**GIỚI THIỆU**

1. **Mô tả bài toán**

Trong cuộc sống hiện đại ngày nay, mỗi ngày đều có hàng triệu thông tin được tạo ra xung quanh chúng ta như kết quả học tập, tài liệu đánh giá,… Điều này làm cho chúng ta trở nên khó khăn trong việc ghi nhớ tất cả các thông tin này, vì thế con người phải lưu lại tất cả thông tin với nhiều phương pháp khác nhau. Nhưng giải pháp này lại có một vấn đề là chúng có nhiều thông tin trùng lặp gây nên kho dữ liệu quá tải, bên cạnh đó là chúng không được sắp xếp theo một thứ tự nào và không có cấu trúc định dạng thống nhất điều này gây nên sự phiền toái trong việc truy xuất thông tin cần thiết cho bản thân. Chính vì vậy, nhằm giải quyết vấn đề này, nhóm tác giả Ratish Puduppully, Li Dong, Mirella Lapata đã nghiên cứu mô hình Data-to-Text Generation with Content Selection and Planning nhằm giải quyết được vấn đề cấu trúc, sắp xếp và chọn lọc dữ liệu để xuất ra thành dạng văn bản với định dạng và thứ tự được sắp xếp theo một tiêu chuẩn được quy định trước, từ đó mọi người có thể truy xuất thông tin cần thiết một cách dễ dàng, nhanh chóng.

1. **Mục tiêu đề tài**

Do nhu cầu ngày càng lớn về việc xử lý dữ liệu với kích thước lớn của thời đại, mô hình Data-to-Text Generation with Content Selection and Planning ra đời. Nhằm mục tiêu giải quyết các vấn đề bất cập về dữ liệu có kích thước lớn như cấu trúc đa dạng, không theo thứ tự, nhiều thông tin trùng lặp,… từ đó mô hình được nghiên cứu và xây dụng để tạo ra những đoạn văn bản có nội dung khái quát, chứa những thông tin quan trọng, gắn kết, theo một cấu trúc văn bản nhất định và đặt biệt không có sự trùng lặp thông tin giúp mọi người có thể truy xuất thông tin một cách dễ dàng, nhanh chóng hiểu được nội dung cần thiết.

1. **Phương pháp giải quyết**

Với một tập dữ liệu lớn, không theo một cấu hay thứ tự nhất định rất khó để truy xuất đúng thông tin mong muốn. Vì vậy chúng ta cần phải địh dạng lại dữ liệu này thành một cầu trúc văn bản theo một thứ tự bằng cách thông qua mô hình Data-to-Text Generation with Content Selection and Planning. Dựa trên tài liệu thu thập được đó là dữ liệu thi đấu của giải bóng rổ NBA của Mỹ, mô hình sẽ được xây dựng với những models được tính toán cẩn thận và cài đặt để có thể biến các dữ liệu rời rạc thành các đoạn văn bản với những thông tin quan trọng cần thiết mang tính nhất quán.

1. **Phương pháp nghiên cứu**

Mô hình Data-to-Text Generation with Content Selection and Planning được nghiên cứu và xây dựng dựa trên:

* Tập dữ liệu về kết quả thi đấu của giải bóng rổ NBA của Mỹ.
* Sử dụng ngôn ngữ Python phiên bản 2.7 cùng Pytorch phiên bản 0.3.1, CUDA 8.0.
* Sử dụng models Neural Content Planning (NPC) kết hợp với Conditional Copy (CC).