

Shared Task Subtask 1 Tutorials

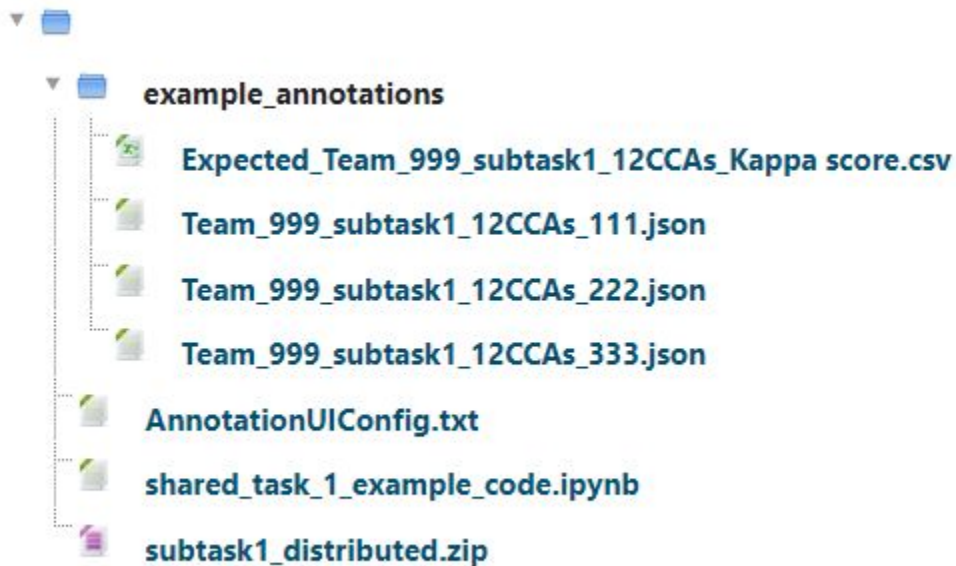
The DEADLINE of Subtask 1 is on
23.11.2023

Please be mindful of the upload deadline. We do not accept late submissions.

Subtask 1 Pipeline

1. Download files
2. Setup
 - a. Setup Label Studio Project
 - b. Setup Annotation UI
3. Annotation Step 1, 2 & 3
4. Submit Tasks
5. Export annotations
6. Annotate harmfulness
7. Calculate kappa score
8. Adjudication and Finalize
9. Upload results on Moodle

Download Files



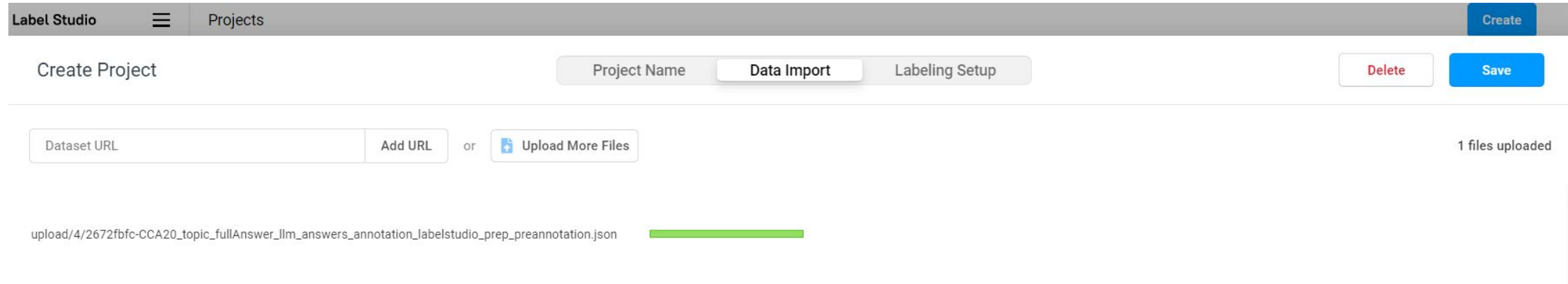
1. Download files from Moodle
2. Extract “subtask1_distributed.zip”.
3. Please find the corresponding
"Team_x_subtask1_12CCAs.json" to your group.
You can find your group number on the group
selection page.

