Requirements for Annotation Tools

- **High productivity** – keep annotators "in the zone"
 - ✓ Intuitive, easy to learn
 - ✓ Keyboard shortcuts
 - ✓ How many clicks? How many eye movements?
 - ✓ Automatic transition between documents
- **Teamwork**
 - ✓ Support for projects, teams, and role-based permissions
 - ✓ Customizable workflows: validation, review, model training, etc.
 - ✓ Collaboration: compare versions, share guidelines
 - ✓ Versioning & audit trails
- **Improve over time**
 - ✓ New features and constant improvements
 - ✓ Learn from annotators actions
 - ✓ High-accuracy, configurable deep learning backend
 - ✓ "Closed-loop" automation: pre-train, annotate, retrain, measure

Let's get started!

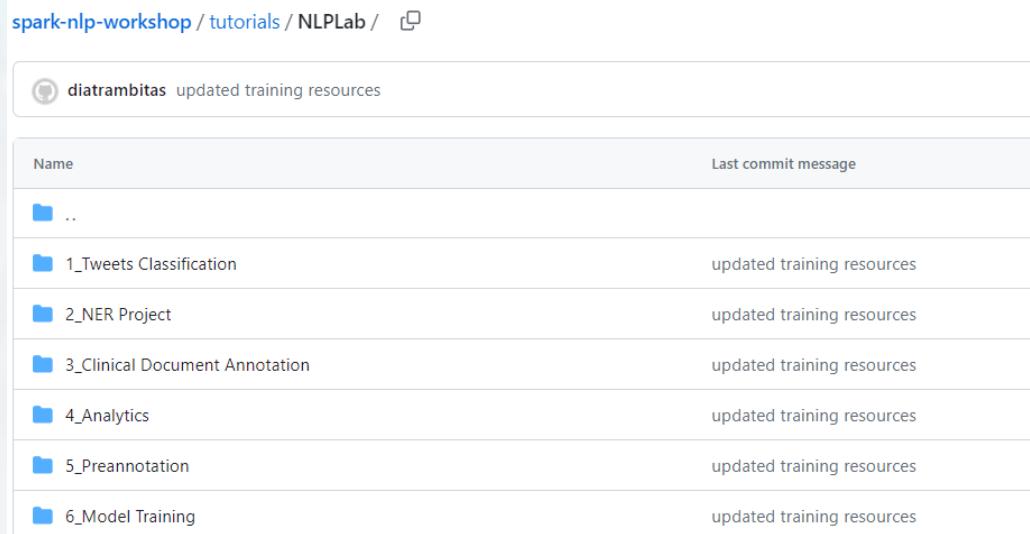
Training Resources

Dedicated **NLP Lab Instance** - <http://3.216.238.207/>

- m4.4xlarge EC2 instance – 16 CPUs and 64GB RAM
- available for 2 weeks
- **Credentials shared via Email** (check also in the spam folder)

Git Hub - <https://github.com/JohnSnowLabs/spark-nlp-workshop>

- Clone the repo on your local machine
- Resources available under **/spark-nlp-workshop/tutorials/NLPLab**



A screenshot of a GitHub repository page titled "spark-nlp-workshop / tutorials / NLPLab". The page shows a list of training resources organized into six folders:

Name	Last commit message
..	
1_Tweets Classification	updated training resources
2_NER Project	updated training resources
3_Clinical Document Annotation	updated training resources
4_Analytics	updated training resources
5_Preannotation	updated training resources
6_Model Training	updated training resources

Outline

1. Introduction to Text Annotation
2. **Annotation Projects Setup in NLP Lab**
3. Annotation Guidelines
4. Preannotation Resources
5. Model Training
6. Conclusions and further resources

Project Setup

1. Project creation steps for Classification project
 - Configuration, choices, hotkeys
2. Project creation steps for NER project
 - Configuration, choices, hotkeys
3. Task import
4. Task assignment
5. Workflow details

Project Creation/Classification

1. Login into NLP Lab
2. Once logged in, in the projects page, select "New Project"



3. In the Project Description page, choose a name for your project and click Next. The name must be unique, you will get a warning in case there is already a project with the name you choose.
4. In the Add team Members page, use the "Search team Members" box to lookup usernames you want added to the project
Note: the users must be created and added to the system prior to assigning them to a project

Click Next. This takes you to the Configuration page. Configuring your project is a guided experience and has 4 steps displayed in the top horizontal bar:

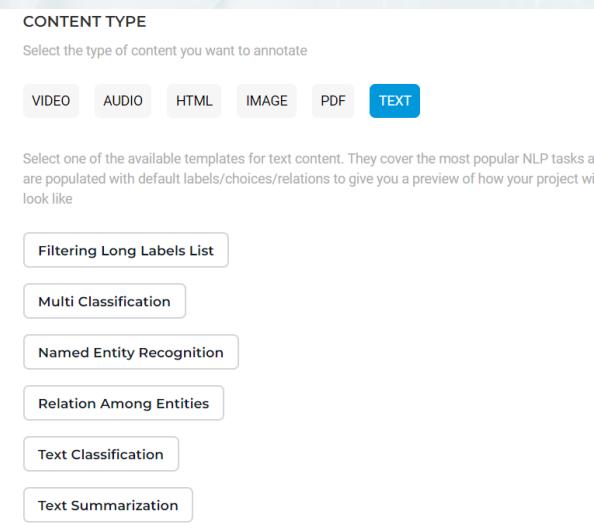


Select Text Classification and click Next.

Project Creation/Classification

For the training session today, we will create 2 different project types : Text Classification and Named Entity Recognition

5. Select Text Classification and click Next.



6. In Step 2 and 3 leave the default settings and click Next.

Project Creation/Classification-cont

7. In Step 4 "Customize Labels", click on Choices. By default, few choices are already populated for you.

LABELS

CHOICES

Choices: Answers (3) ...

Expected Response x Plausible Response x Implausible Response x

+ ADD NEW CHOICE

Add yours then remove existing ones then click the Save Config button. We will be classifying tweets in this project so the choices we add are “positive”, “negative” and “neutral”

Choices: Emotion (4) ...

positive x negative x neutral x

+ ADD NEW CHOICE

This takes us to the Import page where tasks will be imported into the project.

Project Creation/Classification-cont

➤ Customize how you view your choices

By default, the choices are listed vertically. If you prefer to list them horizontally, switch to Code view and add "showInLine" tag set to True

CUSTOMIZE CONFIGURATION

Select from following custom options category

Visual **Code** 

```
1 <View orientation="horizontal" oneClickSubmit="true">
2   <Choices name="Emotion" toName="text" choice="single" showinline| = "true">
3     <Header value="Select Emotion"/>
4     <Choice value="positive"/>
5     <Choice value="negative"/>
6     <Choice value="neutral"/>
7   </Choices>
8   <Text name="text" value="$text"/>
9 </View>
10
```

Select Emotion

positive
 negative
 neutral

Select Emotion positive negative neutral

Project Creation/NER

1. Login into NLP Lab
2. Once logged in,
in the projects page, select "New Project"



3. In the Project Description page, choose a name for your project and click Next. The name must be unique, you will get a warning in case there is already a project with the name you choose.
4. In the Add team Members page, use the "Search team Members" box to lookup usernames you want added to the project

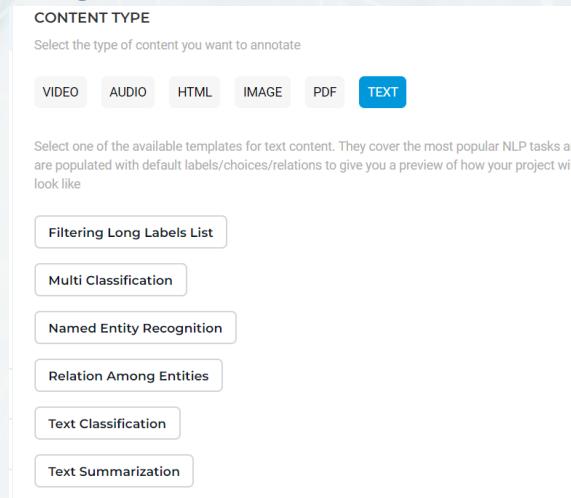
Note: the users must be created and added to the system prior to assigning them to a project

Click Next. This takes you to the Configuration page. Configuring your project is a guided experience and has 4 steps displayed in the top horizontal bar:

Project Creation/NER

For the training session today, we will create 2 different project types : Text Classification and Named Entity Recognition

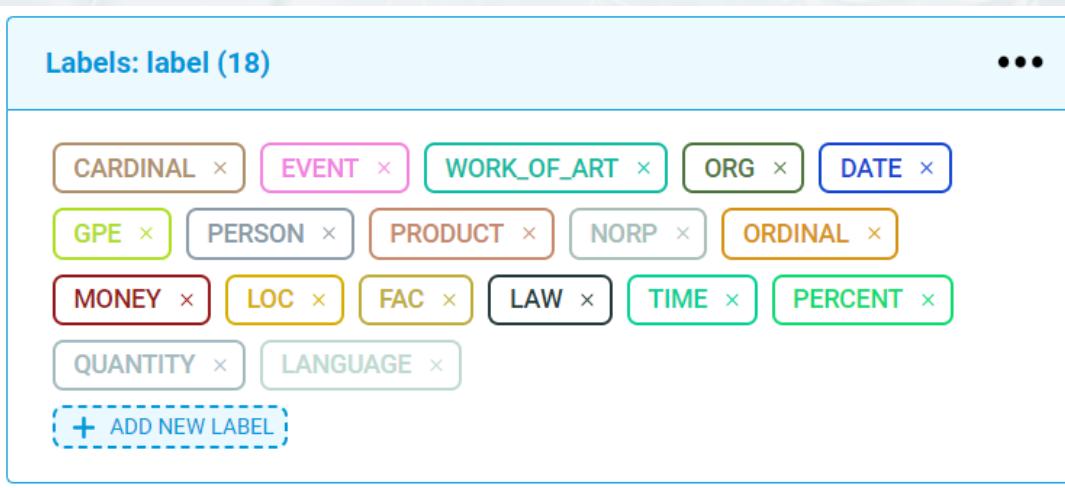
5. Select Named Entity Recognition and click Next.



