**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC CẦN THƠ**

**KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

**BỘ MÔN CÔNG NGHỆ THÔNG TIN**

**NIÊN LUẬN**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

**Đề tài**

**DATA-TO-TEXT GENERATION WITH CONTENT SELECTION AND PLANNING**

**Người hướng dẫn Sinh viên thực hiện**

**Lâm Nhựt Khang Đặng Trung Kiên**

**Mã số: B1706597**

**Khóa: K43**

***Cần Thơ, 12/2020***

**LỜI CẢM ƠN**

Trong thời gian nghiên cứu và hoàn thành đề tài “Data-to-Text Generation with Content Selection and Planning” của nhóm tác giả Ratish Puduppully - Li Dong - Mirella Lapata, em xin gửi lời cảm ơn chân thành và sâu sắc đến cô Lâm Nhựt Khang – người đã tận tình hướng dẫn, giúp đỡ em hoàn thành đề tài.

Học phần Niên luận ngành Công nghệ thông tin là một học phần bổ ích và thú vị trong chương trình giảng dạy. Học phần có tính thực tế cao khi có thể áp dụng ngay sau khi hoàn thành đề tài. Do đó, học phần giúp em có thể tìm hiểu những kiến thức mới, học hỏi thêm nhiều kinh nghiệm quý báu trong ngành học của mình. Nhưng do đây là lần đầu tiên tiếp xúc với việc nghiên cứu các mô hình và đào tạo máy học nên kiến thức bản thân còn nhiều hạn chế vì thế trong quá trình nghiên cứu và thực hiện đề tài này em không tránh khỏi nữa sai sót, kính mong nhận được những ý kiến đóng góp từ cô.

Em xin chân thành cảm ơn!

# MỤC LỤC

[MỤC LỤC I](#_Toc59410546)

[DANH MỤC HÌNH ẢNH II](#_Toc59410547)

[TÓM LƯỢC III](#_Toc59410548)

[GIỚI THIỆU 1](#_Toc59410549)

[**1.** **Mô tả bài toán** 1](#_Toc59410550)

[**2.** **Mục tiêu đề tài** 1](#_Toc59410551)

[**3.** **Phương pháp giải quyết** 1](#_Toc59410552)

[**4.** **Phương pháp nghiên cứu** 1](#_Toc59410553)

[CHƯƠNG 1: ĐẶC TẢ YÊU CẦU 3](#_Toc59410554)

[**1.** **Giới thiệu** 3](#_Toc59410555)

[**1.1.** **Dữ liệu** 3](#_Toc59410556)

[**1.2.** **Chuyển dữ liệu thành văn bản** 3](#_Toc59410557)

[**2.** **Mô hình sử dụng** 4](#_Toc59410558)

[**2.1.** **Mạng nơ-ron** 4](#_Toc59410559)

[**2.2.** **Chọn lọc nội dung và tạo kế hoạch (Content Selection and Planning)** 5](#_Toc59410560)

[CHƯƠNG 2: THIẾT KẾ VÀ CÀI ĐẶT GIẢI PHÁP 7](#_Toc59410561)

[**1.** **Giới thiệu tập dữ liệu** 7](#_Toc59410562)

[**2.** **Thiết kế và cài đặt giải pháp** 11](#_Toc59410563)

[**2.1.** **Mô hình xử lý dữ liệu** 11](#_Toc59410564)

[**2.2.** **Quá trình xử lý dữ liệu** 12](#_Toc59410565)

[CHƯƠNG 3: CÀI ĐẶT GIẢI PHÁP 17](#_Toc59410566)

[**1.** **Cài đặt các thư viện** 17](#_Toc59410567)

[**2.** **Tạo dữ liệu đầu vào** 18](#_Toc59410568)

[**3.** **Xây dựng mô hình** 20](#_Toc59410569)

[**4.** **Tạo văn bản** 22](#_Toc59410570)

[CHƯƠNG 4: ĐÁNH GIÁ VÀ KIỂM THỬ 27](#_Toc59410571)

[**1.** **Kiểm thử** 27](#_Toc59410572)

[**2.** **Kết quả** 27](#_Toc59410573)

[**3.** **Đánh giá** 27](#_Toc59410574)

[KẾT LUẬN 27](#_Toc59410575)

# DANH MỤC HÌNH ẢNH

[Hình 1. Thống kê dữ liệu được tạo ra từng năm 8](#_Toc59409980)

[Hình 2. Mô hình mạng nơ-ron nhân tạo (ANN) 9](#_Toc59409981)

[Hình 3. Mô hình multilayer perceptron 10](file:///F:\KENT\Sinh%20Viên\Niên%20luận%20ngành%20CNTT\MauLuanVan-NienLuan%20(2).docx#_Toc59409982)

[Hình 4. Mô hình hoạt động của Content Selection and Planning 11](#_Toc59409983)

[Hình 5. Tập dữ liệu huấn luyện (training) 12](#_Toc59409984)

[Hình 6. Tập dữ liệu thẩm định (validation) 13](#_Toc59409985)

[Hình 7. Tập dữ liệu thẩm định (validation) 13](#_Toc59409986)

[Hình 8. Một bản ghi đầu vào 16](#_Toc59409987)

[Hình 9. Cơ chế lựa chọn nội dung 18](#_Toc59409988)

[Hình 10. Đoạn trích của một kế hoạch 19](#_Toc59409989)

[Hình 11. Quá trình chuyển đổi dữ liệu ban đầu thành dữ liệu hệ thống 23](#_Toc59409990)

[Hình 12. Sử dụng các thuật toán để chuyển đổi dữ liệu 24](#_Toc59409991)

[Hình 13. Hàm xây dụng dữ liệu 24](#_Toc59409992)

[Hình 14. Hàm lấy dữ liệu đầu vào 24](#_Toc59409993)

[Hình 15. Dữ liệu đã được chuyển đổi 25](#_Toc59409994)

[Hình 16. Model NPC+CC 25](#_Toc59409995)

[Hình 17. Hàm xử lý xây dựng dữ liệu 26](#_Toc59409996)

[Hình 18. Quá trình xây dựng mô hình 27](#_Toc59409997)

[Hình 19. Nội dung kế hoạch 27](#_Toc59409998)

[Hình 20. Hàm lập kế hoạch 28](#_Toc59409999)

[Hình 21. Hàm lập kế hoạch nội dung với các bản ghi từ đầu vào với các chỉ số 29](#_Toc59410000)

[Hình 22. Nội dung đánh giá 30](#_Toc59410001)

[Hình 23. Hàm đánh giá 30](#_Toc59410002)

[Hình 24. Hàm xử lý tóm tắt 31](#_Toc59410003)

# TÓM LƯỢC

Data-to-Text Generation with Content Selection and Planning là mô hình tạo dữ liệu thành văn bản với tập dữ liệu có kích thước lớn và sử dụng mạng nơ-ron. Đồng thời mô hình Data-to-text còn kết hợp với việc chọn lựa nội dung và lập kế hoạch, đây là điểm nổi bật so với các mô hình khác. Để thực hiện được công việc chuyển dữ liệu thành văn bản với một kho dữ liệu bản ghi, mô hình được chia thành hai giai đoạn: tạo kế hoạch nội dung với thông tin nổi bật nào nên được đề cập đến và theo thứ tự nào; sao đó xây dựng văn bản dựa theo kế hoạch nội dung.

