Ứng dụng mô hình chủ đề theo dõi các trang tin tức online

GVHD: Ninh Khánh Duy
Sinh viên : Phạm Thế Tâm

Mục lục

BÁO CÁO CÔNG VIỆC

- Lý thuyết các thuật toán

- Demo các chức năng tìm kiếm tài liệu theo chủ đề và tài liệu liên quan

- Deloy lên server

Kế hoạch :

- Xử lí realtime

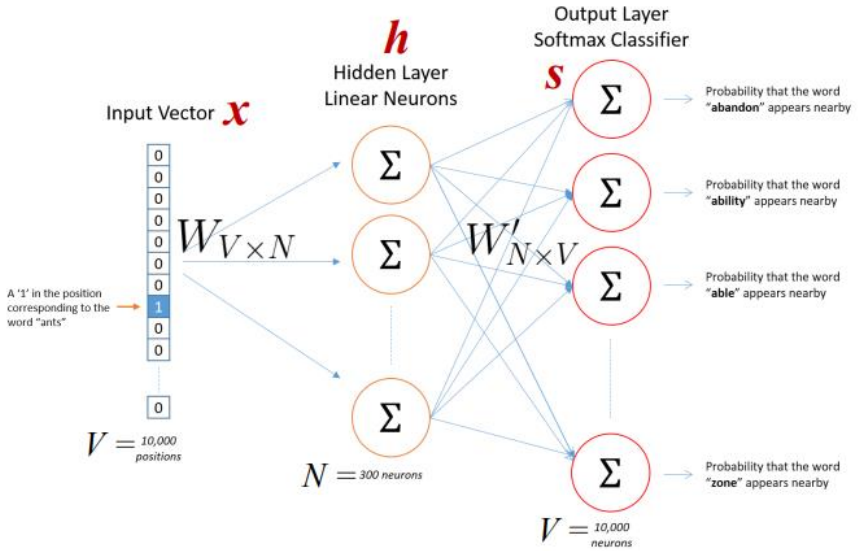
- Xử lí các vấn đề backend

word2vec (skip gram)

Đầu vào: vec-tơ one-hot của một từ

Đầu ra: vec-tơ ngữ cảnh của từ đó

Mục đích: mô hình hóa các từ trong không gian, các từ thường xuất hiện chung ngữ cảnh sẽ nằm gần nhau trong không gian.



doc2vec (DBOW)

Các tham số :

-window size: 15

- minimum count: 10

-vector-size: 300

-epochs: 20-400

UMAP

Các tham số :

UMAP

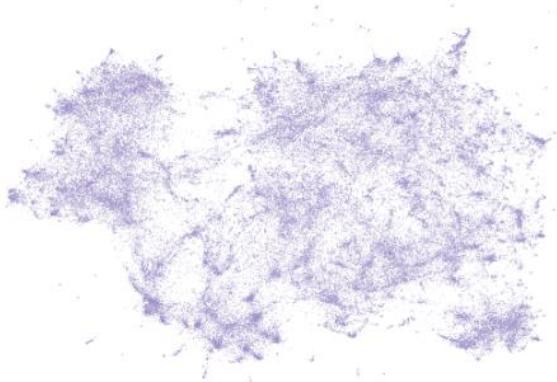
number of nearest neighbour: 7

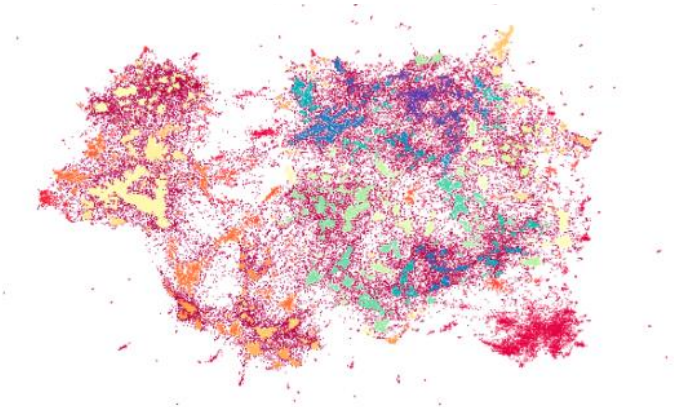
metric: cosine similarity ((tính sự tương đồng giữa các tài liệu)

HDBSCAN:

minimum cluster size : 5 (kích thước nhỏ nhất của một cụm)

metric: euclidean (tính khoảng cách giữa các điểm trong không gian để gom nhóm)