6. In Step 2 and 3 leave the default settings and click Next.

Project Creation/NER-cont

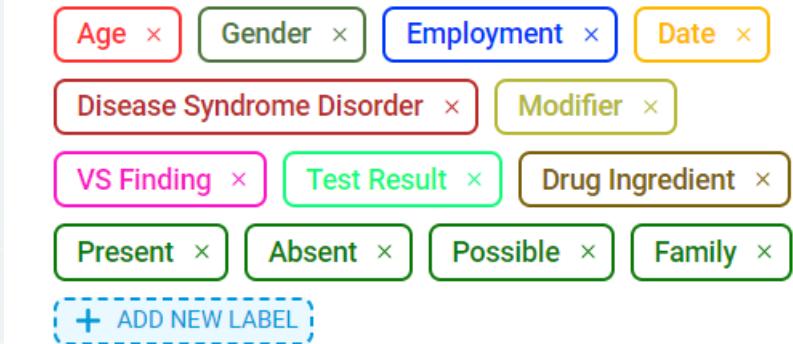
7. In Step 4 "Customize Labels", click on Labels.
By default, few labels are already populated for you.



Add yours then remove existing ones then click the Save Config button.
This takes us to the Import page where tasks will be imported into the project.

Project Creation/NER-cont

- Tip: Remove unwanted labels by editing the xml config file. You can do this by switching modes between Visual to Code



```
<View orientation="horizontal">
  <Labels name="label" toName="text">
    <Label value="Age" background="#ff3333"/>
    <Label value="Gender" background="#4c753d"/>
    <Label value="Employment" background="#0036f8"/>
    <Label value="Date" background="#ffb800"/>
    <Label value="Disease Syndrome Disorder" background="#ba2b2b"/>
    <Label value="Modifier" background="#b4b731"/>
    <Label value="VS Finding" background="#fb16c8"/>
    <Label value="Test Result" background="#13fa7c"/>
    <Label value="Drug Ingredient" background="#7a6109"/>
    <Label value="Present" background="#0b7a09" assertion="true"/>
    <Label value="Absent" background="#0b7a09" assertion="true"/>
    <Label value="Possible" background="#0b7a09" assertion="true"/>
    <Label value="Family" background="#0b7a09" assertion="true"/>
  </Labels>
  <Text name="text" value="$text"/>
</View>
```

Project Creation/NER-cont

- Creating hot keys.

Hot keys are keyboard shortcuts you can set for increasing productivity in the labeling process.
To setup hot keys, switch to Code mode and add hotkey tag:

```
1 <View orientation="horizontal">
2   <Labels name="label" toName="text">
3     <Label value="Age" background="#ff3333" hotkey="A"/>
4     <Label value="Gender" background="#4c753d" hotkey="G"/>
5     <Label value="Employment" background="#0036f8" hotkey="E"/>
6     <Label value="Date" background="#ffb800"/>
7     <Label value="Disease Syndrome Disorder" background="#ba2b2b"/>
8     <Label value="Modifier" background="#b4b731"/>
9     <Label value="VS Finding" background="#fb16c8"/>
10    <Label value="Test Result" background="#13fa7c"/>
11    <Label value="Drug Ingredient" background="#7a6109"/>
12    <Label value="Present" background="#0b7a09" assertion="true"/>
13    <Label value="Absent" background="#0b7a09" assertion="true"/>
14    <Label value="Possible" background="#0b7a09" assertion="true"/>
15    <Label value="Family" background="#0b7a09" assertion="true"/>
16  </Labels>
17  <Text name="text" value="$text"/>
18 </View>
```

Project Creation/NER-cont

➤ Assertion Labels

Assertion labels are tags with contextual information that are assigned to NER extractions. The typical example of assertion status detection is negation identification: in the sentence “the patient has no history of diabetes”, the word “diabetes” is annotated using both an NER label (Disease) and an assertion label (Absent).

These labels are defined using the configuration file as a “regular” label + assertion=“true”

```
2   <Label value="Present" background="#0b7a09" assertion="true"/>
3   <Label value="Absent" background="#0b7a09" assertion="true"/>
4   <Label value="Possible" background="#0b7a09" assertion="true"/>
5   <Label value="Family" background="#0b7a09" assertion="true"/>
```

Using the Annotation Lab, you can annotate assertion in a very simple way. First, choose a NER label and select the part of the text that you want to extract, and then choose an assertion label and select that same part of the document.

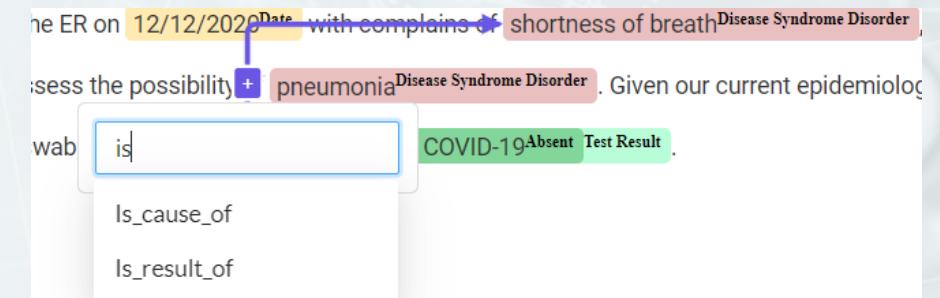
Read more about assertion labels here - <https://www.johnsnowlabs.com/tips-and-tricks-on-how-to-annotate-assertion-in-clinical-texts/>

Project Creation/NER-cont

➤ Creating relations

Relations can be defined using the configuration file by adding a Relation section as depicted in the screenshot below.

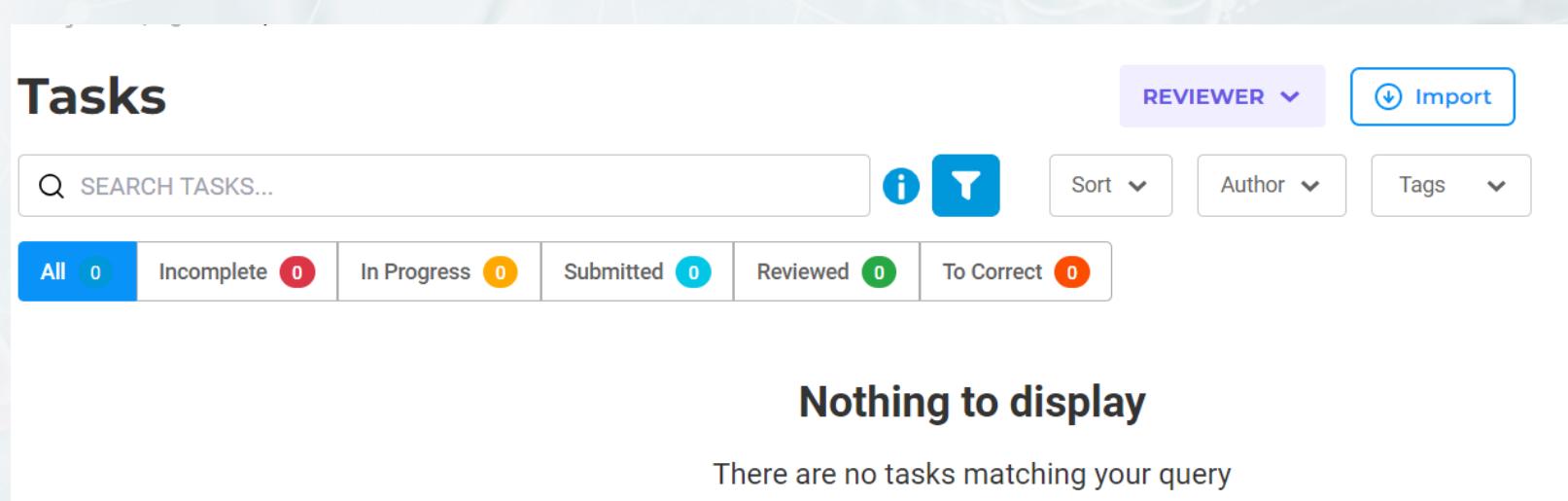
```
5 </Labels>
6
7 <Relations>
8     <Relation value="Is_cause_of"/>
9     <Relation value="Is_result_of"/>
10 </Relations>
```



When creating the relation, the name defined in the configuration file is displayed in the UI.

Task Import and Assignment

When project setup is done, you will land on the Task page. The page is empty, and to import your tasks, click on the Import button – top right.



The screenshot shows the 'Tasks' page interface. At the top, there is a search bar labeled 'SEARCH TASKS...' with a magnifying glass icon. To its right are two buttons: a purple 'REVIEWER' dropdown and a blue 'Import' button with a downward arrow icon. Below the search bar are several filter options: 'Sort' (dropdown), 'Author' (dropdown), and 'Tags' (dropdown). A row of status filters shows counts for 'All' (0), 'Incomplete' (0), 'In Progress' (0), 'Submitted' (0), 'Reviewed' (0), and 'To Correct' (0). The main content area displays the message 'Nothing to display' in bold, followed by the sub-message 'There are no tasks matching your query'.

Task Import and Assignment

There are multiple ways for importing tasks,: using a zip archive with your text files, importing tasks using a csv file, using json file, etc

Import Generate Synthetic Tasks

SUPPORTED IMPORT FORMAT OPTIONS

BULK IMPORT - JSON, CSV, TSV
Multiple tasks will be generated in one step: one for each JSON element or one for each CSV or TSV row.

Click on following buttons to download example import templates.

JSON **CSV** **TSV** **ZIP** **RAR**

SINGLE TASK IMPORT - TXT, JPEG, PDF
Create one task for each imported TXT, JPEG, PDF file.

SAMPLE TASK
Add a sample task to your project to check how annotation works.

