Đại học Bách Khoa Hà Nội
Viện Công nghệ thông tin và Truyền thông

ĐỒ ÁN MÔN HỌC

Phân loại văn bản tin tức

Môn: Học máy
Mã lớp: 95090

Giảng viên hướng dẫn: Nguyễn Nhật Quang

Nhóm thực hiện

Phan Ngọc Lân - 20142505
Nguyễn Duy Mạnh - 20142857

Hà Nội, 5/2017

Mục lục

1. Bài toán

1.1. Mô tả và yêu cầu

Phân loại tin tức là bài toán có ứng dụng rộng rãi không chỉ trong các ứng dụng và dịch vụ tin tức (news aggregator), mà còn có ảnh hưởng trong việc sàng lọc dữ liệu hay tìm kiếm thông tin nói chung. Việc phân loại dặc biệt hữu ích trong các hệ thống thu thập tin tức, trong đó các tin có thể đến từ nhiều nguồn với các cách phân loại khác nhau, không rõ ràng hoặc thậm chí không có phân loại trước. Bài toán

Nhóm phát biểu bài toán phân loại văn bản tin tức như sau:

Đầu vào : 1 văn bản tin tức (bao gồm tiêu đề và nội dung) và 1 số các chủ đề có sẵn.

Đầu ra : Chủ đề phù hợp nhất với văn bản được đưa vào.

Ví dụ : Với tập chủ đề (Sports, World, Business, Sci/Tech),

Input	Output
Tiêu đề : Fears for T N pension after talks **Nội dung :** Unions representing workers at Turner Newall say they are 'disappointed' after talks with stricken parent firm Federal Mogul.	Business
Tiêu đề : The Race is On: Second Private Team Sets Launch Date for Human Spaceflight (SPACE.com) **Nội dung :** SPACE.com - TORONTO, Canada -- A second team of rocketeers competing for the 36.10 million Ansari X Prize, a contest for privately funded suborbital space flight, has officially announced the first launch date for its manned rocket.	Sci/Tech

Bài toán tương tự nhưng khác với bài toán xây dựng chủ đề (topic modeling), trong đó không có 1 tập chủ đề xác định mà thay vào đó lời giải bao gồm 1 tập các chủ đề được tạo ra với số lượng cho trước.

Lời giải cần thỏa mãn 1 số yêu cầu:

- Có độ chính xác, độ hợp lý cao. Trong 1 số trường hợp, có thể chấp nhận 1 lượng sai sót nhất định, nhất là khi nội dung văn bản không thực sự nằm gọn trong 1 chủ

đề cho sẵn. Tuy nhiên, lời giải không thể quá "bất hợp lý" (ví dụ phân loại văn bản tài chính vào thể thao,...).

- Có khả năng thích ứng cao với các văn bản mới.
- Thời gian dự đoán tương đối tốt. Mặc dù tốc độ luôn là 1 ưu tiên, đa số các trường hợp sử dụng của thuật toán nằm trong phần xử lý tại server (khi tổng hợp tin tức trước khi xuất bản), do đó lượng tài nguyên có thể được coi là khá lớn.

1.2. Một số cách tiếp cận phổ biến

Cách tiếp cận phổ biến nhất với bài toán là sử dụng các giải thuật dựa trên xác suất, với điển hình là giải thuật Naive Bayes [1]. Ngoài ra, 1 giải thuật mới được áp dụng gần đây là LDA (Latent Dirichlet Allocation) cũng cho kết quả rất khả quan.

Một cách tiếp cận thứ hai sử dụng các kĩ thuật xử lý ngôn ngữ tự nhiên như các bộ đoán nhận, bộ dịch [2].

Một số nghiên cứu cũng có sử dụng các giải thuật học dựa trên trọng số như SVM hay mạng neuron.

1.3. Hướng giải quyết

Nhóm quyết định sử dụng một mô hình mạng neuron để giải quyết bài toán, với ưu điểm là khả năng chịu nhiễu và làm việc với dữ liệu thô tốt, đồng thời có thời gian chạy ngắn sau giai đoạn huấn luyện. Cụ thể, mô hình sẽ tích hợp mạng hồi quy xử lý từng kí tự trong văn bản để tiến hành phân loại.

Để cài đặt hệ thống, nhóm sử dụng ngôn ngữ lập trình Python và Keras [3], một thư viện Deep Learning mã nguồn mở. Keras cho phép sử dụng 1 trong 2 nền tảng tính toán là TensorFlow (Google) và Theano (cả 2 đều hỗ trợ tăng tốc bằng GPU), cài đặt sẵn các mô hình mạng neuron dưới dạng các lớp có tính module hóa cao và độ tối ưu tốt. Quá trình cài đặt thuật toán do đó khá ngắn gọn, cho phép nhóm tập trung vào thiết kế mô hình và xử lý dữ liệu.

