Assignment

You are going to locate segments in a text using multi-output regression!

Dataset creation:

1. First, create a list of 10 different datetime formats. E.g “%Y-%m-%d %H:%M:%S”, “%A, %B %d, %Y %I:%M %p”, “%Y/%m/%d“ and so on.

2. Consider the following dataset,
   https://www.kaggle.com/datasets/asad1m9a9h6mood/news-articles

3. For each entry, randomly add a datetime segment in the article column. For example, if the article text looks like this,
   “Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse sit amet diam vestibulum, mattis libero non, imperdiet sem. Etiam condimentum, mauris eu faucibus hendrerit, felis felis lobortis nibh, nec tempus neque enim nec tortor. Phasellus porttitor, ligula in imperdiet mollis, nunc ex malesuada risus, lacinia ornare quam tellus vitae elit. Nulla placerat viverra condimentum. Duis id dignissim dolor. Nullam libero erat, porta et libero non, sodales commodo nulla. Ut dapibus convallis nunc non molestie. Cras at finibus enim.”

4. Create a datetime object using today() and generate new datetime by adding/subtracting random timedeltas into/from the today’s datetime. Then format the datetime using a randomly selected format from the list created in the first step. After that, randomly select an index, let's say 20 in the text. Next, insert this formatted datetime string after the selected index. So the final article will look like this,
   “Lorem ipsum dolor sit amet, consectetur **10-04-2022 5:00 PM** adipiscing elit. Suspendisse sit amet diam vestibulum, mattis libero non, imperdiet sem. Etiam condimentum, mauris eu faucibus hendrerit, felis felis lobortis nibh, nec tempus neque enim nec tortor. Phasellus porttitor, ligula in imperdiet mollis, nunc ex malesuada risus, lacinia ornare quam tellus vitae elit. Nulla placerat viverra condimentum. Duis id dignissim dolor. Nullam libero erat, porta et libero non, sodales commodo nulla. Ut dapibus convallis nunc non molestie. Cras at finibus enim.”

5. Add three new columns - is_deadline, start, end. For the example above, is_deadline will be 1, start will be 20 and end will be (20+len(**10-04-2022 5:00 PM**)). This is how you can create a positive sample. Also create some negative samples where is_deadline is 0 and both start and end also 0. See the table below for examples of both positive and negative samples.

article	date	heading	news_type	is_deadline	start	end

Lorem ipsum dolor sit amet, consectetur **10-04-2022 5:00 PM** adipiscing elit	2/2/2015	Nullam libero erat, porta	business	1	20	35
nunc ex malesuada risus, lacinia ornare quam tellus vitae elit. Nulla placerat viverra condimentum. Duis id dignissim dolor.	2/4/2015	malesuada risus, lacinia ornare quam	business	0	0	0

Use the articles from the dataset to create both positive and negative examples. Your dataset must contain at least 2500 examples.

Find Segment:
Now the dataset is created, using multi-output regression to try to predict the existence of the deadline segment and if it is present then its start and end index.

** Create a Github repo for this task and make it public. Make sure your repo contains all the relevant code and data files. When complete send the Github link by replying to the email that you received this task from.