ADD SAMPLE TASK

IMPORTING RESOURCE FILES

Text Source
Import JSON/CSV/TSV/TXT/PDF files following the above illustrated templates.

Images, Audio, Video Files
Images, audio, video and other files may be hosted on external servers with http/https access. In this case, your JSON/CSV/TSV/TXT/PDF import file should contain http/https URLs pointing to those resources.

Overwrite completions/predictions
When checked, if the import file includes tasks already available in the project (tasks with same id and same text content) those will overwrite the existing tasks.

OCR Document **OCR Server: --Select--**
When checked, the text included in each imported PDF/Image file will be automatically extracted and used for creating one task.
This option is only available when OCR is enabled

Drag and drop your files here or click for import
JSON, CSV, TSV,TXT, Zip, RAR, PDF, Audio, Video, Images

OR paste a file URL or task in JSON format here
`http://example.com/tasks.json OR {"image": "http://my.image.jpg"}` **Import**

OR upload from S3
Path to s3 folder to use for import, e.g. "s3://bucket/folder"

S3 Access Key **S3 Secret Key** **Session Token (Required for MFA Account)**
S3 Access Key e.g. ***** Session Token (Required for MFA Account)
***** **Import**

Task Import and Assignment

Download any sample file by clicking on the specific icon.

Click on following buttons to download example import templates.

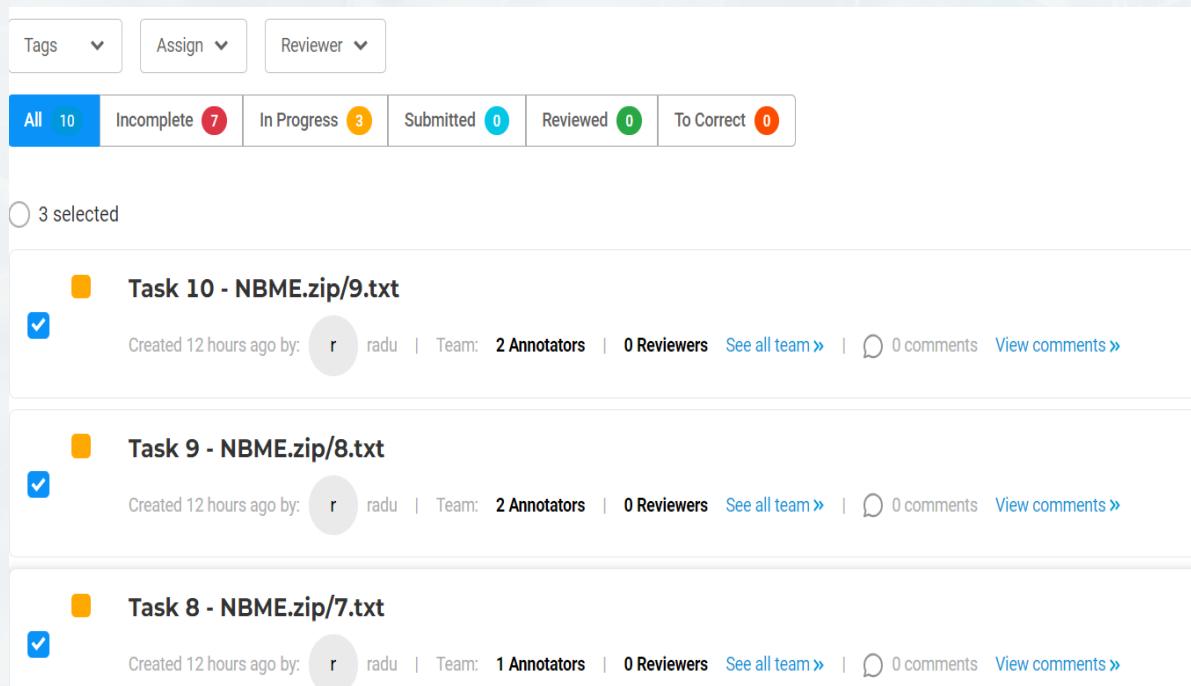
JSON CSV TSV ZIP RAR

The screenshot below shows a sample file used to import tweets in the classification project used as part of this training sessionCSV file used for this demo:

text	title
I'd have responded, if I were going	t1
Sooo SAD I will miss you here in San Diego!!!	t2
my boss is bullying me...	t3
what interview! leave me alone	t4
Sons of ****, why couldn't they put them on the releases we already bought	t5

Task Import and Assignment

Once you import the tasks, choose to Explore tasks; this takes you to the Tasks page. As a project owner you will start assigning tasks to Annotators. select one or more tasks, click on Assign drop down and select the name of the user to assign selected tasks. Repeat this step to continue assigning tasks.

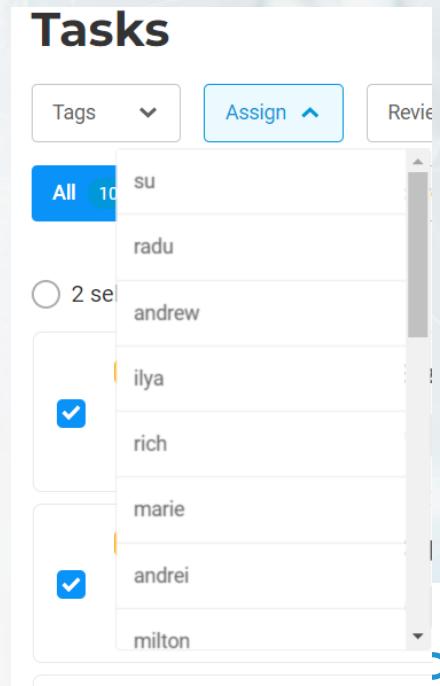


Tags Reviewer

All 10 Incomplete 7 In Progress 3 Submitted 0 Reviewed 0 To Correct 0

3 selected

- Task 10 - NBME.zip/9.txt
Created 12 hours ago by:  radu | Team: 2 Annotators | 0 Reviewers See all team » | 0 comments View comments »
- Task 9 - NBME.zip/8.txt
Created 12 hours ago by:  radu | Team: 2 Annotators | 0 Reviewers See all team » | 0 comments View comments »
- Task 8 - NBME.zip/7.txt
Created 12 hours ago by:  radu | Team: 1 Annotators | 0 Reviewers See all team » | 0 comments View comments »



Tags Reviewer

All 10

- su
- radu
- andrew
- ilya
- rich
- marie
- andrei
- milton

Analytics of NLP Lab - Tasks



Analytics

Tasks

Team Productivity

Last Updated: 37 minutes ago
Updated by: andrei

Total tasks

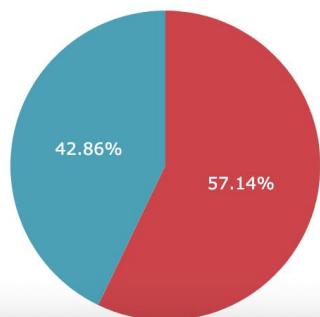
7

Inter-Announcer Agreement

Total tasks in last 30 days

7

Tasks by status



- Incomplete
- Submitted
- In Progress
- Reviewed

Tasks Created By

andrei

100%

- different charts
- overall annotation progress

Click [here](#) for in-depth explanation

Analytics of NLP Lab - Productivity



Analytics

Last Updated: 37 minutes ago
Updated by: andrei

Tasks Team Productivity Inter-Annotator Agreement

Total completions **18**

Completions by Status

A pie chart divided into two equal halves, orange and blue, each labeled 50%. A legend indicates: Starred Completions (blue) and Draft Completions (orange).