Data-to-Text Generation with Content Selection and Planning is data-to-text generation the use of large-scale datasets and neural network models. Data-to-Text also incorporates content selection and planning, this is the highlight compared to other models. To do the data-to-text generation the use of large-scale datasets, the model generate task into two stages: generate a content plan highlighting which information should be mentioned and in which order; then generate the document while taking the content plan into account.

# 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.

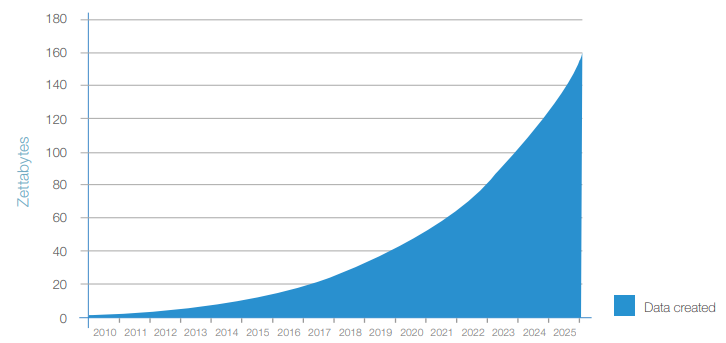
1. **Phương pháp nghiên cứu**

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).

# CHƯƠNG 1: ĐẶC TẢ YÊU CẦU

1. **Giới thiệu**
   1. **Dữ liệu**

Mỗi ngày trôi qua sẽ có một lượng thông tin khổng lồ được tạo ra khiến cho con người chúng ta không thể ghi nhớ hết tất cả. Vì thế, chúng ta cần phải lưu trữ lại thành những dữ liệu để có thể trích xuất khi cần thiết giúp cho việc xử lý lượng thông tin lớn một cách dễ dàng. Dữ liệu bao gồm hình ảnh, bản ghi âm, video, bảng thống kê…, có nội dung có thể chuyển sang dạng văn bản. Những dữ liệu này có là dạng phi ngôn ngữ thường được lưu dưới dạng một cấu trúc cố định như tệp cở sở dữ liệu quan hệ hoặc bảng tính.

Hình . Thống kê dữ liệu được tạo ra từng năm

* 1. **Chuyển dữ liệu thành văn bản**

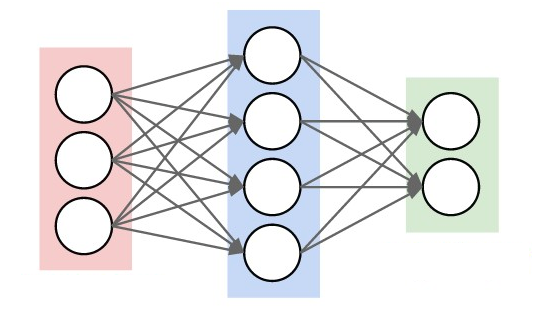
Trong sự phát triển ngày càng lớn về công nghệ trí tuệ nhân tạo dẫn đến nhu cầu tạo dữ liệu thành văn bản với tập dữ liệu quy mô lớn và mô hình mạng nơ-ron được đào tạo từ đầu đến cuối ngày càng cao. Do đó, đã có nhiều nhà phát triển đã triển khai nhiều công trình nghiên cứu vấn đề data-to-text nhầm đáp ứng nhu cầu trên. Nhưng những nghiên cứu đó lại không có bất kỳ lựa chọn hoặc lập kế hoạch nội dung nào hoặc lựa chọn nội dung và lập kế hoạch. Vì thế mục đích của bài báo là trình bày kiến trúc mạng nơ-ron kết hợp lựa chọn nội dung có kế hoạch.

Để có thế thực thi được phương pháp kết hợp mạng nơ-ron và việc chọn nội dung dùng cho việc lập kế hoạch, nhóm tác giả đã chia thành hai nhiệm vụ chính: Lập kế hoạch với nội dung quan trọng được chọn ra từ dữ liệu đồng thời sắp xếp thứ tự dựa vào nội dung kế hoạch; và tạo tài liệu văn bản dựa trên kế hoạch ban đầu. Từ đó, đó tạo ra những lợi thế cho việc tạo ra tài liệu nhiều câu:

* Đại diện cho một nhóm dữ liệu cấp cao của cấu trúc tài liệu cho phép bộ giải mã tập trung vào các nhiệm vụ. Từ đó làm cho việc lập kế hoạch và hiện thực bề mặt hóa dễ dàng hơn.
* Làm cho quá trình tạo dữ liệu thành tài liệu văn bản dễ hiểu hơn bằng cách tạo ra một đại diện trung gian.
* Giảm dư thừa trong đầu ra vì có ít khả năng nội dung kế hoạch chứa cùng một thông tin dữ liệu ở nhiều nơi.

1. **Mô hình sử dụng**
   1. **Mạng nơ-ron**

Mạng nơ-ron nhân tạo (Artificial Neural Networks - ANN) được xây dựng, mô phỏng và mô hình hóa dựa trên sự hoạt động mạng lưới thần kinh não người bằng  một chuỗi các [thuật toán](https://vietnambiz.vn/giao-dich-thuat-toan-algorithmic-trading-la-gi-nhung-dac-diem-can-luu-y-20191205025413985.htm)được đưa ra để nỗ lực tìm kiếm các [mối quan hệ](https://vietnambiz.vn/tuong-quan-dong-bien-positive-correlation-la-gi-nhung-dac-diem-can-luu-y-20191227113029107.htm) cơ bản trong một tập hợp dữ liệu. Một mạng nơ-ron nhân tạo là tập hợp của nhiều nơ-ron nhân tạo được nối kết với nhau, và xử lý thông tin bằng cách truyền theo các kết nối và tính giá trị mới tại các nút.



Hình . Mô hình mạng nơ-ron nhân tạo (ANN)

Mạng nơ-ron là sự kết hợp của của các tầng perceptron hay còn được gọi là perceptron đa tầng (multilayer perceptron) là mạng truyền thẳng, tất cả các nơron chỉ có thể được kết nối với nhau theo một hướng tới một hay nhiều các nơron khác trong lớp kế tiếp (trừ các nơ-ron ở output layer). Multilayer perceptron gồm có 3 tầng: tầng vào (input layer), tầng ẩn (hidden layer), tầng ra (output layer). Mỗi nút trong mạng là một sigmoid nơ-ron nhưng hàm kích hoạt của chúng có thể khác nhau thường cùng dạng với nhau.