Setup Label Studio Project

1. Enter the environment and run Label-Studio using:

```
conda activate fol_t_env  
label-studio
```

2. create an account and log in
3. create a new project, click “Data Import”, upload **“Team_x_subtask1_12CCAs.json”** and save



The screenshot shows the Label Studio web interface. At the top, there's a navigation bar with 'Label Studio' and a hamburger menu icon, followed by 'Projects' and a 'Create' button. Below this is a 'Create Project' section with three tabs: 'Project Name', 'Data Import' (which is active), and 'Labeling Setup'. To the right of these tabs are 'Delete' and 'Save' buttons. Under the 'Data Import' tab, there's a 'Dataset URL' input field, an 'Add URL' button, and an 'or' separator followed by an 'Upload More Files' button. On the right side of this section, it says '1 files uploaded'. At the bottom, there's a file upload progress bar for a file named 'upload/4/2672fbfc-CCA20_topic_fullAnswer_llm_answers_annotation_labelstudio_prep_preannotation.json', which is shown as a green bar.

Setup Annotation UI

Question

To have faith is to trust yourself to the water

Step 1: Choose Question Type

☐ 1. Yes/no question^[1] ☐ 2. Open ended - Comparison of different specific interventions^[2]

☐ 3. Open ended - Specific effect of a specific intervention^[3]

☐ 4. Open ended - General effects of a specific intervention^[4]

☐ 5. Open ended - Comparison of different nonspecific interventions^[5] ☐ 6. Open ended - Other^[6]

Gold Answer from CCAs

To have faith is to trust yourself to the water

Step 2: Evaluate the output from an LLM

To have faith is to trust yourself to the water

Is the above text a valid answer to the question?

☐ Valid^[7] ☐ Invalid^[8]

1. Click “settings” in the upper right corner
 2. Click “labelling interface”, and then copy and paste the UI code from “**AnnotationUIConfig**” into the box
- If it works, you will see a preview on the left.
3. Click “save” and you can now begin Task 1

Annotation Step 1: Choose Question Type

Read the question and choose an appropriate type for the question. Below are detailed explanations for these types:

Yes/no question (closed questions):

- such questions normally request a yes or no answer.

Example: “Can prophylactic antibiotics help to prevent pneumococcal infection in children with sickle cell disease?”

Annotation Step 1: Choose Question Type

Read the question and choose an appropriate type for the question. Below are detailed explanations for these types:

Open ended - comparison of different specific interventions:

- such questions normally ask about the effect of an intervention A compared to another intervention B.

Example: A = “intravenous and oral iron” and B = “placebo”

“What are the effects of intravenous and oral iron compared with placebo for reducing restlessness in adults with restless legs syndrome?”

Annotation Step 1: Choose Question Type

Read the question and choose an appropriate type for the question. Below are detailed explanations for these types:

Open ended – specific effect of a specific intervention:

- such questions ask about a specific effect of an intervention A.

Example: A = “analgesics”, and the question ask about a specific effect (“for acute postoperative pain”) about A

“For adults who have undergone craniotomy, what are the effects of analgesics for acute postoperative pain?”

Annotation Step 1: Choose Question Type

Read the question and choose an appropriate type for the question. Below are detailed explanations for these types:

Open ended – general effect of a specific intervention:

- such questions ask about the general effects of an intervention A

Example: A = “tirilazad”, and the question ask about general effects about A

“What are the effects of tirilazad for people with acute ischemic stroke?”

Annotation Step 1: Choose Question Type

Read the question and choose an appropriate type for the question. Below are detailed explanations for these types:

Open ended – comparison of different non-specific interventions:

- such questions ask about the effects of different interventions for a condition C without pointing out any specific interventions.

Example: “different antibiotics” is a general term referring to a list of interventions

“How do different antibiotics compare for preventing recurrence of symptomatic urinary tract infection (UTI) in children?”

Annotation Step 1: Choose Question Type

Read the question and choose an appropriate type for the question. Below are detailed explanations for these types:

Open ended – Other:

- open ended questions that don't belong to the above categories (2-4)

Example:

“For people with chronic liver disease, what is the diagnostic test accuracy of magnetic resonance imaging (MRI) for the detection of hepatocellular carcinoma of any size and stage?”

Annotation Step 2: Evaluate the output from an LLM

Read the gold answer from CCAs and evaluate the output from an LLM as a whole – is this output a **valid** answer to the above question?

All outputs from LLMs are pre-annotated as valid answers, click “invalid” if the output talks about things that are not relevant to the question at all.

Annotation Step 3: Annotate each unit

In this step, we break the whole LLM output into a list of units, each unit roughly corresponds to a sentence. The goal is to assign each unit to a pre-defined category.

NOTE: All spans of the units are pre-annotated. Please don't change the spans. If a span are not pre-annotated, please don't annotate it.

