Harmony PDF challenge for Doxa

Situation

Harmony has developed a cutting-edge functionality that allows users to upload a PDF document, which the system then processes to identify and extract the text of questionnaire questions. This technology represents a significant advancement in the field of document processing and data extraction. You can try Harmony at harmonydata.ac.uk. However, handling PDFs is hard due to varying formats. The PDF extraction needs improvement.

Objective of the Competition

The main objective of this competition is to build upon Harmony's existing technology to create a more efficient, accurate, and robust tool for extracting questionnaire questions from a variety of documents. Participants are encouraged to innovate and develop solutions that can handle a wide range of document formats and structures.

Problem description

We have questionnaires like the following and we need to identify the questions ("Little interest or pleasure in doing things") but not the other text.

PATIENT HEALTH QUESTIONNAIRE-9 (PHQ-9)

Over the last 2 weeks, how often have you been bothered by any of the following problems? *(Use "✔" to indicate your answer)*	**Not at all**	**Several days**	**More than half the days**	**Nearly every day**
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3

Your annotation task is to make training data to improve Harmony.

I am supplying the PDFs converted to **plain text** form for you.

The training code

The repository with the training code is as follows. We have one example Conditional Random Fields model that we have trained.

https://github.com/harmonydata/pdf-questionnaire-extraction

It's important that your model is small and doesn't need lots of memory to run, as it should be easy to deploy. Ideally not much bigger than the existing CRF model.

The data

You can download everything you need from:

https://harmonyapistorage.z33.web.core.windows.net/20240822_annotated-files.zip

The files have been annotated with XML-esque tags <q>...</q> and <o>...</o>.

<q> marks a question ("How often have you felt nervous?")

<o> marks a question option ("very much so", "somewhat", etc)

Example un-annotated file

```
Nearly
every
day

1. Little interest or pleasure in doing things 0 1 2 3

2. Feeling down, depressed, or hopeless 0 1 2 3
```

Example file with tags

```
Nearly
every
day

1. <q>Little interest or pleasure in doing things</q> 0 1 2 3

2. <q>Feeling down, depressed, or hopeless</q> 0 1 2 3
```