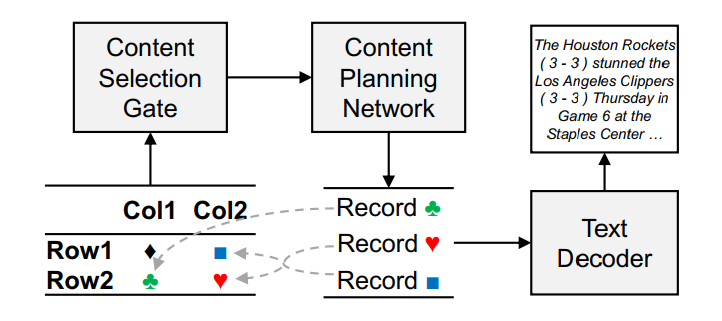
Hình . Mô hình multilayer perceptron

Mạng lưới thần kinh nhân tạo đang mở đường cho các ứng dụng thay đổi cuộc sống được phát triển để sử dụng trong tất cả các lĩnh vực của nền kinh tế. Các nền tảng trí tuệ nhân tạo được xây dựng trên mạng ANNs đang phá vỡ những cách làm truyền thống. Từ việc dịch các trang web sang các ngôn ngữ khác đến việc có một cửa hàng tạp hóa trợ lý ảo trực tuyến để trò chuyện với chatbot để giải quyết vấn đề, các nền tảng AI đang đơn giản hóa các giao dịch và làm cho các dịch vụ có thể truy cập được với chi phí không đáng kể.

* 1. **Chọn lọc nội dung và tạo kế hoạch (Content Selection and Planning)**

Kiến trúc tổng thể của mô hình lập lựa chọn và lập kế hoạch nội dung được chia thành hai giai đoạn chính:

* Giai đoạn 1: Lựa chọn nội dung và lập kế hoạch hoạt động dựa trên các bản ghi đầu vào của cơ sở dữ liệu, từ đó tạo ra một kế hoạch nội dung xác định những bản ghi nào sẽ được diễn đạt bằng lời trong tài liệu và theo thứ tự nào.
* Giai đoạn 2: Tạo văn bản tạo ra văn bản đầu ra với kế hoạch nội dung làm đầu vào; ở mỗi bước giải mã, mỗi thành phần trong mô hình chú ý đến các vector được biểu diễn ở từng bản ghi trong kế hoạch nội dung.



Hình . Mô hình hoạt động của Content Selection and Planning

Content Selection and Planning là một hệ thống xử lý dữ liệu vô cùng phức tạp, dữ liệu được xử lý trải qua nhiều quá trình tính toán khác nhau bao gồm:

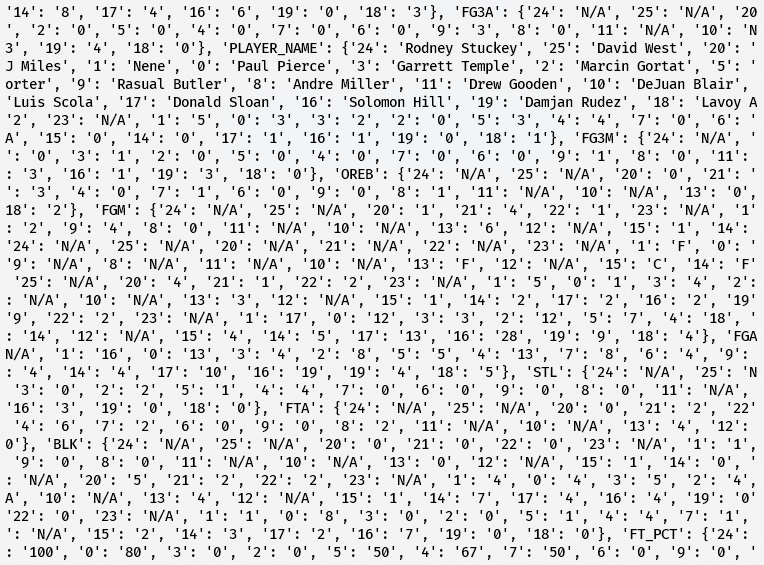
* Bộ mã hóa bản ghi: tại đây dữ liệu sẽ được chuyển hóa thành các dạng véc-tơ, sau đó sẽ được các multilayer perceptron trong mô hình sử dụng.
* Cổng lựa chọn nội dung: Từ các véc-tơ thu được ở bộ mã hóa bản ghi, các nhà nghiên cứu thực hiện các phép tính để thu được những thông tin tối ưu, tránh gây trùng lập.
* Lập nội dung kế hoạch: Lập bản đồ văn bản trong phần tóm tắt dựa vào các thực thể ở đầu vào (mối quan hệ giữa các thực thể). Mỗi kế hoạch là một chuỗi các con trỏ với mỗi mục nhập trỏ đến các véc-tơ và sử dụng bộ giải mã LSTM để tạo mã thông báo.
* Tạo văn bản: Mã hóa kế hoạch nội dung thành các đơn vị LSTM, dựa vào đó sử dụng mạng nơ-ron tuần hoàn sử dụng các đơn vị LSTM và LSTM sao chép để tăng hiệu năng xử lý.
* Đào tạo và suy luận: Được đào tạo để tối đa hóa khả năng ghi nhật kí của mô hình. Đối với mỗi giai đoạn, họ sử dụng tìm kiếm chùm tia để đạt được kết quả tốt nhất.

# CHƯƠNG 2: THIẾT KẾ VÀ CÀI ĐẶT GIẢI PHÁP

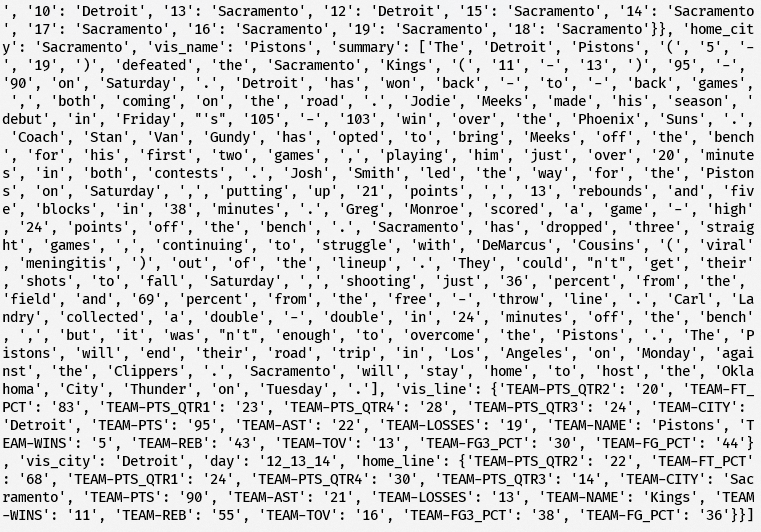
1. **Giới thiệu tập dữ liệu**

Mô hình sử dụng tập dữ liệu Rootowire:

<https://github.com/harvardnlp/boxscore-data.git>

Tập dữ liệu Rotowire là các bản tóm tắt các trận đấu bóng rổ NBA (do con người lập nên) được trình bài một cách hợp lý dựa vào sự canh chỉnh điểm số thành từng dòng tương ứng. Rotowire có 4853 bản tóm tắt riêng biệt, bao gồm các trận đầu NBA được chơi từ 1/1/2014 đến 29/3/2017, trong đó có một số trận đấu có nhiều bản tóm tắt. Các bản tóm tắt đã được chia ngẫu nhiên thành những tệp JSON bao gồm: 3398 tập huấn luyện (training), 727 tập thẩm định (validation) và 728 tập kiểm thử (test).

