# 3 mô hình GloVe

Thống kê sự xuất hiện của từ trong văn thể là nguồn thông tin chính có sẵn cho tất cả các phương pháp học không giám sát and mặc dù nhiều phương pháp tồn tại, vẫn còn nhiều câu hỏi làm cách nào ý nghĩa được tạo ra từ sự thống kê đó, và làm thể nào kết quả word vector có thể trình bày ý nghĩa đó. Trong phần này, chúng ta sẽ làm sáng tở câu hỏi. chúng ta sử dụng hiểu biết của chúng ta để xây dựng một mô hình mới cho trình bày từ gọi là GloVe, Global Vectors, bởi vì global corpus thống kê được thu lại trực tiếp bởi mô hình.

Đầu tiên chúng ta hình thành một vài ký hiệu. hay để ma trận word-word co-occurrence đếm được ký hiệu là X, XiJj trên bảng thể hiện số lần từ j xuát hiện canh từ i. hãy để Xi= … là số lần xuất hiện của ất kì từ nào trong văn thể của từ i. cuối cùng, P…. là tiể lệ từ j xuất hiện cạnh từ i

Bảng 1. Tỉ lệ cùng xuất hiện của từ ice và steam được chọn từ ngữ cảnh của 6 tỉ văn thể. Chỉ có tỉ lệ những từ không phân biệt như water và fashion bị hủy, vì vậy (số lượng lấn hơn 1 ) …

Chúng ta bắt đầu với ví dụ đơn giảng chỉ ra rằng khía cạnh chắc chắn của nghĩa có thể được trích xuất trực tiếp từ xác xuất sự xuất hiện cùng nhau. Xem xét 2 từ I và j chỉ ra ràng thể hiện một khía cạnh đặc biệt; cho tính cụ thể, giả sử chúng ta quan tâm đến đoạn văn về nhiệt động lục học, mà chúng ta có thể lấy I = ice và j = steam. mối quan hệ của những từ này có thể được diễn đặt bằng xác suất xuất hiện cùng nhau với các từ tham dò khác, k. đối với những từ k liên quan tới ice nhưng không liên quan tới steam, cho k = solid, chúng ta mong đợi ràng tỉ lệ Pik/Pjk sẽ lớn. tương tự cho từ k liên quan tới steam nhưng không liên quan tới ice, cho k=gas tỉ lệ sẽ nhỏ. Nêu k những từ như water hoặc fashion, nó có thể liên quan tới cả ice và steam, hoặc không, do đó tỉ lệ sẽ tiến gần về 1. Bảng 1 cho ta thấy xác xuất của những tỉ lệ trên 1 văn thể lớn và số lần xác nhận như mong đợi. so sánh tỉ lệ thô, tỉ lệ tốt hơn có thể phân biệt những từ liên quan (solid và gas) từ những từ không liên quan (water và fashion) và nó cũng tốt để phân biệt với những từ liên quan.

Lập luận ở trên gợi ý rằng điểm bắt đầu cho word vector là nên hoặc với tỉ lệ xác xuất cùng xuất hiện thay vì tỉ lệ của bản thân chúng. Ghi nhận ràng tỉ lệ Pik/Pjk dự trên ba từ I,j và k mô hình tổng quat nhất

…

Trong đó … là word vectors và … là ngữ cảnh word vector có vai trờ sẽ được bàn đến sau. Trong phương tirnh2 này, phía tay phải dùng để trích xuất từ văn thể và F có thể dự trên một vài thông số chưa xác định. Số lượng có thể của F là lớn, nhưng bằng việc thực hành mong muốn chúng ta có thể có một vài lựa chọn riêng biệt. đầu tiên, chúng ta muốn F để mã hóa thông tin trình bày tỉ lệ Pik/Pjk trong không gian word vector. Do đó không gian word vector kế thừa trực tiếp, cách tự nhiên nhất để làm việc này là với 2 vector khác biệt. Với mục đích này, chúng ta có thể giới hạn sự xem xét tới hàm F điều này dự trên sự khác biệt của 2 từ mục tiêu

…

Tiếp theo, chúng ta chú ý những tham số của F trong phương trình là những vector trong khi vế phải là vô hướng. trong khi F có thể được lý chức năng phức tạp bởi tham số mạng neural, làm như vậy sẽ làm rõ cấu trúc tuyến tính của chúng ta. Để tránh rắc rối này, chúng ta có thể bắt đầu trước điểm chấm của tham số.

…

Cái mà ngăn cản F từ việc trôn chiều của vector một cách không mong muốn. tiếp theo, chú ý rằng ma trận word-word co-occurrence, sự phân biệt giữa từ và ngữ cảnh là tùy tiện và điều đó là tự do để đổi vai trò của 2 từ. Làm như vậy nhất quán, chúng ta không chỉ đổi chỗ … mà còn …. Mô hình cuối cùng nên được cố điện theo cách gán nhãn này nhưng phương trình thì không. Tuy nhiên, tính đối xứng có thể được khôi phục qua 2 cách. Đầu tiên, chugn1 ta yêu cầu ràng F là một đồng số giữa … và …

…