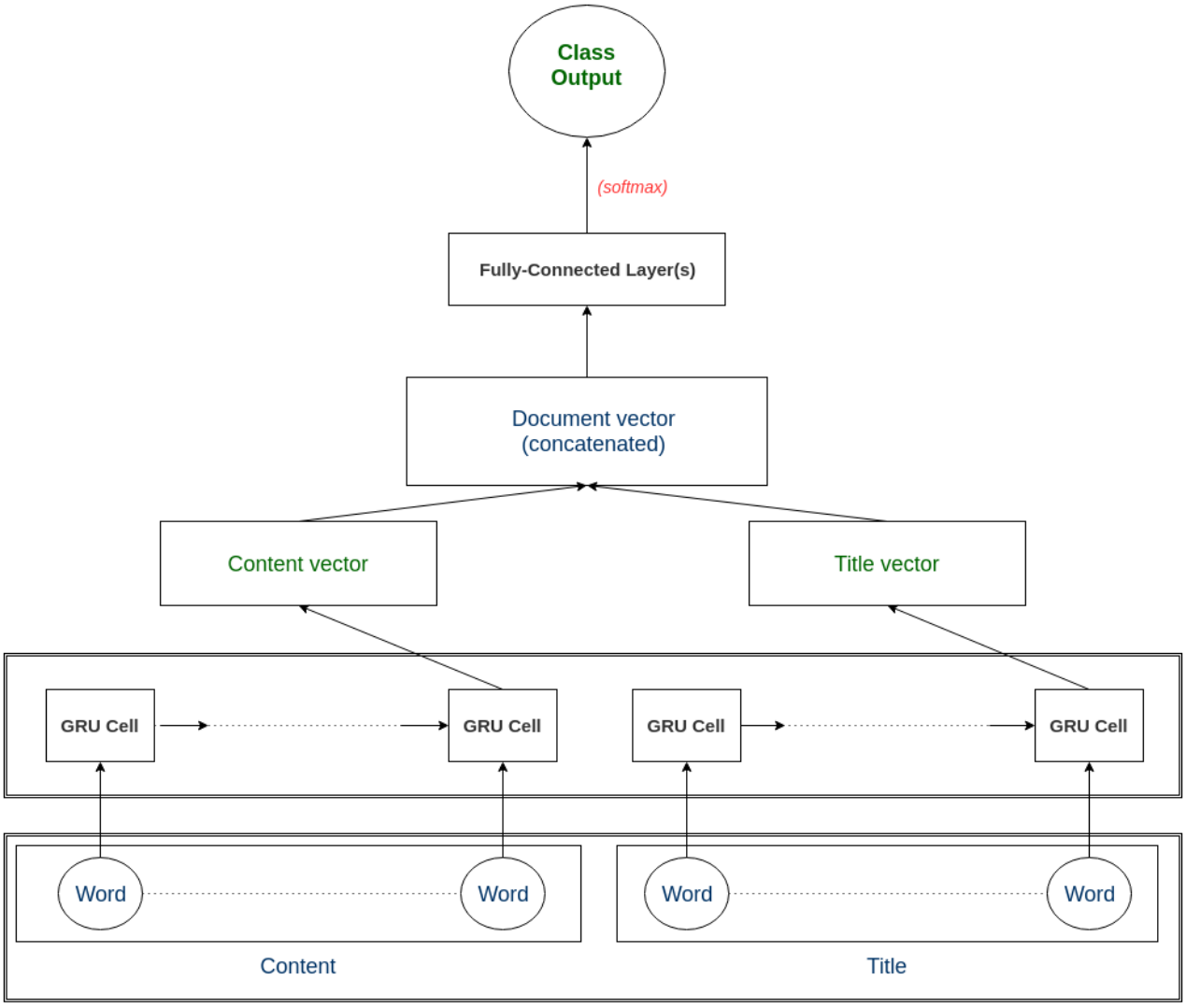
2. Mạng neuron - Mạng hồi quy

2.1. Mạng neuron (Artificial Neural Network – ANN)

2.2. Mạng hồi quy (Recurrent Neural Network – RNN)

3. Thiết kế, cài đặt

3.1. Mô hình mạng



3.2. Tập dữ liệu

3.3. Quá trình huấn luyện

3.4. Cài đặt chi tiết

4. Kết quả

5. Kết luận và hướng phát triển

Tài liệu tham khảo

1: Chase, Zach; Genain, Nicolas; Karniol-Tambour, Orren, Learning Multi-Label Topic Classification of News Articles,
2: Dennis Ramdass, Shreyes Seshasai, Document Classification for Newspaper Articles, 2009