Hình . Tập dữ liệu huấn luyện (training)

Hình . Tập dữ liệu thẩm định (validation)



Hình . Tập dữ liệu thẩm định (validation)

Trong đó có các trường dữ liệu bao gồm:

* home\_name: tên đội chủ nhà (unicode).
* home\_city: Tên thành phố đội chủ nhà (unicode).
* vis\_name: tên đội khách (unicode).
* vis\_name: tên thành phố của đội khác (unicode).
* day: ngày thi đầu được định dạng %MM\_%DD\_%YY (unicode).
* summary: Mã của tóm tắt trận đầu.
* home-line: Dòng ghi nhận điểm của đội chủ nhà.
* vis-line: Dòng ghi nhận điểm của đội khách.
* box-core: Điểm hộp số.

Đối với trường dữ liệu Line-score (điểm theo dòng) bao gồm các thuộc tính như sau:

* TEAM-NAME: Tên đội bóng (unicode).
* TEAM-CITY: Thành phố cả đội bóng (unicode).
* TEAM-AST: Số lần hỗ trợ của đội bóng (integer as unicode).
* TEAM-FG3\_PCT: Tỷ lệ 3 con trỏ theo nhóm (integer as unicode).
* TEAM-FG\_PCT: Tỷ lệ bàn thắng được thực hiện bởi đội bóng (integer as unicode).
* TEAM-FT\_PCT: Tỷ lệ 3 điểm được thực hiện bởi đội bóng (integer as unicode).
* TEAM\_LOSSES: Đội thua (integer as unicode).
* TEAM-PTS: Tổng điểm của đội bóng (integer as unicode).
* TEAM-PTS\_QTR1: Điểm trong hiệp đầu tiên của đội bóng (integer as unicode).
* TEAM-PTS\_QTR2: Điểm trong hiệp thứ hai của đội bóng (integer as unicode).
* TEAM-PTS\_QTR3: Điểm trong hiệp thứ ba của đội bóng (integer as unicode).
* TEAM-PTS\_QTR4: Điểm trong hiệp thứ tư của đội bóng (integer as unicode).
* TEAM-REB: Tổng điểm Rebounds của đội (integer as unicode).
* TEAM-TOV: Tổng điểm Turnovers (integer as unicode).
* TEAM-WINS: Đội bóng thắng cuộc (integer unicode).

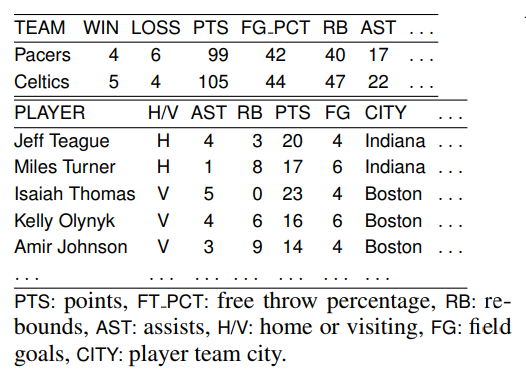
Đối với trường dữ liệu Box-core gồm có các thuộc tính sau:

* AST – cầu thủ hỗ trợ (row\_number -> integer as unicode).
* BLK - cầu thủ vị trí blocks (row\_number -> integer as unicode).
* DREB – cầu thủ ngăn chặn rebounds (row\_number -> integer as unicode).
* FG3A – cầu thủ tấn công 3 điểm (row\_number -> integer as unicode).
* FG3M – cầu thủ tạo 3 điểm (row\_number -> integer as unicode)
* FG3\_PCT – Tỷ lệ cầu thủ tạo 3 điểm (row\_number -> integer as unicode).
* FGA – cầu thủ với mục tiêu tấn công (row\_number -> integer as unicode).
* FGM – cầu thủ với mục tiêu ghi điểm (row\_number -> integer as unicode).
* FG\_PCT – Tỷ lệ cầu thủ ghi điểm (row\_number -> integer as unicode).
* FIRST\_NAME – Tên cầu thủ (row\_number -> unicode).
* FTA – Cầu thủ ném phạt (row\_number -> integer as unicode).
* FTM – Cầu thủ thực hiện ném phạt (row\_number -> integer as unicode).
* FT\_PCT – Tỷ lệ phần trăm cầu thủ ném phạt (row\_number -> integer as unicode).
* MIN – Số phút cầu thủ đã ra sân (row\_number -> integer as unicode).
* OREB – Cầu thủ phản hồi rebounds (row\_number -> integer as unicode).
* PF – Cầu thủ phạm lỗi cá nhân (row\_number -> integer as unicode).
* PLAYER\_NAME – Tên đầy đủ của cầu thủ (row\_number -> integer as unicode).
* PTS – Điểm của cầu thủ (row\_number -> integer as unicode).
* REB – Tổng điểm rebounds của cầu thủ (row\_number -> integer as unicode).
* SECOND\_NAME – Tên lót của cầu thủ (row\_number -> integer as unicode).
* START\_POSITION – Vị trí cầu thủ (row\_number -> unicode).
* STL – Cầu thủ lấy bóng (row\_number -> integer as unicode).
* TEAM\_CITY – Thành phố của đội bóng (row\_number -> unicode).
* TO – Cầu thủ turnovers (row\_number -> integer as unicode).

1. **Thiết kế và cài đặt giải pháp**
   1. **Mô hình xử lý dữ liệu**

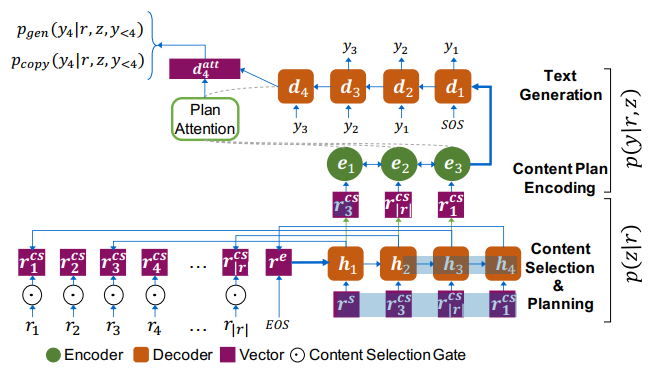
Đầu vào của mô hình là một bảng các bảng ghi (hình ). Mỗi bản ghi sẽ được gọi là rj, được biểu diễn dưới dạng , trong mỗi bản ghi gồm có 4 thuộc tính:

* Loại thuộc tính (rj,1 ; …; LOSS, CITY): là thuộc tính tên từng cột trong bản ghi bao gồm các thuộc tính như WIN, LOSS, PTS…
* Đối tượng (rj,2 ;…; Pacers, Miles Turner): là những thuộc tính liên quan đến đội bóng thi đầu và cầu thủ của đội bóng.
* Dữ liệu (rj,3 ;…; 11; Indiana): là thuộc tính liên quan đến thành phố và tổng ba chỉ số AST, RB, FG của cầu thủ.
* Dữ liệu (rj,4 ;…, H): là thuộc tính chỉ ra là đội khách hay đội nhà (cột H/V).