Average number of draft completion per task: 3.0

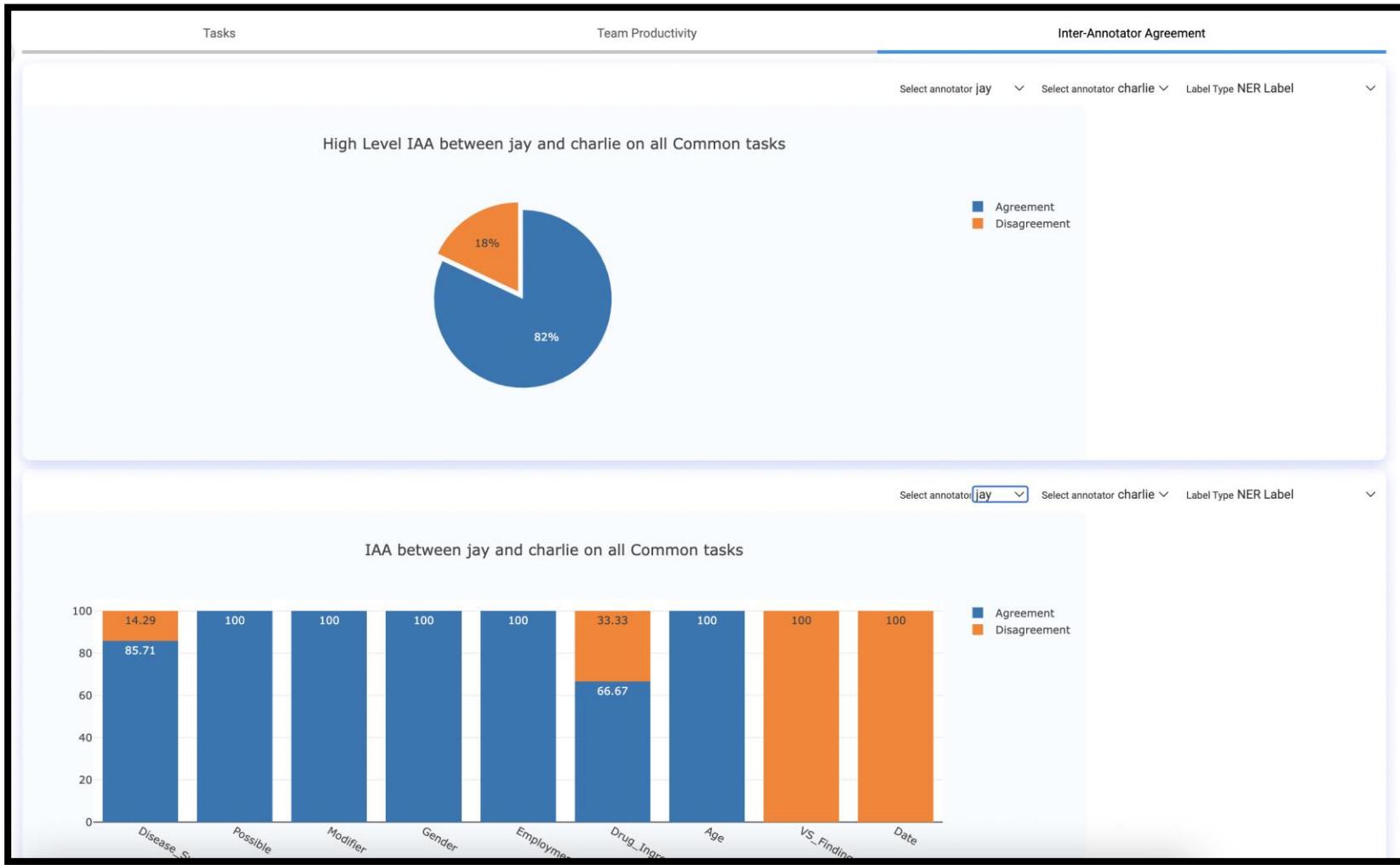
Time Period

Completions By Annotator

- different tools
- productivity across team

Click [here](#) for in-depth explanation

Analytics of NLP Lab – IAA



- level of agreement
- tools for assessing IAA

Click [here](#) for in-depth explanation

Outline

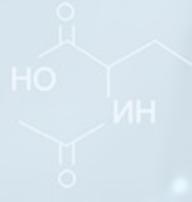
1. Introduction to Text Annotation
2. Annotation Projects Setup in NLP Lab
- 3. Annotation Guidelines**
4. Preannotation Resources
5. Model Training
6. Conclusions and further resources



Annotation Guidelines (AG)

Andrei Feier, MD, PhD

Clinical Lead Annotator



Contents

- 1. Stakeholders and Roles**
- 2. Structure and Content of AG**
- 3. Best Practices**
- 4. Annotation Errors**
- 5. Project Analysis**



Stakeholders and Roles



Stakeholders and Roles



Annotator Lead

- oversees the entire project - > guidelines align with the project's objectives
- keeps track of changes with a Change Log
- manager role in projects

Annotators

- usually medical professionals (MDs, nurses etc.)
- assess/curate data/annotate
- annotator/reviewer role in projects

Stakeholders and Roles



Data Scientist

- technical stakeholders who will use the annotated data for training
- provide requirements for annotation granularity and specificity

Quality Assurance (Reviewers)

- ensure the quality and consistency of the annotations and AGs
- usually, an experienced annotator from the team is assigned as reviewer



Structure and Content of AG

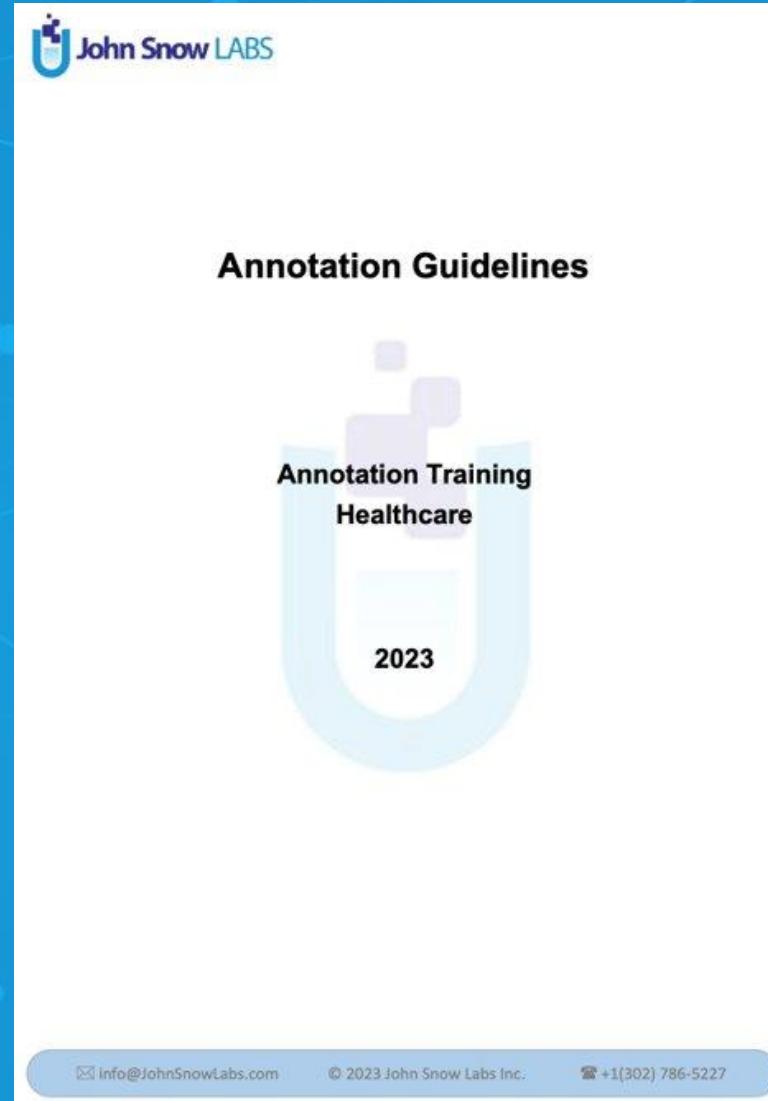


Contents

1. Cover Page
2. Table of Contents
3. Introduction
4. Taxonomy
5. NER Labels
6. Assertion Labels
7. Relations
8. Text Classification
9. Appendix
10. Changelog

Contents

1. [Cover Page](#)
2. Table of Contents
3. Introduction
4. Taxonomy
5. NER Labels
6. Assertion Labels
7. Relations
8. Text Classification
9. Appendix
10. Changelog



Contents

1. Cover Page
2. **Table of Contents**
3. Introduction
4. Taxonomy
5. NER Labels
6. Assertion Labels
7. Relations
8. Text Classification
9. Appendix
10. Changelog

	John Snow LABS
Content	
Introduction.....	3
Taxonomy	3
Entity Labels	4
Age	4
Gender.....	4
Employment.....	5
Date.....	5
Disease_Syndrome_Disorder.....	6
Modifier.....	6
VS_Finding	7
Test_Result.....	7
Drug_Ingredient.....	7
Assertion Labels	9
Present	9
Absent	9
Possible	10
Family	10
Assertion Table.....	11
Relations	12
Is_diagnosis_date_of	12
Is_modifier_of.....	12
Is_cause_of	13
Is_result_of.....	13
Relation Table.....	14
Text Classification	15
Gender.....	15
Type of Text.....	15

Contents

1. Cover Page
2. Table of Contents
3. Introduction
4. Taxonomy
5. NER Labels
6. Assertion Labels
7. Relations
8. Text Classification
9. Appendix
10. Changelog



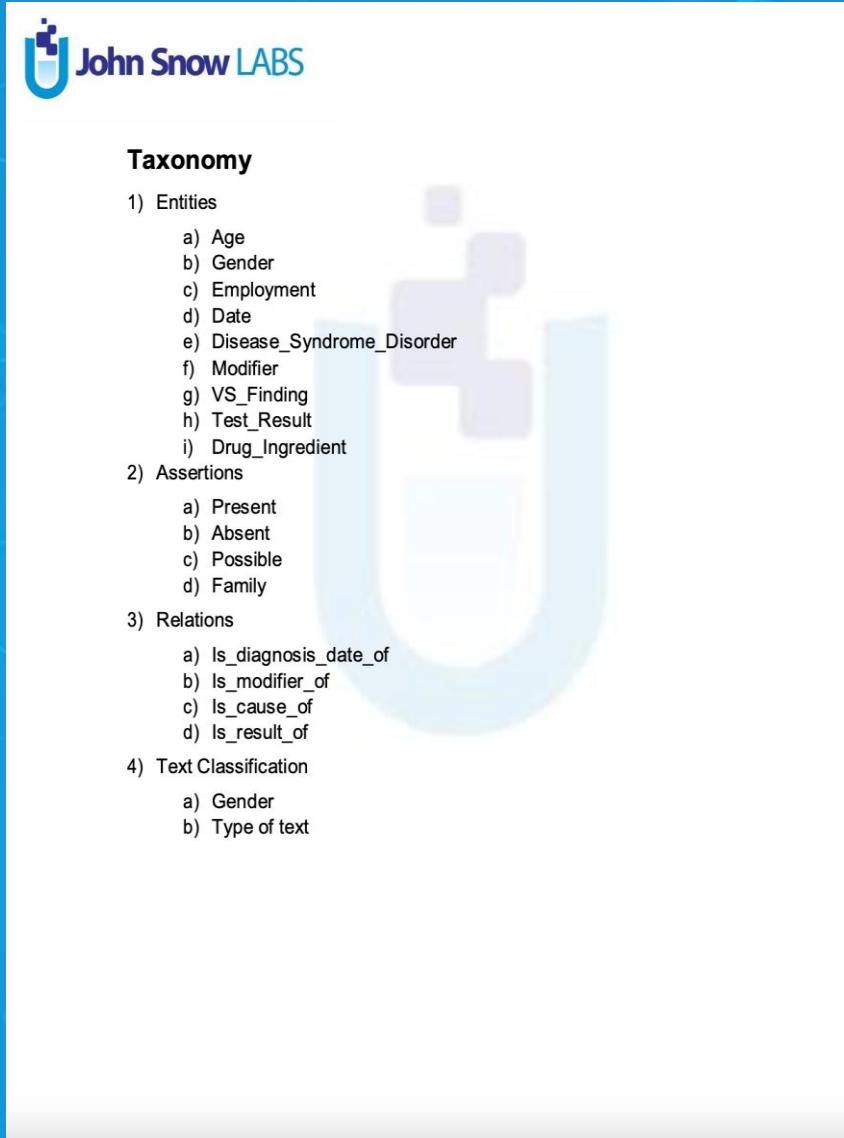
Introduction

The goal of annotations is to train Natural Language Processing (NLP) models to extract relevant information from different kinds of documents.

