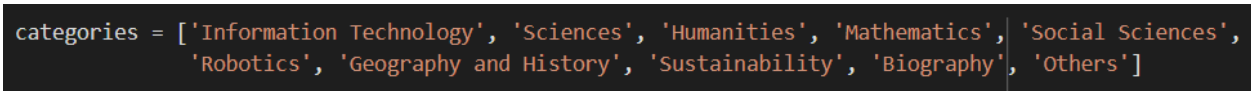
1. The Starting Point

At first we spoke to the PO in order to understand how many levels were needed, and until what granularity (how many levels down). It was decided to use the existing first-level classification, as it is very widely used in other platforms, and it seemed as a well-known top category for content creators.



```
categories = ['Information Technology', 'Sciences', 'Humanities', 'Mathematics', 'Social Sciences',
              'Robotics', 'Geography and History', 'Sustainability', 'Biography', 'Others']
```

Therefore we decided for a limit of ten first-level categories, and 10 subcategories (second level) for each first-level category as a structure, and to limit the scope. Also we decided for this because we thought that we are gaining usability for the searchers, at a price of granularity.

To begin creating the ontology, we sourced data through web scrapers. Tofocus the search, we used a list of “search words” for the crawler, which were then used as the mock first-level category/classification (approx. 2300 data points). These search words were the first-level categories’ list from above.

From each data point we got:

- The first-level category (because that is what the uploader or content creator would be selecting in the upload form) was filled with the search word
- The short description (excerpt)

Lastly, we did not use the current ontology because of several reasons:

- There is very little data on the database backing it up
- The ontology hence, does not represent the categories of the data that is on the platform
- We do not know what logic was followed to create this ontology

2. The Approach

<u>First-level category:</u>
The first-level category is selected by the uploader/content creator directly when uploading the content. Therefore, there is no model necessary for this stage. A possible alternative would be to run zero-shot classification with the given list of first-level categories. However, we decided against it to give the content creator “power” to decide in which category it should fall.

<u>Second-level category:</u>
To create meaningful subcategories that represent the content on the platform, an initial dump of content needs to run through the keyBERT mono algorithm. This will extract a list of the most meaningful keywords for each data point (content) sorted from the most meaningful to the least.

With human intervention, these keywords need to be labeled as a subcategory. The most practical way is doing this for every first-level category separately (in our case we used

“Robotics”). So for example a content whose resulting keywords are [cat, feline, lion, paws], could be labeled by a human into a subcategory called “feline”.

After having labeled the list of keywords into a maximum of 10 subcategories (per category), the zero-shot classification (ZSC) model is used. The ZSC model forces the content it analyzes into one of the given (labeled) subcategories. If it does not find a fitting subcategory, we decided for it to be classified into a subcategory called “others”.

We recommend that every 4-6 months the “others” subcategory to be reviewed and to run the content through the keyBERT mono algorithm and review the resulting keywords for each data point. This allows for the stakeholder to see if there are any subcategories that are relevant but not part of these 10 defined ones. We think this is the key for having a dynamic ontology, which is representative of the content that actually is on the platform. As an example, maybe a subcategory was created called “molecular biology”. However, there is more content under “others” that could be classified into “microbiology”. So the stakeholder would have to make a decision if to keep the more representative label (“microbiology”) or to keep “molecular biology”. Alternatively, maybe some content could be grouped together into a broader term.

The reason why ZSC was chosen is because it works extremely well with data that it has never seen. There is no control on what type of content might be uploaded, so it is important that the model can handle uncertainty very well, or “unseen” data. Most other methods need pretraining, which also implies having a lot of data points to come up with a meaningful result. Those other methods normally do not handle data very well that they have not been trained on. And that is the benefit of ZSC - it is proven to work better than supervised classification models.

We experimented with using the LDA model and also other keyword extractors, but the results were less satisfactory. We cannot compare models in terms of numeric accuracy, but we can compare the models based on the qualitative output. And there is where ZSC and keyBERT outperformed the other models. The results were more meaningful.

3. Other Categories

We just did the subcategory labeling for the first-level category “Robotics”, although we have data points for the other topics as well. However, since there is human intervention needed in order to label the resulting keywords, this is time-intensive and also requires a certain expertise in the field. Therefore, we decided to just show one category as a proof-of-concept.

As mentioned before, the ideal start for this ontology is to have a relevant number of data points to be analyzed in each first-level category, and then do the labeling of the keywords to create the 10 subcategories for each category. After that initial work, only the “others” subcategory can be analyzed more in depth, and maybe weights could be given to each subcategory in order to see which ones are becoming less relevant because another subcategory has more content related to it.

4. **Miscellaneous:**
    a. Mentoring sessions: Namrata and Ekaterina
    b. Other teams: Ceylan, Cayla, Manju, Silviya