

AI Engineer | Interview Tasks

ML/DL Task - Clustering and Title Generation

You are provided with a set of news articles collected from several websites. The task is to create a python module using machine learning/NLP/Deep learning techniques to cluster the documents with similar content and generate a title for the cluster which represents these news articles. Number of clusters should be customisable by the user and it can contain anywhere from 1 to 100 cluster points. You are free to use any kind of open-source code/model but should give proper justification on why you choose the same. The title generating should not be an extractive one, it should be *generative* i.e any selected sentence among the articles cannot be considered as the title for the cluster.

The candidate is free to crawl for additional content or download and use any publicly available dataset for training the model.

Self-trained models will get additional weightage in evaluation.

Provided: News articles dataset with 1k documents along with metadata in json format.

Evaluation criteria: At the time of evaluation, we will test multiple clusters to evaluate the performance of the module using unseen documents.

How to submit the solution:

1. **ZIP of the solution** should contain
 - a. A runnable solution (code, comments, model etc)
 - b. Proper documentation (approach, challenges faced etc)
 - c. Requirements.txt(dependencies and versions)
 - d. Reference links/papers if any

-----Confidential-----