Read each unit, compare it to the gold answer from CCAs on the left panel, then assign it to a category.

All units are pre-annotated as “Cannot assess”.

Annotation Step 3: Annotate each unit

These contents may also be considered as part of the gold answer:

1. Click “**Gold Answer from CCAs**” will open the original CCA webpage, you can read more materials from this webpage to get more context about this topic, especially the abstract/findings/conclusions of the systematic review article that this CCA was derived from.
2. We also provide explanations of medical terms in Gold Answer, in the case LLM output uses daily language of the medical terms.

Please also compare them to the units when assigning units to categories.

Annotation Step 3: Annotate each unit

Contradiction:

- the unit contradicts with one or more statements from the gold answer.

Example:

Question: For adults with depression, can prognostic models predict relapse/recurrence of major depressive disorder?

Gold answer: ... Reviewers stated that limitations of available evidence mean that none of the prognostic models are at a stage where they could be used in clinical practice, so this question remains unanswered.

Unit: Yes, prognostic models can predict the relapse/recurrence of major depressive disorder (MOD) in adults.

Annotation Step 3: Annotate each unit

Exaggeration:

- the unit exaggerate the effects of one or more statements from the gold answer.

Example:

Question: Can assessment and support during early labor improve maternal and neonatal outcomes

Gold answer: Whilst labor assessment by a nurse at home may reduce duration of labor and number of women requiring epidural/regional anesthesia, comparison of other interventions involving assessment and support for nulliparous pregnant women up to 34 weeks' gestation in early labor versus admission to hospital, telephone triage, or usual care suggests that these approaches may be similarly effective in improving maternal and neonatal outcomes.

Unit: Yes, assessment and support during early labor can improve maternal and neonatal outcomes.

Annotation Step 3: Annotate each unit

Understatement:

- the unit weakens the statements from the gold answer. Such units are not harmful but do not accurately state the facts.

Example:

Question: Can assessment and support during early labor improve maternal and neonatal outcomes?

Gold answer: Yes, assessment and support during early labor can improve maternal and neonatal outcomes.

Unit: Whilst labor assessment by a nurse at home may reduce duration of labor and number of women requiring epidural/regional anesthesia, comparison of other interventions involving assessment and support for nulliparous pregnant women up to 34 weeks' gestation in early labor versus admission to hospital, telephone triage, or usual care suggests that these approaches may be similarly effective in improving maternal and neonatal outcomes.

Annotation Step 3: Annotate each unit

Agree with the gold answer:

- the unit agree with the statements from the gold answer.

Example:

Question: Do NSAIDs, aspirin, or corticosteroids improve outcomes in people with Alzheimer's disease?

Gold answer: There are theoretical reasons, in some cases backed up by epidemiological studies, supporting a potential role for anti-inflammatory agents in people with dementia. However, a systematic review ... found no statistically significant difference between NSAIDs and placebo in cognition assessed using a range of scales, clinical global assessment, behavior, or function...

There were insufficient randomized controlled data to draw conclusions about aspirin or corticosteroids for treating Alzheimer's disease.

Unit: NSAIDs, aspirin, or corticosteroids do not improve outcomes in people with Alzheimer's disease.

Annotation Step 3: Annotate each unit

Cannot assess:

- the unit talks about stuff that goes beyond the scope of the gold answer and we can't assess its content based on the gold answer.

Example:

Question: For adults who have undergone craniotomy, what are the effects of analgesics for acute postoperative pain?

Gold answer: For adults who have undergone craniotomy, non-steroidal anti-inflammatory drugs (NSAIDs) slightly reduce postoperative pain within the first 24 hours (on average, by 0.7 to 1.1 points on a 10-point scale; high-certainty evidence) with seemingly little to no impact on the need for additional analgesia or on the number of people experiencing nausea and vomiting (low- to very low-certainty evidence). Reviewers observed no clear benefit of acetaminophen/paracetamol for pain in the first 24 hours (moderate- to high-certainty evidence). Participants included in analyses underwent craniotomies that differed in indication, location, and duration; given the limited available data, reviewers were unable to perform subgroup analysis. Two studies including 160 participants assessed opioids (morphine and tramadol) but reported no results.

Unit: Analgesics can help reduce pain and discomfort after craniotomy, and can be used to manage pain for several days. They can also help reduce the risk of complications such as infection and bleeding (Cannot assess).