Annotation Guidelines (AG) define the entities to be extracted and the way they have to be extracted, with the aim of assuring consistency between annotators. Extraction of entities, relations and assertions will be explained in these AG.

Contents

1. Cover Page
2. Table of Contents
3. Introduction
4. **Taxonomy**
5. NER Labels
6. Assertion Labels
7. Relations
8. Text Classification
9. Appendix
10. Changelog



The screenshot shows a slide titled "Taxonomy" from the John Snow LABS presentation. The slide features a blue header with the "John Snow LABS" logo. Below the title, there is a large, semi-transparent watermark of a stylized human figure composed of interconnected nodes and lines. The main content is organized into four numbered sections: 1) Entities, 2) Assertions, 3) Relations, and 4) Text Classification. Each section contains a list of specific labels or types.

Taxonomy	
1)	Entities
a)	Age
b)	Gender
c)	Employment
d)	Date
e)	Disease_Syndrome_Disorder
f)	Modifier
g)	VS_Finding
h)	Test_Result
i)	Drug_Ingredient
2)	Assertions
a)	Present
b)	Absent
c)	Possible
d)	Family
3)	Relations
a)	Is_diagnosis_date_of
b)	Is_modifier_of
c)	Is_cause_of
d)	Is_result_of
4)	Text Classification
a)	Gender
b)	Type of text

Contents

1. Cover Page
2. Table of Contents
3. Introduction
4. Taxonomy
5. **NER Labels**
6. Assertion Labels
7. Relations
8. Text Classification
9. Appendix
10. Changelog

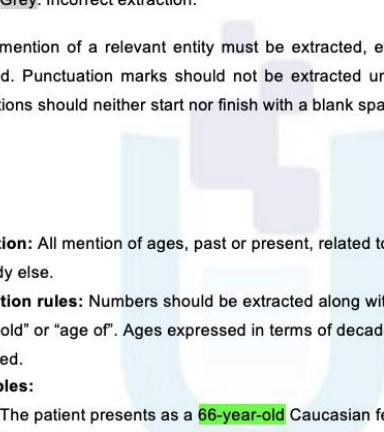
 John Snow LABS

Entity Labels

For each kind of entity, these AG provide a definition, specific extraction rules, examples, relations with other variables and assertion labels. Examples only include the annotations that are relevant for the kind of entity that is being explained. The following color reference is used:

- **Green**: Correct extraction.
- **Grey**: Incorrect extraction.

Every mention of a relevant entity must be extracted, even if it is repeated or | negated. Punctuation marks should not be extracted unless stated otherwise. Extractions should neither start nor finish with a blank space.



Age

Definition: All mention of ages, past or present, related to the patient or with anybody else.

Extraction rules: Numbers should be extracted along with expressions such as "years old" or "age of". Ages expressed in terms of decades should also be extracted.

Examples:

- a) The patient presents as a **66-year-old** Caucasian female in stable health.
- b) The patient was diagnosed in his **50s**.

Assertion labels : None.

Relations : None.

Gender

Definition: Gender-specific nouns, excluding family members.

Contents

1. Cover Page
2. Table of Contents
3. Introduction
4. Taxonomy
5. NER Labels
6. Assertion Labels
7. Relations
8. Text Classification
9. Appendix
10. Changelog

 John Snow LABS

Assertion Labels

Assertion labels are used to indicate an attribute of an entity. The assertion label is placed on top of the entity label, example, Entity **Assertion**.

The following are considerations when adding assertions:

- Entities should only be assigned one assertion label only.
- For the annotation of Assertion, it should be considered only the information found in the sentence that includes the asserted entity.

Not all the combinations of entity and assertion are possible. A table of all the entities and possible assertions is included at the end of this section.

Present

Definition: Entities referring to the patient that are currently present and not negated.

Extraction rules: Use this assertion label only for entities extracted as Disease_Syndrome_Disorder.

Example:

a) He is a 60 years old gentleman with **diabetes** **Present Disease_Syndrome_Disorder**. (Disease_Syndrome_Disorder + Present Assertion).

Absent

Definition: Label added to negated entities.

Extraction rules: Absent entities are found in phrases that include words such as *no*, *without*, *lack*, etc.

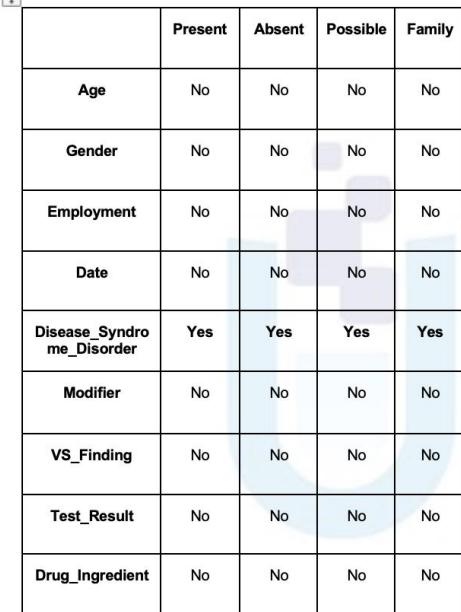
Examples:

a) The ultrasound showed that the patient does not have **psoriasis** **Absent Disease_Syndrome_Disorder**. (Disease_Syndrome_Disorder Entity + Absent Assertion).

b) She is neither **Diabetic** **Absent Disease_Syndrome_Disorder** nor diagnosed with **Obesity** **Absent Disease_Syndrome_Disorder**. (Disease_Syndrome_Disorder Entity + Absent Assertion).

 John Snow LABS

Assertion Table

+ 

	Present	Absent	Possible	Family
Age	No	No	No	No
Gender	No	No	No	No
Employment	No	No	No	No
Date	No	No	No	No
Disease_Syndrome_Disorder	Yes	Yes	Yes	Yes
Modifier	No	No	No	No
VS_Finding	No	No	No	No
Test_Result	No	No	No	No
Drug_Ingredient	No	No	No	No

✉ info@JohnSnowLabs.com © 2023 John Snow Labs Inc. ☎ +1(302) 786-5227

Contents

1. Cover Page
2. Table of Contents
3. Introduction
4. Taxonomy
5. NER Labels
6. Assertion Labels
7. **Relations**
8. Text Classification
9. Appendix
10. Changelog

 John Snow LABS

Relations

Relations are used to link two related entities. To create relations between entities, use the Create Relation button of the annotation tool. Relations are NOT created for entities present in different sentences, or are 2 or more sentences apart. Some relations require a relation label that is found in the relation section of the annotation tool. Also, some relations require assignation of direction that is represented by an arrow in the relation section of the annotator tool. A table with all the possible relations is included at the end of this section.

Is_diagnosis_date_of

Definition: This relation is used to associate a Disease_Syndrome_Disorder entity and a Date entity.

Extraction rules: The Disease_Syndrome_Disorder entity and the relevant date associated with it are extracted and related using the relation label **is_diagnosis_date_of** only when the date refers to the diagnosis of the medical problem.

Examples:

a) She was diagnosed with **hypertension** in **1987**. **Hypertension** | (Disease_Syndrome_Disorder entity) and **1987** (Date entity) are related with **is_diagnosis_date_of** label.

Is_modifier_of

Definition: This relation is used to associate a Disease_Syndrome_Disorder entity and a Modifier.

Extraction rules: The Disease_Syndrome_Disorder and the relevant modifier are extracted and related using the relation label **is_modifier_of**.

Examples:

a) He has been experiencing **chronic migraine** for five years. **Migraine** (Disease_Syndrome_Disorder entity) and **Chronic** (Modifier entity) are related with **is_modifier_of** label.

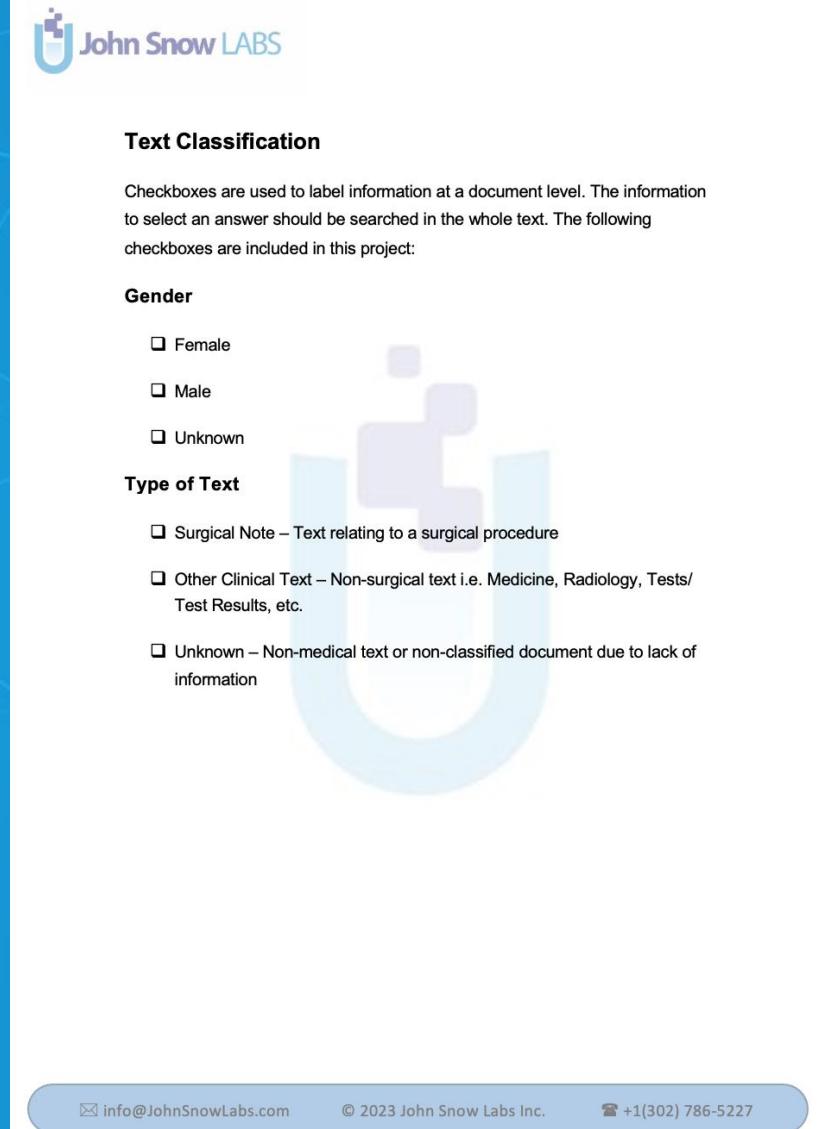
 John Snow LABS

Relation Table

	Entity 1	Entity 2	Label Needed	Direction Needed
Is_diagnosis_date_of	Disease_Syndrome_Disorder	Date	Yes	No
Is_modifier_of	Disease_Syndrome_Disorder	Modifier	Yes	No
Is_cause_of	Disease_Syndrome_Disorder	Disease_Syndrome_Disorder	Yes	Yes
Is_result_of	VS_Finding	Test_Result	Yes	No