Hình . Một bản ghi đầu vào

Đầu ra của mô hình là một tài liệu được gọi là y bao gồm các thành phần y = y1 · · · với là độ dài của tài liệu.

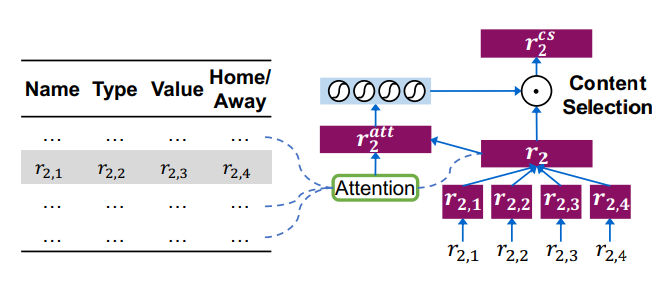
Hình. Mô hình lựa chọn và lập kế hoạch nội dung

Từ dữ liệu đầu vào là các bản ghi đã được lựa chọn nội dung để mã hóa dưới dạng các véc-tơ. Các véc-tơ sẽ được đi qua cổng tuyển chọn nội dung để tìm lọc ra các nội dung hữu ích, tránh sự trùng lập nội dung không cần thiết tạo thành một bản ghi mới. Tiếp theo, bản ghi mới này sẽ được sắp xếp theo trình tự kế hoạch tương ứng với thứ tự các véc-tơ trong bản ghi ban đầu. Đồng thời tại bước này, sử dụng bộ giải mã LSTM để tạo ra mã thông báo tương ứng với các vị trí trong đầu vào. Sau khi đã đồng bộ được thứ tự các véc-tơ, quá trình sẽ giải mã các thông tin văn bản dựa trên mạng nơ-ron tuần hoàn với các đơn vị LSTM, đó là sự đối chiếu của bản ghi với sự dự đoán từ trước. Đồng thời, mô hình cũng tăng cường thêm với bộ giải mã với cơ chế sao chép trực tiếp từ các phần giá trị của bản ghi trong kế hoạch nội dung.

* 1. **Quá trình xử lý dữ liệu**
     1. **Bộ mã hóa bản ghi (Record Encoder)**
* Đầu vào của mô hình là một bảng các bản ghi không có thứ tự, mỗi bản ghi được biểu diễn dưới dạng các tính năng *.*
* Công việc tiếp theo được thực hiện đó là nhúng các tính năng thành véc-tơ và sử dụng multilayer perceptron để lấy một biểu diễn véc-tơ rj cho mỗi bản ghi theo công thức:

Trong đó “ ; ” là chỉ phép nối véc-tơ, với n là tham số, ReLU là chức năng kích hoạt chỉnh lưu.

* + 1. **Cổng lựa chọn nội dung (Content Selection Gate)**
* Ngữ cảnh của một bản ghi rất quan trọng trong việc xác định giá trị của bản ghi đó có hữu ích hay không so với các bản ghi khác trong bảng. Vì thế mô hình sử dụng cơ chế cổng lựa chọn nội dung nhằm xác định tính hữu dụng của bản ghi trong bảng theo từng ngữ cảnh.
* Bước đầu tiên, mô hình tính điểm chú ý () qua đầu vào bảng, và sử dụng chúng để thu được một véc-tơ lưu ý () cho mỗi bản ghi , theo công thức:

Trong đó , là ma trận tham số, = 1

Hình . Cơ chế lựa chọn nội dung

* Tiếp theo cơ chế lựa cổng lựa chọn nội dung được áp dụng cho các véc-tơ . Đồng thời tìm được bản ghi đại diện mới được xác định dựa theo công thức:

Trong đó ⊙ biểu thị phép nhân khôn ngoan nguyên tố và cổng ∈ kiểm soát lượng thông tin xuất ra từ .

Nói cách khác, mỗi phần tử trong  được cân nhắc bởi phần tử tương ứng của cổng lựa chọn nội dung .

* + 1. **Kế hoạch nội dung (**Content Planning**)**
* Kế hoạch nội dung được xây dựng bằng cách lập bảng đồ văn bản trong phần tóm tắt vào các thực thể trong bảng đầu vào là giá trị và kiểu (mối quan hệ giữa các thực thể).
* Một kế hoạch là một chuỗi các con trỏ với mỗi mục nhập trỏ đến bản ghi đầu vào .



Hình . Đoạn trích của một kế hoạch

* Cho *z = z1 …*  biểu thị trình tự lập kế hoạch nội dung với mỗi đơn vị *zk* trỏ đến một bản ghi đầu vào và xác suất *p(z|r)* được phân rã được tính theo công thức:

Trong đó

* Vì các mã thông báo đầu ra của giai đoạn lập kế hoạch nội dung tương ứng với các vị trí trong trình tự đầu vào nên mô hình sử dụng mạng con trỏ (Pointer Nerwork). Từ quá trình này, mô hình sẽ chú ý mã thông báo trình tự của các dữ liệu đầu vào thay vì phải tạo ra các biểu diễn có trọng số của các dữ liệu đầu vào đã mã hóa.
* Tiếp đến, mô hình được áp dụng bộ giải mã LSTM để tạo mã tương ứng với các vị trí ở đầu vào. Tại bước này, trạng thái đầu tiên của bộ giải mã được khởi tạo bởi công thức là giá trị trung bình của véc-tơ bản ghi.
* Mô hình được xây dựng bằng cách quan tâm đến giá trị các bản ghi đầu vào. Với mỗi bản ghi tại bước *k*, gọi *hk* là trạng thái ẩn của LTSM:

Trong đó xác suất được chuẩn hóa thành 1 và là các tham số. Khi trỏ đến bản ghi , mô hình sử dụng véc-tơ tương ứng làm đầu vào của đơn vị LSTM tiếp theo trong bộ giải mã.

* + 1. **Tạo văn bản (Text Generation)**
* Xác suất xuất ra văn bản *y* có điểu kiện có kế hoạch nội dung *z* và bảng đầu vào *r* được mô hình hóa như sau:

Trong đó đó , mô hình sử dụng kiến trúc bộ mã hóa và giải mã với cơ chế chú ý để tính .

* Bước đầu tiên của quá trình, mô hình sẽ mã hóa kế hoạch nội dung *z* thành bằng LSTM hai chiều. Bởi vì kế hoạch nội dung là một chuỗi các bản ghi đầu vào, vì thế mô hình cung cấp trực tiếp các véc-tơ bản ghi tương ứng làm đầu vào cho các đơn vị LTSM, chúng chia sẻ bản ghi mã hóa với trạng thái đầu tiên.
* Bộ giải mã văn bản dựa trên mạng nơ-ron tuần hoàn làm việc với các đơn vị LSTM được khởi tạo với các trạng thái ẩn ở bước cuối cùng trong bộ mã hóa. Lúc giải mã bước *t*, đầu vào của các đơn vị LSTM là sự nhúng dự đoán trước đó . Gọi là trạng thái ẩn của đơn vị LTSM thứ *t*. Xác suất dự đoán từ vựng đầu ra được tính bằng công thức:

Trong đó , , ,, là các tham số và là đầu ra kích thước từ vựng.