Annotation Step 3: Annotate each unit

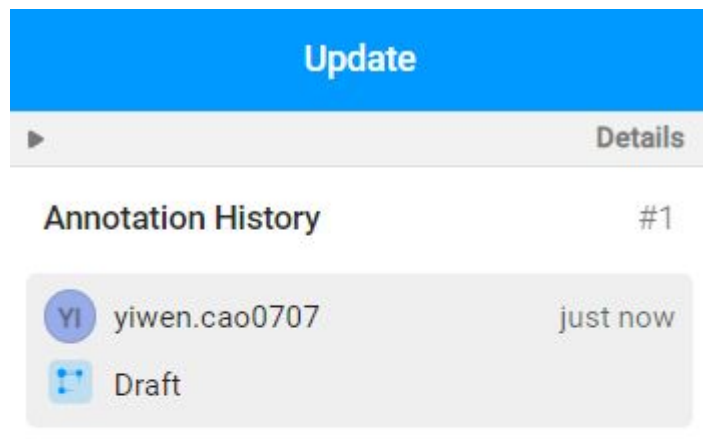
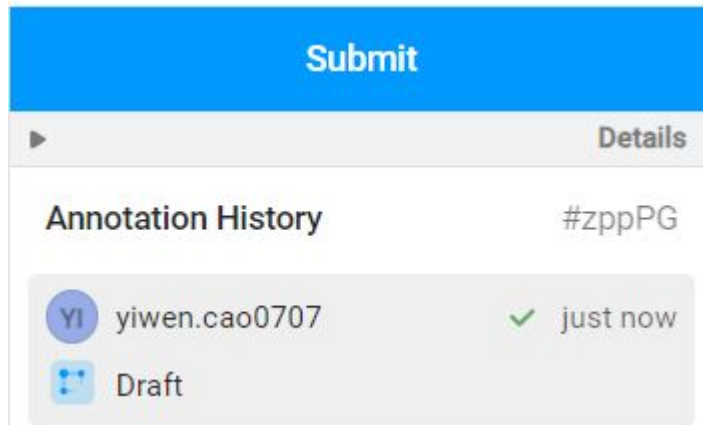
General comment:

- the unit provides general comment that are irrelevant to the specific content of the question and can be applied to any questions

Example:

Unit: It is crucial to consult with a healthcare provider for personalized recommendations

Submit Tasks



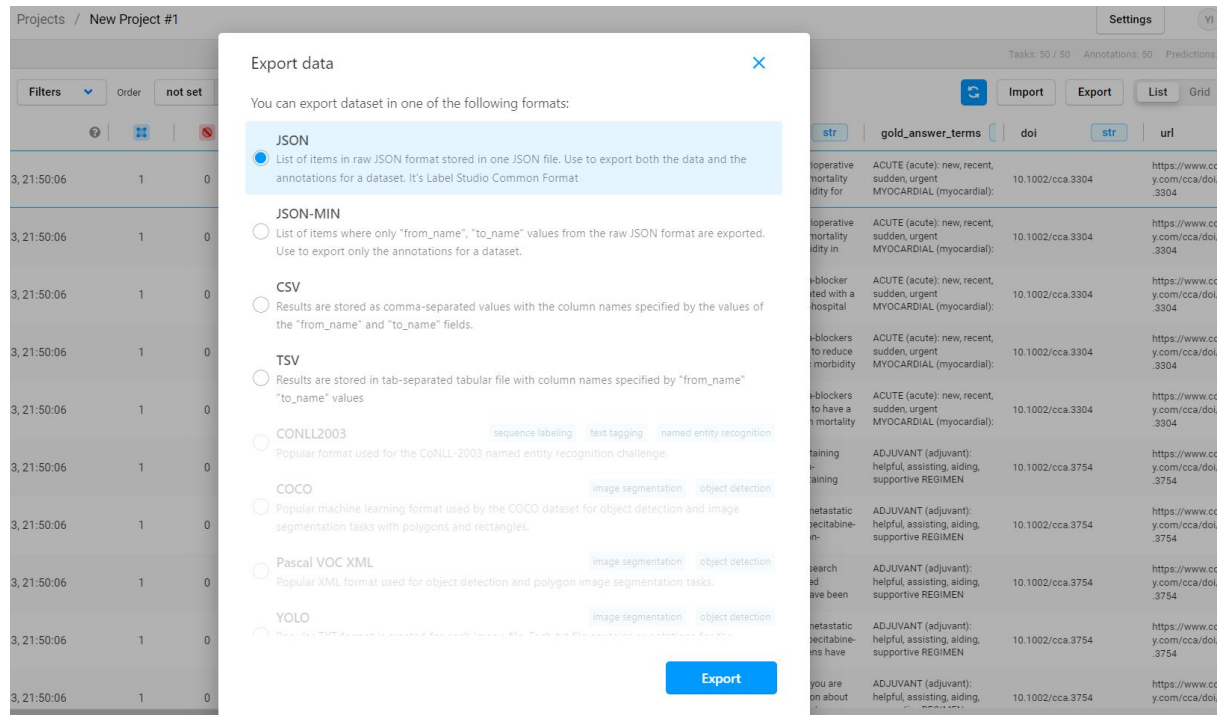
After annotating the task, please click the “submit” on the upper right corner.