Contents

1. Cover Page
2. Table of Contents
3. Introduction
4. Taxonomy
5. NER Labels
6. Assertion Labels
7. Relations
8. **Text Classification**
9. Appendix
10. Changelog



The screenshot shows a page from the John Snow LABS documentation titled "Text Classification". The page includes the company logo at the top left. Below the title, there is a descriptive paragraph about checkboxes for document-level labeling. Under the heading "Gender", there are three checkboxes: "Female", "Male", and "Unknown". To the right of these checkboxes is a graphic of three stylized human figures in light blue, purple, and yellow. Under the heading "Type of Text", there are three checkboxes: "Surgical Note – Text relating to a surgical procedure", "Other Clinical Text – Non-surgical text i.e. Medicine, Radiology, Tests/ Test Results, etc.", and "Unknown – Non-medical text or non-classified document due to lack of information". At the bottom of the page is a footer bar containing links for email, copyright information, and a phone number.

info@JohnSnowLabs.com

© 2023 John Snow Labs Inc.

+1(302) 786-5227

Contents

1. Cover Page
2. Table of Contents
3. Introduction
4. Taxonomy
5. NER Labels
6. Assertion Labels
7. Relations
8. Text Classification
9. Appendix
10. Changelog



Change Log



Version	Revision Date	Revision Description	Responsible for AG Updates
1.0	June 23 rd , 2023	Updates to implement following the consensus meeting – 22.07.2023	@Annotator 1
2.0	June 27 th , 2023	Updates to implement following the consensus meeting – 27.07.2023	@Annotator 2

Examples #1

Age

Definition: All mention of ages, past or present, related to the patient or with anybody else.

Extraction rules: Numbers should be extracted along with expressions such as “years old” or “age of”. Ages expressed in terms of decades should also be extracted.

Examples:

- a) The patient presents as a **66-year-old** Caucasian female in stable health.
- b) The patient was diagnosed in his **50s**.
- c) John is a **65** years old male.
- d) He was diagnosed with obesity at the age of **25**.

Assertion labels : None.

Relations : None.

Examples #2

Employment

Definition: Mentions of jobs or occupations included in the text.

Extraction rules: Extract terms that are related to any specific jobs or employment, whether related to the patient or not. Do not extract words such as "works", "working" or "employed".

Examples:

- a) She is an office manager for a gravel company.
- b) She will also see a nutritionist and a social worker.
- c) He works as a financial officer.

Assertions: None.

Relations: None.

Examples #3

Disease_Syndrome_Disorder

Definition: Extract all the diseases, syndromes and any relevant condition mentioned in the document.

Extraction rules: Extract all mentions of medical conditions and diseases, including those related to the patient or to a family member. Do not include in the extraction modifiers such as “chronic”, “mild” or “severe” (this kind of words should be extracted using the label Modifier).

Examples:

- a) The patient has **Alzheimer** diagnosed back in 2012.
- b) He was diagnosed with **colon cancer**.
- c) A diagnosis of **chronic kidney disease** was established in the past.
- d) He was diagnosed twice with **chronic depression**.

Modifier

Definition: Terms that modify the medical problem.

Extraction rules: Extract words that indicate severity (such as “mild” or “severe”), duration (such as “chronic” or “acute”) or any other feature of the entities.

Examples:

- a) He has been experiencing **chronic** back pain for five years.
- b) Patient with history of **recurrent** angina.



Best Practices

Best Practices - Annotation Rules



Centralized

one document should include all the rules
(the AGs)



Consistent

contradiction should be avoided



Specific

ambiguity should be avoided



Cross-check Annotations

Use multiple annotators for the same data
("seed corpus")



Clear Objectives

objectives will shape your annotation
guidelines



Explicit

i.e written (even if in detail!)

Best Practices - AG Development



Define goals clearly

What information is needed from texts?

e.g.: "identifying all gender-related words
(NER)"

vs.

"classifying the document based on gender"
(Text Classification +/- NER)



Communication is key

Define communication channels (Slack,
Teams etc.)

Keep track of decisions made (Change Log)

Encourage annotators to avoid private
messaging



Iterative process

Test your AGs

Seed Corpus annotations (pilot)

data scientists/NLP engineers

Best Practices - NER



clinical problem of persistent coughing Clinical_Problem

VS

clinical problem of persistent Modifier coughing Symptom

Define the level of **granularity** of your taxonomy

e.g.: **Clinical_Problem** label vs. **Modifier + Symptom**

Treatment will involve the drug Tarceva Cancer_Treatment

VS

Treatment will involve the drug Tarceva Drug

Each label should have **clear boundaries** (avoid overlapping)

e.g.: "Treatment will involve the drug Tarceva"

(**Cancer_Treatment** vs **Drug**)

Best Practices - NER

concerning evidence of metastasis to the liver^{Metastasis}

a primary symptom indicative of a potential respiratory issue^{Possible Clinical_Problem}

vs

diagnosed with lung disease, specifically non-s⁺ small cell lung can
concerning evidence of metastasis^{Metastasis} to the liver^{Site}

Gender

Female^L Male^{L 1.00} Unknown^L

Extractions should not include more than 2-3 words!

Combine NER with other NLP features

- Consider merging/dropping/splitting entities

assertion labels

e.g.: “*There is metastasis in the lung*” (*Metastasis* >
Metastasis + Site)

relations

text classification

Best Practices - Assertion

Keep Assertion Taxonomy simple!

the patient was diagnosed with lung disease^{Present} Disease

vs

the patient was diagnosed with lung disease^{Disease}

medical history reveals an allergy to penicillin^{Allergen} Drug

- Assign a default assertion status

e.g.: **"The patient has cancer". (Disease + Present assertion label by Default)**

How to decide between multiple possible assertion labels

e.g.: **"No family history of cancer". (Disease + Absent/Family?)**

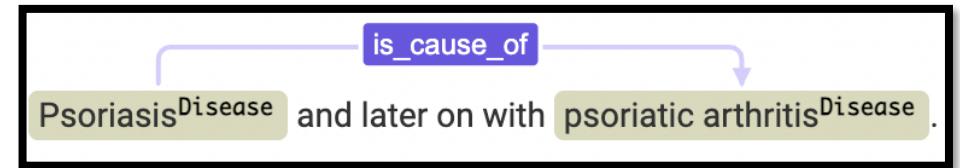
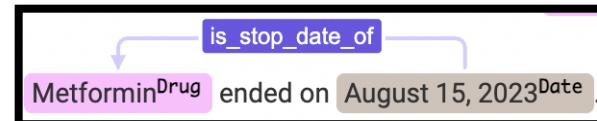
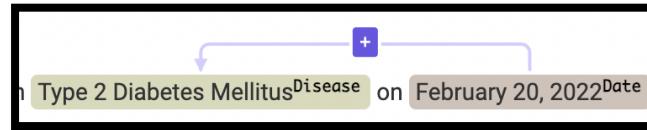
Create your own assertion labels

e.g.: **"medical history reveals an allergy to penicillin" (Drug NER label + Allergen assertion label)**

- Define each assertion label

Best Practices - Relations

Define entities that can be related!



- Add Relation Labels Names only when there are different ways of relating two entities

Disease & Date (1 relation): no need to add a Label
is_date_of_diagnosis

- Drug & Date (3 relations): label needed*
(is_start_date_of, is_stop_date_of, is_generic_date_of)

Add direction only if necessary

is_cause_of (Disease & Disease) needs direction to differentiate cause and consequence

- Define the **distance** between entities to use relation label



Frequent Annotation Errors

Cause of Error

1. Accidental Human Error

Example: "female" extracted as Age

2. Wrong Extraction

Example: "young" extracted as Age (if AG specify that only numeric values should be extracted).

3. Normal Disagreement

Example: different granularity to extract "late 40s" as Age.