* Mô hình được tăng cường thêm cho bộ giải mã với cơ chế sao chép, tức là khả năng sao chép các từ trực tiếp từ các phần giá trị của bản ghi trong kế hoạch nội dung, được thử nghiệm với phương pháp joint (Gu et al. 2016) và các phương pháp sao chép có điều kiện (Gulcehre et al. 2016). Cụ thể, với biến cho mỗi bước thời gian để cho biết liệu mã thông báo đã được sao chép () hay chưa được sao chép (). Xác suất tạo ra được tính bằng công thức:

Trong đó bị loại bỏ ra ngoài lề.

* Với phương pháp sao chép joint, xác suất sao chép từ các giá trị bản ghi và tạo từ vựng được chuẩn hóa toàn cầu dựa trên công thức:

Trong đó chỉ ra rằng có thể sao chép từ , .

* Với các phương pháp sao chép có điều kiện, biến được tính như một cổng chuyển đổi, sau đó được sử dụng để lấy xác suất đầu ra:

Trong đó và , , là các tham số.

* + 1. **Huấn luyện và suy luận (Training and Inference)**
* Mô hình được đào tạo để tối đa hóa khả năng ghi nhật kí của kế hoạch nội dung cho các bảng ghi *r* và văn bản xuất ra có giá trị cho kế hoạch nội dung và các bảng ghi:

Trong đó *D* đại diện cho các ví dụ đào tạo (bản ghi đầu vào, kế hoạch nội dung, và tóm tắt trò chơi).

* Trong quá trình suy luận, đầu ra cho đầu vào *r* được dự đoán bằng công thức:

Trong đó và đại diện cho các ứng cử viên của kế hoạch nội dung và văn bản đầu ra tương ứng. Đối với mỗi giai đoạn, mô hình sử dụng tìm kiếm chùm tia để thu được kết quả tốt nhất

# CHƯƠNG 3: CÀI ĐẶT GIẢI PHÁP

1. **Cài đặt các thư viện**

Giống như các chương trình máy học khác, Data-to-Text Generation with Content Selection and Planning cũng cần có môi trường hoạt động bằng cách cài đặt các thư viện:

* Numpy
* Nltk
* Tensoflow

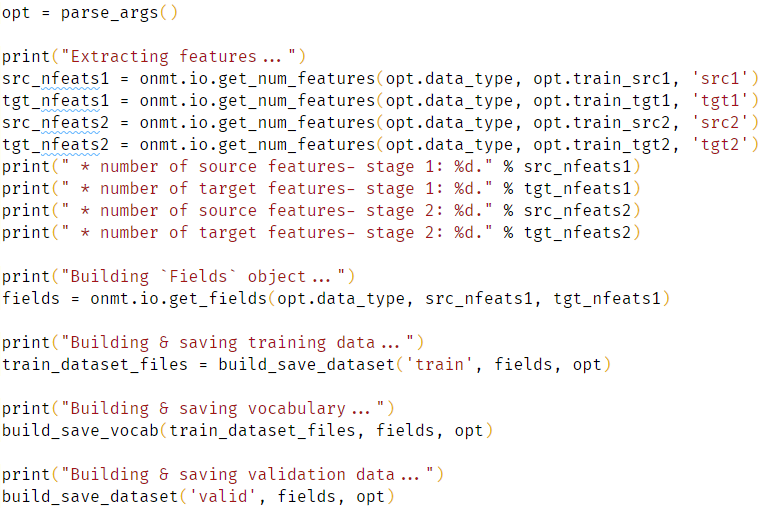
Ngoài ra, mô hình còn sử dụng python phiên bản 2.7 và Pytorch phiên bản 0.3.1, CUDA 8.0. Bên cạnh đó còn có các model được thiết lập sẵn như: AudioEncoder, Conv2Conv, ConvMultiStepAttention, Gate, CopyGenerator, GlobalAttention, ImageEncoder, PointerAttention, SRU, StackedRNN, StructuredAttention, Tranformer, UtilClass, WeightNorm.

1. **Tạo dữ liệu đầu vào**

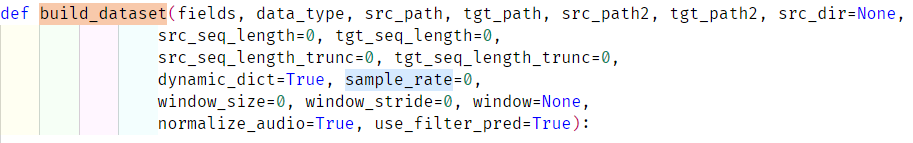
Đầu tiên phải khai báo biến $BASE để làm biến môi trường trong suốt quá trình cài đặt mô hình. Đồng thời tạo thư mục preprocess để lưu trữ dữ liệu đã chuyển đổi thành các véc-tơ.

Hình . Quá trình chuyển đổi dữ liệu ban đầu thành dữ liệu hệ thống

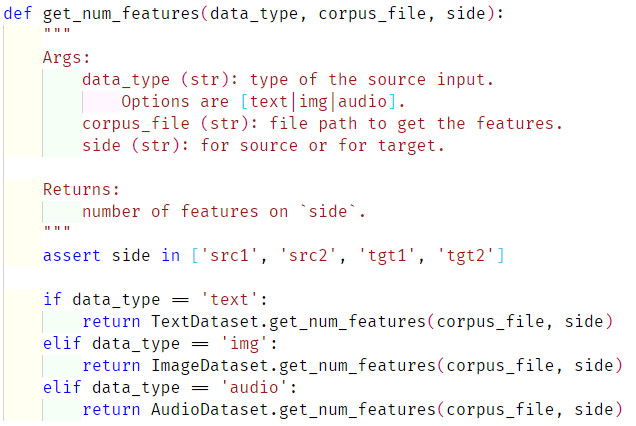
Dựa vào những thuật toán đã được trình bày trong thư mục onmt.io của dự án, mô hình sẽ áp dụng những thuật toán này vào nhiệm vụ mã hóa dữ liệu.



