Đại học Khoa học tự nhiên

Môn học: Khoa Học Dữ Liệu

Báo cáo đồ án cuối kỳ

Word2Vec

Giảng viên
Nguyễn Ngọc Đức

Sinh viên
20424008 - Dương Mạnh Cường

Ngày 9 tháng 2 năm 2022

Mục lục

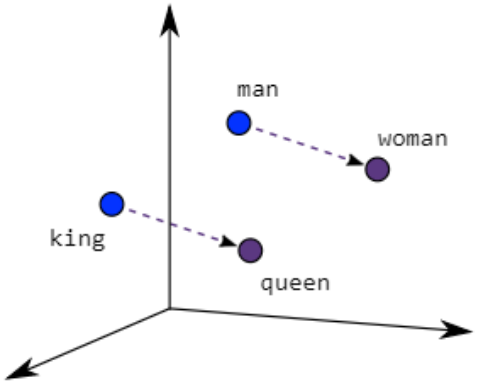
Neural network yêu cầu đầu vào ở dạng **numeric**. Cho nên, khi ta có dữ liệu dạng **text**, ta cần chuyển đổi chúng thành dữ liệu dạng numeric.

Có nhiều phương pháp khác nhau để chuyển đổi dữ liệu dạng text sang numeric mà phổ biến nhất là:

- Term frequency-inverse document frequency (TF-IDF).
- Bag of words (BOW).

Tuy nhiên, điểm yếu của hai phương pháp trên là chúng **không nắm bắt được ngữ nghĩa của từ**, nói cách khác là chúng không hiểu được ý nghĩa của từ.

Có nhiều cách khắc phục nhược điểm này, mà một trong những cách đó là sử dụng **Word2Vec** bằng cách đại diện cho từng từ bằng một **vector** trong không gian n chiều. Lúc này, các từ có nghĩa tương đồng nhau sẽ nằm gần nhau.



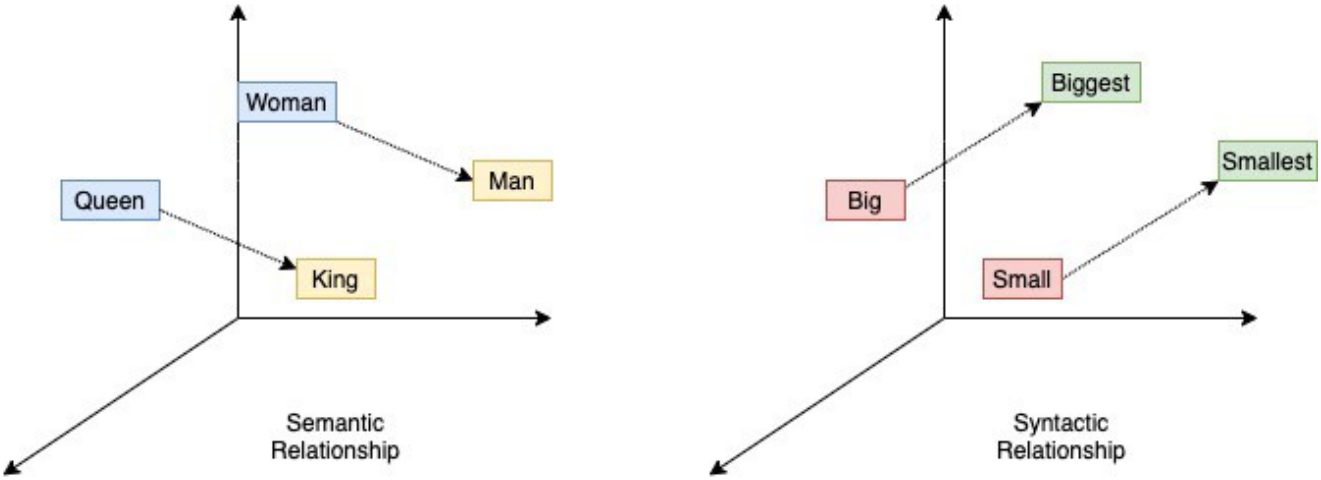
Hình 1: Các từ có ý nghĩa tương đồng nhau nằm gần nhau.

1 Word2Vec model

Word2Vec là một trong những phương pháp **word embedding** được sử dụng phổ biến.

Word embedding là cách ta biểu diễn các **word vector** trong không gian vector.

Các word vector được tạo ra bởi Word2Vec model có khả năng nắm bắt được các **semantic** *(ngữ nghĩa)* và **syntactic** *(ý nghĩa cú pháp)* của từ.



Hình 2: Ví dụ về **semantic** và **syntactic**.

Ví dụ có câu: *"Archie used to live in New York, he then moved to Santa Clara. He loves apples and strawberries."*.

Word2Vec model sẽ phát sinh các vector cho từng từ trong văn bản. Nếu chúng ta trực quan các vector này trong không gian vector tương ứng, chúng ta có thể thấy các từ tương tự nhau sẽ nằm gần nhau.

Hình 3: Các từ trong câu ví dụ được biểu diễn trong không gian vector.

Nhận xét

- Ở đây các cặp từ như *apples* - *strawberries*, *New York* - *Santa Clara* có ý nghĩa tương đồng nhau nên nằm gần nhau.