Kinds of NER Errors

1. Missing Extraction

“The patient is 30 years old” [0]

2. Wrong Label

“The patient is 30 years old” [Gender]

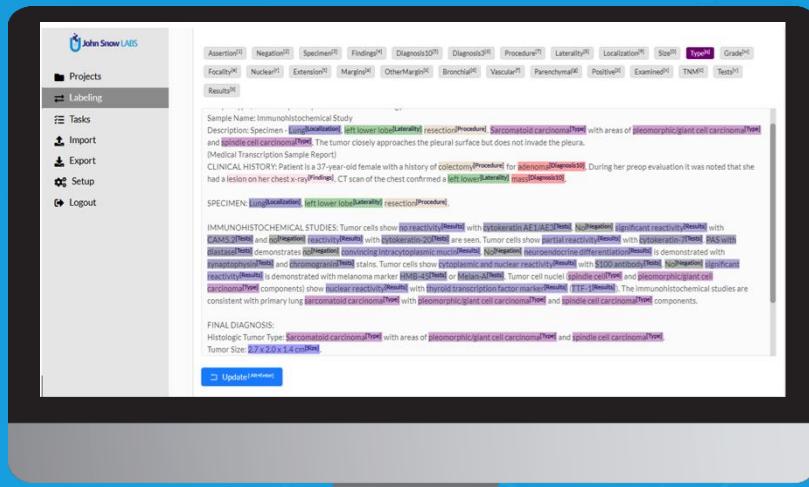
3. Wrong Granularity

“The patient is 30 years old” [Age]

4. Wrong Extraction

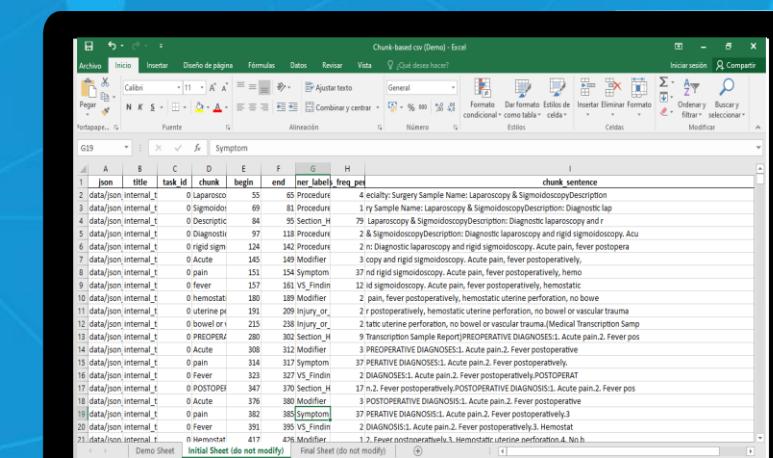
“The patient is young” [Age] (redundant if AG specify “only numeric values”)

Data Review



The screenshot shows the John Snow LABS labeling interface. On the left, a sidebar lists 'Projects', 'Labeling' (which is selected), 'Tasks', 'Import', 'Export', 'Setup', and 'Logout'. The main area displays a medical report for a 'Sample Name: Immunochemical Study'. The report details a 'Specimen' (left lower lobectomy resection) with 'Sarcomatoid carcinoma' and 'pleiomorphic giant cell carcinoma'. It includes a 'CLINICAL HISTORY' section about a 37-year-old female patient with a history of colostomy and a 'IMMUNOHISTOCHEMICAL STUDIES' section mentioning various markers like CK7, CK20, S-100, and Melan-A. A 'FINAL DIAGNOSIS' section concludes with 'Sarcomatoid carcinoma' and 'Tumor Size: 2.7x2.0x1.4 cm'. At the bottom right is a blue 'Update' button.

Manual Reviews



The screenshot shows a Microsoft Excel spreadsheet titled 'Chunk-based cov (Demo) - Excel'. The data is organized into columns labeled A through H. Column A contains numerical IDs, column B contains labels like 'SYMPTOM', 'A', 'D', 'E', 'F', 'G', and 'H'. Column C is labeled 'task_id' and column D is labeled 'chunk'. Columns E and F represent 'begin' and 'end' coordinates for text segments. Column G is labeled 'ner_label' and column H is labeled 'freq_per'. The data rows describe symptoms and associated clinical findings, such as 'Laparoscopy & Sigmoidoscopy' and 'Acute pain, fever postopera'. The last few rows provide context for 'POSTOPERA', 'DIAGNOSES1', and 'DIAGNOSES2'.

	A	B	C	D	E	F	G	H	
1	1	SYMPTOM							chunk sentence
2	2	A	task_id	chunk	begin	end	ner_label	freq_per	
3	3	data/json/internal.t	0	Laparosc	55	65	Procedure		4:entity: Surgery Sample Name: Laparoscopy & Sigmoidoscopy/Description
4	4	data/json/internal.t	0	Sigmoidos	69	81	Procedure		1:ry Sample Name: Laparoscopy & Sigmoidoscopy/Description: Diagnostic lap
5	5	data/json/internal.t	0	Descriptic	84	95	Section_H		2: & Sigmoidoscopy/Description: Diagnostic laparoscopy and r
6	6	data/json/internal.t	0	Diagnosti	97	118	Procedure		3:Diagnostic laparoscopy and rigid sigmoidoscopy. Acu
7	7	data/json/internal.t	0	rigid sigm	128	142	Procedure		4:n: Diagnostic laparoscopy and rigid sigmoidoscopy. Acute pain, fever postopera
8	8	data/json/internal.t	0	Acute	145	149	Modifier		5:copy and rigid sigmoidoscopy. Acute pain, fever postoperativ
9	9	data/json/internal.t	0	pain	154	154	Term		6:copy and rigid sigmoidoscopy. Acute pain, fever postoperativ, hemato
10	10	data/json/internal.t	0	hemostat	157	161	VS_Findin		7:acute pain, fever postoperativ, hemostat
11	11	data/json/internal.t	0	uterine pi	191	209	Injury_Ot		8:acute pain, fever postoperativ, hemostat
12	12	data/json/internal.t	0	bowel or i	215	238	Injury_Ot		9:acute pain, fever postoperativ, no bowel or vascular trauma
13	13	data/json/internal.t	0	PREPOPER	260	302	Section_H		10:acute pain, fever postoperativ, no bowel or vascular trauma/(Medical transcription Samp
14	14	data/json/internal.t	0	Acute	308	312	Modifier		11:PREOPERATIVE DIAGNOSES1. Acute pain.2. Fever pos
15	15	data/json/internal.t	0	pain	314	317	Symptom		12:PREOPERATIVE DIAGNOSES1. Acute pain.2. Fever postoperativ
16	16	data/json/internal.t	0	Fever	323	327	VS_Findin		13:DIAGNOSES1. Acute pain.2. Fever postoperativ/POSTOPERA
17	17	data/json/internal.t	0	POSTOPERA	347	370	Section_H		14:DIAGNOSES1. Acute pain.2. Fever postoperativ
18	18	data/json/internal.t	0	Acute	376	380	Modifier		15:acute pain, fever postoperativ, Acute pain.2. Fever postoperativ
19	19	data/json/internal.t	0	pos	380	380	Term		16:acute pain, fever postoperativ, Acute pain.2. Fever postoperativ
20	20	data/json/internal.t	0	Fever	391	395	VS_Findin		17:POSITIVE Diagnos1. Acute pain.2. Fever postoperativ.3
21	21	data/json/internal.t	0	menstru	417	426	Modifier		18:DIAGNOSES1. Acute pain.2. Fever postoperativ.3. Hemostat

CSV Reviews



Project Analytics – good example

Exercise – NER Annotation following AG

1. Create new project with the configuration specified [here](#)
2. Import tasks from json file
3. Assign all tasks to you as annotator
4. Annotate tasks according to [AGs](#)
5. Submit completion as Ground Truth (starred)

Outline

- 1. Introduction to Text Annotation**
- 2. Annotation Projects Setup in NLP Lab**
- 3. Annotation Guidelines**
- 4. Preannotation Resources**
- 5. Model Training**
- 6. Conclusions and further resources**

The Hub Of Resources

1. Integration with NLP Models Hub

- ✓ Check benchmarking information
- ✓ Download any resource with one click

2. Models

- ✓ Private repository of models
- ✓ Trained with NLP Lab, Manually uploaded, Downloaded from NLP Models Hub

3. Rules (examples available [here](#))

- ✓ Private repository of rules
- ✓ Rules editing interface

4. Prompts (examples available [here](#))

- ✓ Private repository of prompts
- ✓ Prompt editing interface

Playground

1. Deploy resource for quick tests on custom documents
2. Edit rules and prompts on the fly

John Snow LABS

Projects
Hub
Settings

Model Testing: ner_jsl

Insert Text 75/300 words

The patient is a pleasant 17-year-old gentleman who was playing basketball today in gym. Two hours prior to presentation, he started to fall and someone stepped on his ankle and kind of twisted his right ankle and he cannot bear weight on it now. It hurts to move or bear weight. No other injuries noted. He does not think he has had injuries to his ankle in the past. He was given adderall and accutane.

TEST

Results

The patient is a pleasant 17-year-old **Age** **gentleman** **Gender** who was playing basketball **today** **RelativeDate** in gym. Two hours prior **RelativeDate** to presentation, **he** **Gender** started to **fall** **Injury_or_Poisoning** and someone stepped on **his** **Gender** **ankle** **External_body_part_or_region** and kind of **twisted** **his** **right** **ankle** **Injury_or_Poisoning** and **he** **Gender** **cannot** **bear** **weight** **Symptom** on it now. It **hurts** to move or bear weight **Symptom**. No other **Injuries** **Injury_or_Poisoning** noted. **He** **Gender** does not think **he** **Gender** has had **Injuries** **Injury_or_Poisoning** in the past. **He** **Gender** was given **adderall** **Drug_BrandName** and **accutane** **Drug_BrandName**.

Model: ner_jsl

DETAIL OF MODEL

Edition clinical/models/ner_jsl_en_4.2.0_3.0_166618137...
 Task Named Entity Recognition
 Source Published on NLP Models Hub
 Uploaded by John Snow Labs
 Upload date -
 Trained with embeddings -

BENCHMARKING

name	f1	fn	fp	precision	recall
VS_Finding	0.8679	26.0	37.0	0.8484	0.8884
Direction	0.9144	264.0	418.0	0.897	0.9324
Respiration	0.928	4.0	5.0	0.9206	0.9355
Cerebrovas...	0.8532	12.0	20.0	0.823	0.8857
Family_Hist...	0.9872	1.0	1.0	0.9872	0.9872
Heart_Dise...	0.9033	50.0	47.0	0.906	0.9006

Clusters

1. Up-to-date information on the deployed servers
2. Computation resource management
3. Access deployed playground

Settings / Clusters

Clusters

Auto Refresh

License Info

1 floating license is available. Each floating license allows one healthcare, finance, or legal training and/or pre-annotation job according to its scope. There are no restrictions for running open-source parallel jobs as long as your system has enough resources and has been configured to allow multiple parallel jobs during setup.