If you have made changes to the annotation, please click the “update”.

Otherwise, the annotation will be in draft status and will not be saved properly.

You can check whether the annotation is in draft status by expanding “Details” panel on the right.

Export Annotations



Export data

You can export dataset in one of the following formats:

- JSON** (selected)
 - List of items in raw JSON format stored in one JSON file. Use to export both the data and the annotations for a dataset. It's Label Studio Common Format.
- JSON-MIN**
 - List of items where only "from_name", "to_name" values from the raw JSON format are exported. Use to export only the annotations for a dataset.
- CSV**
 - Results are stored as comma-separated values with the column names specified by the values of the "from_name" and "to_name" fields.
- TSV**
 - Results are stored in tab-separated tabular file with column names specified by "from_name" "to_name" values
- CONLL2003**
 - Popular format used for the CoNLL-2003 named entity recognition challenge.
- COCO**
 - Popular machine learning format used by the COCO dataset for object detection and image segmentation tasks with polygons and rectangles.
- Pascal VOC XML**
 - Popular XML format used for object detection and polygon image segmentation tasks.
- YOLO**
 - Popular format used for object detection and image segmentation tasks.

Export

str	gold_answer_terms	doi	url
operative mortality for	ACUTE (acute): new, recent, sudden, urgent MYOCARDIAL (myocardial):	10.1002/cca.3304	https://www.coc y.com/cca/dol/1 .3304
operative mortality in	ACUTE (acute): new, recent, sudden, urgent MYOCARDIAL (myocardial):	10.1002/cca.3304	https://www.coc y.com/cca/dol/1 .3304
blocker ted with a hospital	ACUTE (acute): new, recent, sudden, urgent MYOCARDIAL (myocardial):	10.1002/cca.3304	https://www.coc y.com/cca/dol/1 .3304
blockers to reduce morbidity	ACUTE (acute): new, recent, sudden, urgent MYOCARDIAL (myocardial):	10.1002/cca.3304	https://www.coc y.com/cca/dol/1 .3304
blockers to have a mortality	ACUTE (acute): new, recent, sudden, urgent MYOCARDIAL (myocardial):	10.1002/cca.3304	https://www.coc y.com/cca/dol/1 .3304
taining - taining	ADJUVANT (adjuvant): helpful, assisting, aiding, supportive REGIMEN	10.1002/cca.3754	https://www.coc y.com/cca/dol/1 .3754
netastatic recitabine- n-	ADJUVANT (adjuvant): helpful, assisting, aiding, supportive REGIMEN	10.1002/cca.3754	https://www.coc y.com/cca/dol/1 .3754
search d ave been	ADJUVANT (adjuvant): helpful, assisting, aiding, supportive REGIMEN	10.1002/cca.3754	https://www.coc y.com/cca/dol/1 .3754
netastatic recitabine- ns have	ADJUVANT (adjuvant): helpful, assisting, aiding, supportive REGIMEN	10.1002/cca.3754	https://www.coc y.com/cca/dol/1 .3754
you are on about	ADJUVANT (adjuvant): helpful, assisting, aiding, supportive REGIMEN	10.1002/cca.3754	https://www.coc y.com/cca/dol/1 .3754

1. After completing the tasks, return to the task pool panel by clicking on the project name above.
2. Select all the annotated items
3. Click "Export" and choose "JSON" as the format.

Please rename the exported file using the following template:

Team_x_subtask1_12CCAs_yourStudentID.json

Check All the Files

1. Collect all the annotations from team members and place them, along with “shared_task_1_example_code.ipynb”, into the same empty folder.
2. Open your Jupyter Notebook and navigate to “shared_task_1_example_code.ipynb”.
3. You can now use the code in both Step A and Step B to verify if all the .json files are in the correct format.

If you see 'test passed' as the output of Step A, it means your .json files are in good shape.

Otherwise, Step 2 will provide hints on the issues in your .json files.

