Problem Setting

Sites are requesting an ad service in order to obtain ads to display to their visitors. This process generates input data (site requests) that each contain a list of keywords from the sites. Each input request contains a single list, and each site sends the same keywords list.
To be useful, this data needs to be mapped to site thematic or vertical (e.g. here is Google's verticals list).

The goal is to reliably achieve this mapping (keywords → site thematic).

Our task has three levels.
The 1st one is the only one mandatory. Consider the next two as giving you a huge bonus for your application.

Task level 1 (mandatory)

Describe in free text how you would achieve the goal.
What is your approach to the problem?
How will you source data for your purposes?
What are edge cases and considerations to take into account?
What is challenging and why?

Submit to us: A document describing your idea and thought process.

Task level 2 (optional)

Create a keywords crawler and scrape the keywords from top 20 sites for the following verticals (commonly you can find site keywords in the meta keywords or description tags)
"Sports" site rankings
"TV Movies and Streaming" site rankings
"File Sharing and Hosting" site rankings

Clean up and prepare the resulting dataset to be ready for feature engineering and modelling.

Submit to us: Provide your code and documentation about your data preparation approach. (e.g. a Jupyter notebook with all the details)

Task level 3 (optional)

Using your generated dataset, train a model that classifies the vertical of a list of keywords.
Run the model on the provided testing sample of keyword lists and predict their site verticals.

Submit to us: Provide your code and documentation about your approach, alongside a file with your predictions on the testing sample data. (e.g. a Jupyter notebook with all the details)