Hình . Sử dụng các thuật toán để chuyển đổi dữ liệu

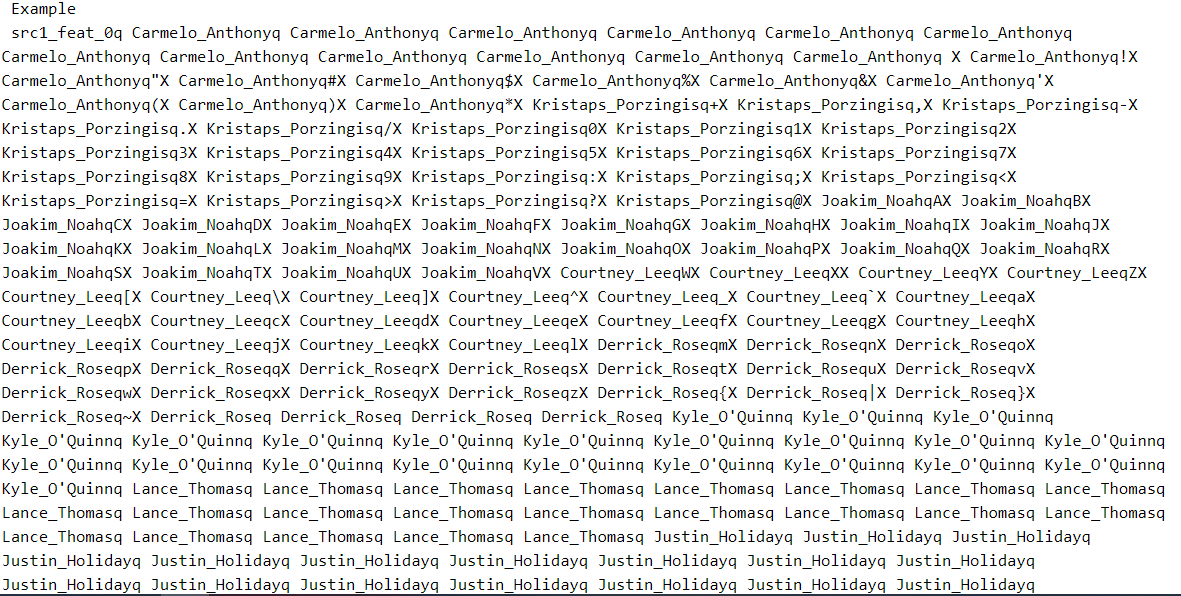


Hình . Hàm xây dụng dữ liệu



Hình . Hàm lấy dữ liệu đầu vào

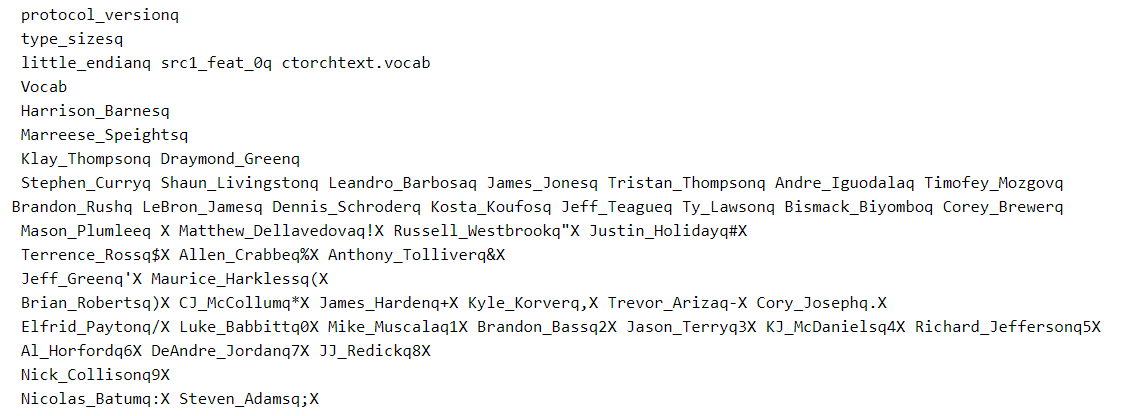
Sau khi thực hiện kết thúc quá trình mã hóa dữ liệu, tập dữ liệu sẽ được chia thành ba tập tin nằm trong thư mục preprocess: roto.train.1, roto.valid.1, roto.vocab.



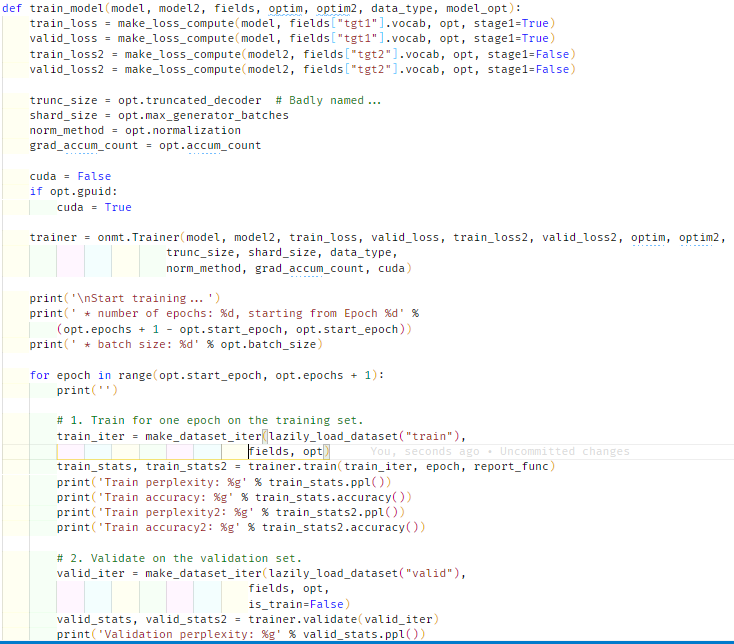
Hình . Dữ liệu đã được chuyển đổi

1. **Xây dựng mô hình**

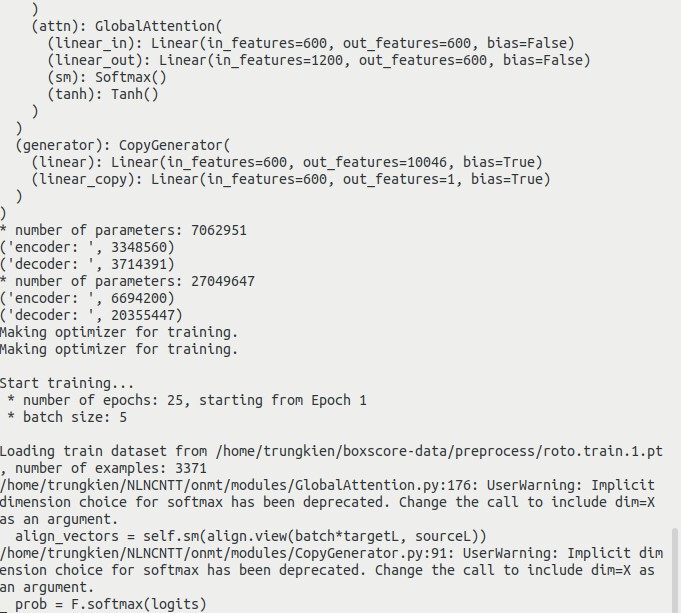
Sau khi đã có dữ liệu đã được mã hóa, mô hình sẽ được xây dụng dựa trên model kết hợp giữa Neural Content Planning (NPC) + Conditional Copy (CC).