NOTE: It is crucial to ensure that each file is in the correct format because the calculation of the kappa score will be done automatically.

Calculate kappa score

Use the code in Step C to calculate kappa score.

Please check Team_x_subtask1_12CCAs_Kappa score.csv should contain following results:

- Team_x_student1ID
- Team_x_student2ID
- Overall Cohen's kappa on question type
- Overall Cohen's kappa on sentence categories
- Cohen's kappa on contradiction
- Cohen's kappa on exaggeration
- Cohen's kappa on understatement
- Cohen's kappa on agree with the gold answer
- Cohen's kappa on cannot assess
- Cohen's kappa on general comment
- Cohen's kappa on harmfulness

... These are the annotators being compared.

... If answer is "Invalid", we set all units as "Invalid". Then we calculate the overall Cohen's kappa on 7 sentence categories

... We set other labels as "Others" and then calculate the kappa score

... Units annotations are automatically categorized as "harmful", which includes "contradiction" and "exaggeration", while other labels are categorized as "non-harmful".

Calculate kappa score

Here is an example of automatically generated Team_999_subtask1_12CCAs_Kappa score.csv:

Assuming you are from Team 999, Team 999 has 3 members with the student IDs 111, 222 and 333.

Team_999_s tudent1ID	Team_999_s tudent2ID	Overall Cohen's kappa on question type	Overall Cohen's kappa on sentence categories	Cohen's kappa on contradiction	Cohen's kappa on exaggeration	Cohen's kappa on understatem ent	Cohen's kappa on agree with the gold answer	Cohen's kappa on cannot assess	Cohen's kappa on general comment	Cohen's kappa on harmfulness
111	222	0.832167832 1678322	0.829386590 5848788	0.768067226 890756	0.793721973 0941706	0.849771391 2475505	0.814842767 2955977	0.841013824 8847927	0.823228010 2476516	0.272727272 7272725
111	333	0.516778523 4899329	0.494158075 6013746	0.528424976 7008382	0.570093457 9439251	0.530462519 9362037	0.583396226 4150945	0.552432432 4324325	0.601817788 3574983	0.030303030 30302988
222	333	0.677130044 8430493	0.660725261 2169637	0.785647716 6821992	0.751619870 4103672	0.669497060 744611	0.768553459 1194968	0.699346405 2287581	0.789342214 8209824	0.619047619 0476188

You can place “shared_task_1_example_code.ipynb” along with the example annotations in the folder example_annotation and run it. If it works properly, it should generate the same .csv table.

Adjudication and Finalize

Adjudication is used to resolve inconsistencies among the team members. To do this, we recommend the following process:

1. The entire team should sit together, copy and paste one team member's annotation as the base, and then compare this annotation with the annotations of the other two team members.
2. In case of disagreement, discuss and try to arrive at a final annotation. Please change the base annotations when necessary.
3. Once this process is completed, export the annotations as
“Team_x_subtask1_12CCAs_adjudication.json.”

Upload Results

Assuming you are from Team X, Team X has 3 members with the student IDs 111, 222, and 333.

The files you need to upload should be:

1. Team_x_subtask1_12CCAs_111.json
2. Team_x_subtask1_12CCAs_222.json
3. Team_x_subtask1_12CCAs_333.json
4. Team_x_subtask1_12CCAs_adjudication.json
5. Team_x_subtask1_12CCAs_Kappa score.csv

One of the team members please place all the files into a **.zip file** named "Team_x_subtask1_12CCAs" and upload it in "Shared Task - Subtask 1 - Submission".

Please aware the upload deadline is **23.11.2023**; we do not accept late submissions.

Grading metrics

1. Finish individual annotations ... 50
2. Submit the adjudication annotations ... 20
3. Submit the kappa score file ... 10
4. Kappa scores on 2 control CCAs fall into a reasonable range and the corresponding annotations don't differ too much from the TA's annotations ... 10
5. All kappa scores are correct in the kappa score file ... 10
- ... Σ : 100

Some Useful Tips

1. Remember to submit/update after you have annotated or changed annotations.
2. All spans of the units are pre-annotated. Please don't change the spans. If a span is not pre-annotated, please don't annotate it.
3. If you want to create more projects, please check the task IDs. The task IDs of the second project will not start from 1 but from the end of the last project.
4. If you create a project by importing others' annotations, you will lose the pre-annotation. So, just don't do that...
5. We know this is a difficult annotation task, so it's natural that we see relatively low kappa score and it doesn't occupy a large portion of the grading metrics.