ID	Server Name	License Used/Scope	Usage	Status	Deployed By	Deployed At	Action
3	playground	Floating License Legal:Inference, Legal:Training, Finance:Inference, Finance:Training, Ocr:Inference, Ocr:Training, Healthcare:Inference, Healthcare:Training	Playground	Busy	dia	1 minute ago	

Reuse Resources in your projects

1 Content Type ————— 2 Define what to annotate ————— 3 Reuse Resources ————— 4 Customize Labels

Model **Rules** **Prompts**

USE LABELS FROM AVAILABLE MODELS
Select any class/label from an existing model and add it to your project

Select All

SEARCH LABELS...

ASSERTION_JSL_AUGMENTED (8)▼

CLASSIFICATION_CLASSIFIERDL_USE_EMOTION (4)▼

NER_CLINICAL (3)▲

TREATMENT **PROBLEM** **TEST**

NER_DL (4)▼

NER_ISI (79)▼

+ ADD TO PROJECT CONFIGURATION

PREVIEW WINDOW
Preview your taxonomy and annotate a sample task to see the obtained output.

Age Gender Employment Date Medical_Problem Modifier Vital_Sign
Vital_Sign_Result Drug Present Absent Possible Family VS_Finding
Vital_Signs_Header Drug_BrandName PatientName Email PhoneNumber Name

The patient is a pleasant 17-year-old gentleman who was playing basketball today in gym. Two hours prior to presentation, he started to fall and someone stepped on his ankle and kind of twisted his right ankle and he cannot bear weight on it now. It hurts to move or bear weight. No other injuries noted. He does not think he has had injuries to his ankle in the past. He was given adderall and accutane.

Gender

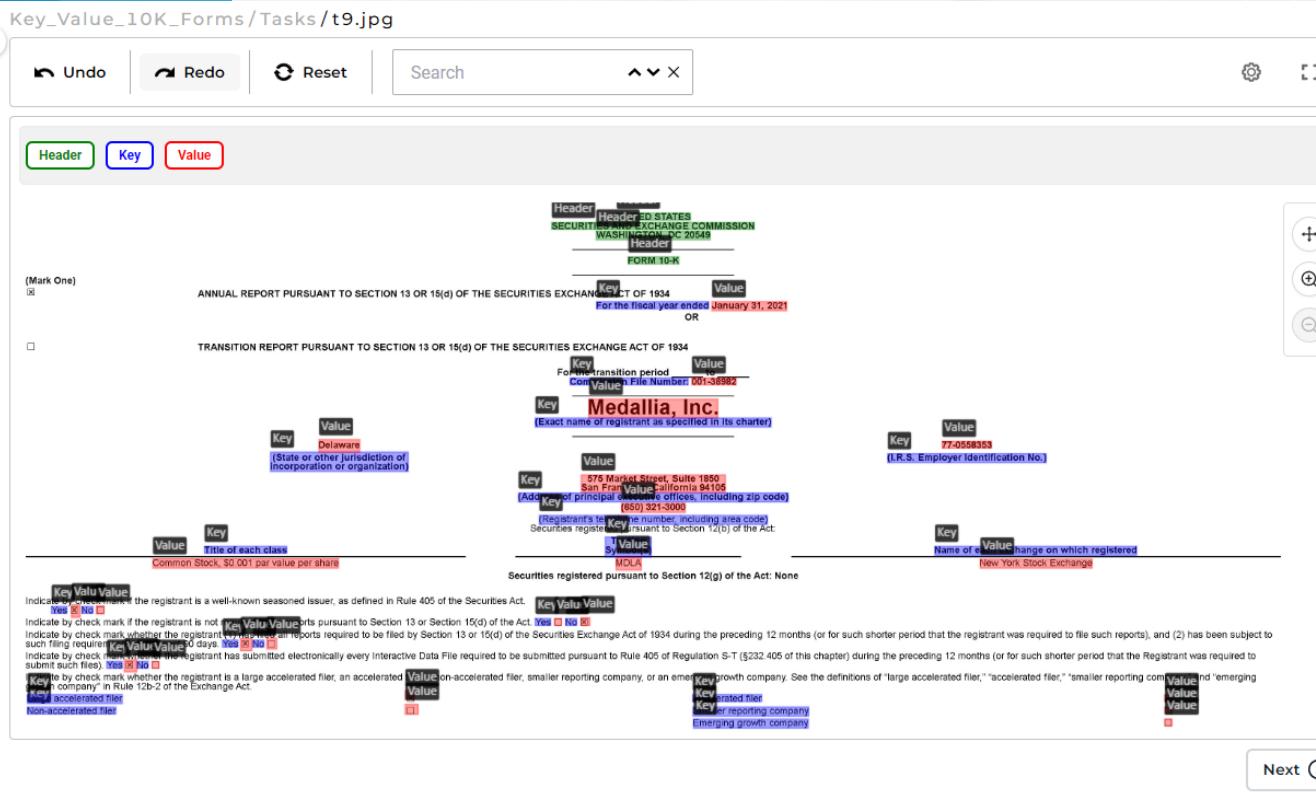
Female Male Unknown

Kind of Text

Exercise – Clinical Data Annotation

1. Create a project with the configuration described [here](#).
2. Assign all tasks to you as annotator.
3. Run preannotations on all tasks.
4. Click on one task to access the annotation screen.
5. Check the pre-annotations + click Edit and if necessary, update the annotations according to your AG.
6. Save your work and submit the new version of annotations.

Support for PDF and Image



The screenshot shows a document processing interface with the following key elements:

- Header:** Key_Value_10K_Forms/Tasks/t9.jpg
- Toolbar:** Undo, Redo, Reset, Search, and various configuration icons.
- Annotations:** A sidebar on the right lists 50 regions (rectangles) extracted from the document, each with a red number, a preview icon, and a "ro" status.
- Regions:** The sidebar also includes sections for CONNECTED WORDS (0) and RELATIONS (0).
- Document Content:** The main area displays the 10-K filing from Medallia, Inc., with annotations highlighting specific fields like the company name, address, and stock information.

- OCR license required
- [Select Visual NER template](#)
- [Create OCR server for pdf/image import](#)

[Sample configuration](#)

[Sample tasks](#)

Outline

- 1. Introduction to Text Annotation**
- 2. Annotation Projects Setup in NLP Lab**
- 3. Annotation Guidelines**
- 4. Preannotation Resources**
- 5. Model Training**
- 6. Conclusions and further resources**

Model Training

Projects / FoodProcessing / Training & Active Learning

Training & Active Learning

1. TRAINING SETTINGS

The settings chosen below also apply to Active Learning.

Training Type

License Type

Embeddings

3. TRAINING & ACTIVE LEARNING

Active Learning



Active Learning feature will automatically train a new model when the selected number of new completions is reached. The training will be triggered only when a license is available and the model server count is within the limit.

Active Learning

Save

Train Model

Test Configuration

Upload Training Script

History

Wizard
OFF

Last training succeeded

2. TRAINING PARAMETERS

Epoch

25

Learning Rate

0.001

Learning Rate Decay

0.005

Dropout

0.5

Batch

16

Train / Test data

Split dataset using Test/Train tags

Random Split Data Set

Validation Split 0.2

Confusion Matrix

Generate Confusion Matrix

Filter Completions By

submitted

Filter Tasks by Tag for Training

Select Tags

Click here to go to Training Resource Management

3 easy steps:

1. Choose the model type and the embeddings
2. Set training params including the tasks to use
3. Optional – turn on the active learning for future trainings

Example project

- [Food Processing](#)

Outline

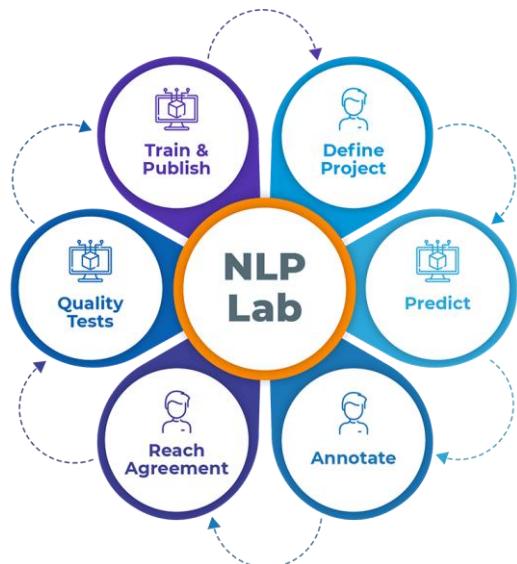
- 1. Introduction to Text Annotation**
- 2. Manual Annotation**
- 3. Annotation Guidelines**
- 4. Project Setup and Management**
- 5. Preannotations**
- 6. Conclusions and further resources**

Conclusions – NLP Lab Facilitates



- 1 Manual annotation process
- 2 Best practices for efficient and coherent annotations
- 3 Project templates and project configuration to suite your project
- 4 Reuse pretrained resources for faster and more efficient results
- 5 Detect annotation errors with analytics
- 6 Train models to learn common tasks

Learning Resources



Docs

<https://nlp.johnsnowlabs.com/docs/en/alab/quickstart>

Tutorials

https://nlp.johnsnowlabs.com/docs/en/alab/step_by_step_tutorials

Blog

<https://www.johnsnowlabs.com/nlp-lab-blog/>

Quick Install

- [AWS Marketplace](#)
- [Azure Marketplace](#)
- [OCI Marketplace](#)
- [On-Premise](#)

Support

- [Slack](#)
- [Email](#)

Questions?

- dia@johnsnowlabs.com
- andrei@johnsnowlabs.com
- radu@johnsnowlabs.com