Hình . Model NPC+CC



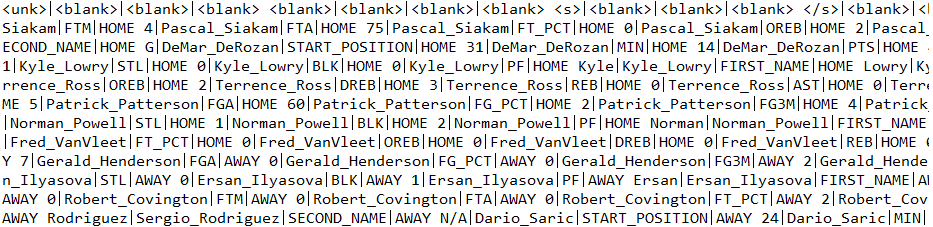
Hình . Hàm xử lý xây dựng dữ liệu



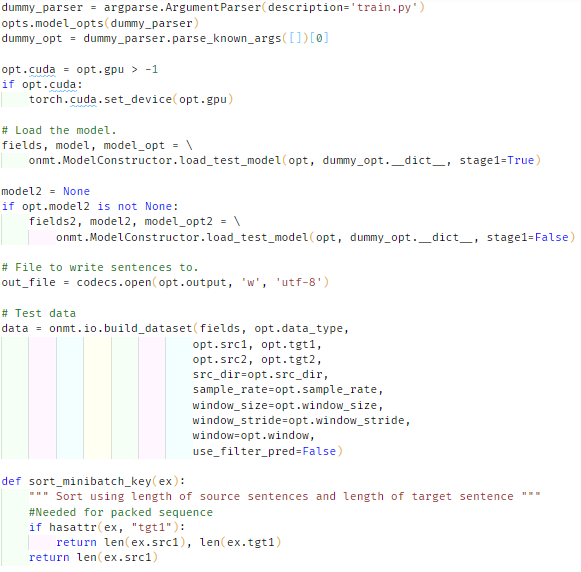
Hình . Quá trình xây dựng mô hình

1. **Tạo văn bản**

Để tạo được văn bản, chương trình cần tạo ra kế hoạch nội dung dựa vào mô hình đã được xây dựng bằng model NPC+CC.

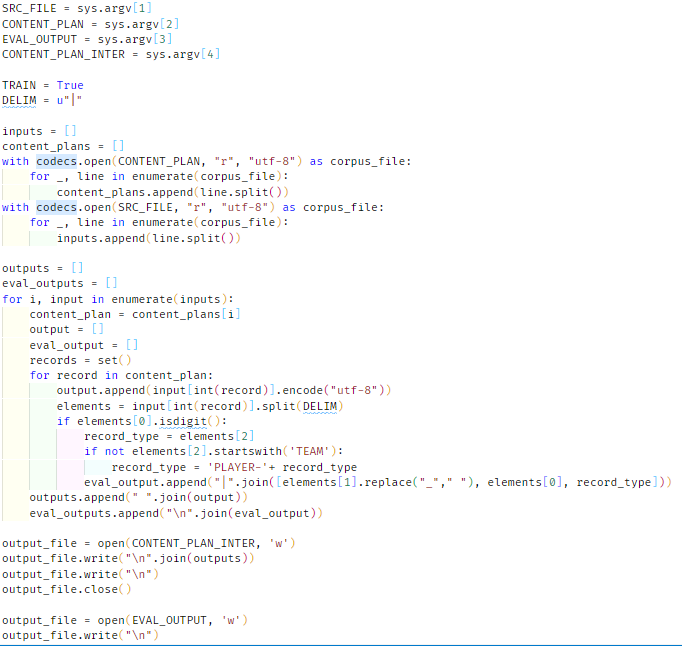


Hình . Nội dung kế hoạch



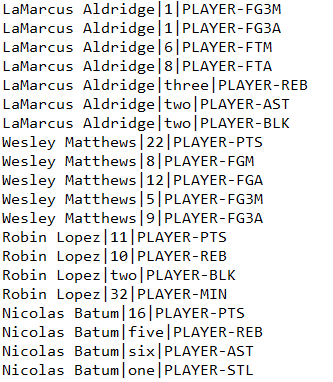
Hình . Hàm lập kế hoạch

Tiếp theo sẽ lập kế hoạch nội dung với các bản ghi từ đầu vào với các chỉ số.

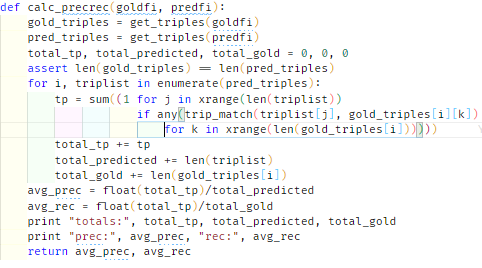


Hình . Hàm lập kế hoạch nội dung với các bản ghi từ đầu vào với các chỉ số

Sau đó sẽ đánh giá độ chính xác của kế hoạch nội dung trong giai đoạn đầu tiên.

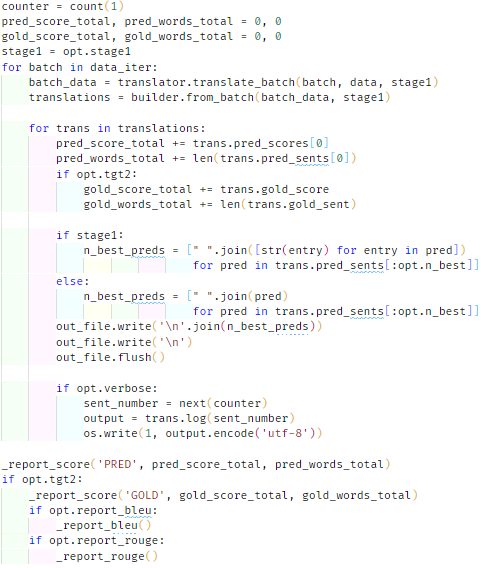


Hình . Nội dung đánh giá



Hình . Hàm đánh giá

Bước cuối cùng trong quá trình là tóm tắt dữ liệu đầu ra thành văn bản.



Hình . Hàm xử lý tóm tắt

# CHƯƠNG 4: ĐÁNH GIÁ VÀ KIỂM THỬ

1. **Kiểm thử**

* Xử lý được dữ liệu ban đầu trở thành dữ liệu dưới dạng các vec-tơ bằng cách mã hóa chúng.
* Xây dựng mô hình với tập dữ liệu đã được mã hóa theo model NCP+CC.
* Chuyển đổi dữ liệu từ mô hình xây dừng thành văn bản thông qua các model tính toán của hệ thống.

1. **Kết quả**

* Chuyển hóa được dữ liệu đầu vào thành dữ liệu véc-tơ thông qua mã hóa.
* Tạo được nội dung kế hoạch xử lý dữ liệu
* Thu được những dữ liệu quan trọng, đã qua sàng lọc xử lý.

1. **Đánh giá**

Bước đầu đã hoàn thành yêu cầu được đề ra đó là xử lý dữ liệu đầu vào, mã hóa và tuyển chọn theo kế hoạch được vạch ra. Nhưng do một số yếu tố khách quan về môi trường hoạt vì vậy mô hình chưa thể được xây dựng để có thể dựa vào đó chuyển dữ liệu đã tuyển chọn thành dạng văn bản tóm tắt.

**KẾT LUẬN**

Từ dữ liệu thu thập được từ giải thi đấu bóng rổ NBA, chương trình đã có thể chuyển từ dữ liệu ban đầu thành dạng véc-tơ để có thể xử lý, đồng thời cũng đã mã hóa thành các đầu vào. Từ đó đã tạo ra được kế hoạch nội dung để có thể xây dựng và chuyển hóa dữ liệu thành văn bản. Tuy nhiên do quá trình thực hiện phát sinh nhiều vấn đề ở hệ thống nên việc xây dựng mô hình vẫn chưa thể hoàn thiện.