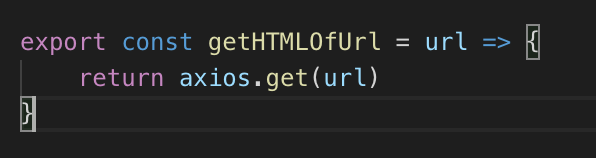
Đề tài: Nghiên cứu, xây dựng các công cụ tự động trích rút tin theo mẫu từ các URL và phân loại tin tức theo chủ đề

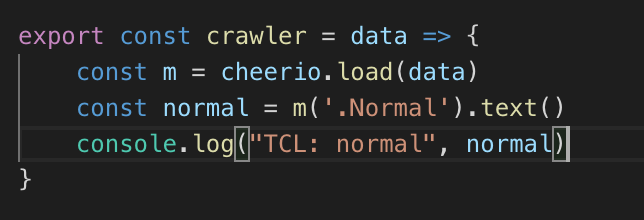
1. Công cụ: Trích rút tin từ URL

* Ngôn ngữ Nodejs
* Thư viện sử dụng: axios, cheerio

1. Axios

* Thư viện này cho phép chúng ta down HTML của 1 trang web về
* 

1. cheerio

* Thư viện này cho phép chúng ta lấy dữ liệu từ HTML trong các thẻ được định sẵn
* 

1. Bài toán phân loại tin tức theo chủ đề
2. Thu thập dữ liệu

* Các tin tức sẽ lấy từ trang web Express.vn
* Các url này sẽ được lưu trong 1 file (url.txt)
* Trong file: mỗi dòng sẽ lưu url, thể loại của tin

1. Trích rút, xử lý dữ liệu

* Sử dụng thư viện để trích rút tin tức từ các url về
* Sau đó các lưu các thông tin này dưới dạng json, hoặc text để cho quá trình trainning

1. Trainning: sử dụng thư viện node-fasttext

* Yêu cầu tranning có dạng 

1. ứng dụng

* ngôn ngữ sử dụng: Nodejs
* Ứng dụng gồm 3 phần:
  + Trích rút URL
  + Tranning phân loại URL
  + Testing