Thuật toán

Doc2vec

Tạo các vector nhúng tài liệu và vector từ trong cùng không gian

UMAP

Giảm chiều dữ liệu cho các vector tài liệu

HDBSCAN

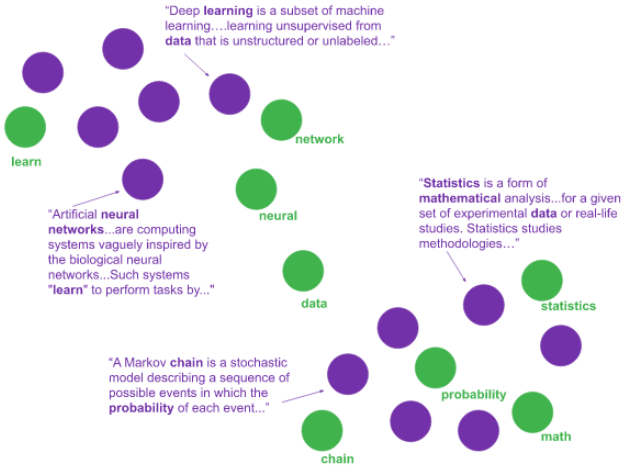
Tìm các cụm dày dặc tài liệu

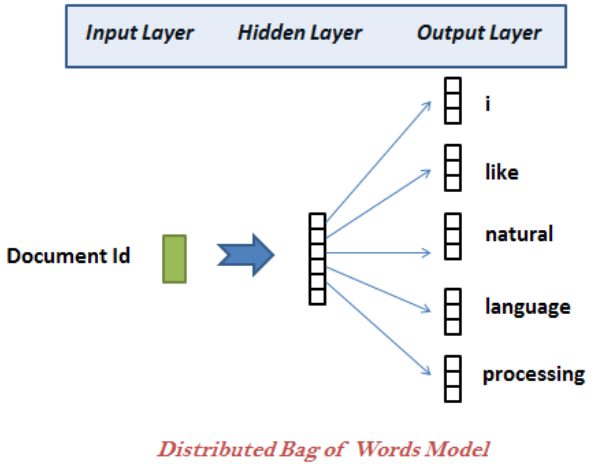
Tính trọng tâm của các cụm tài liệu -> vector chủ đề

TÌm các từ mô tả chủ đề

Thuật toán

1, Tạo các vector từ và tài liệu trong cùng không gian (*sử dụng Doc2vec*)

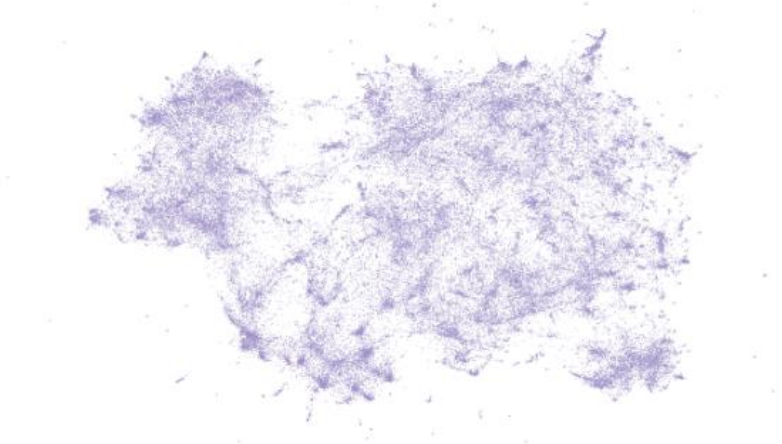




Distributed Bag of Words Model

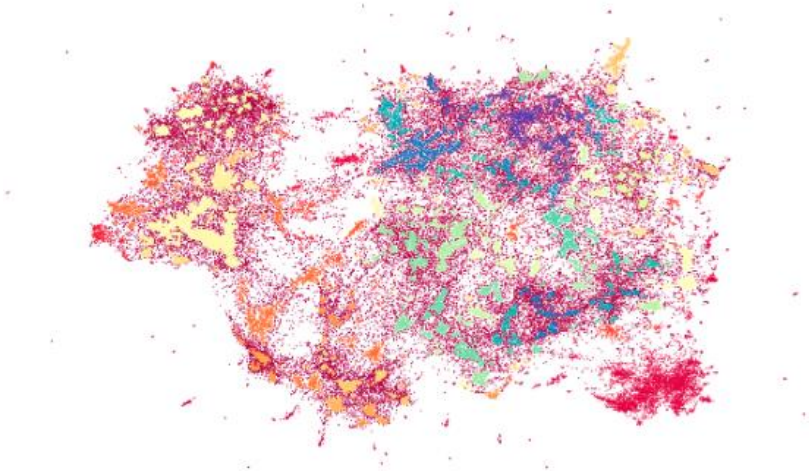
Thuật toán

2, Giảm chiều dữ liệu của các vector tài liệu (sử dụng Umap)



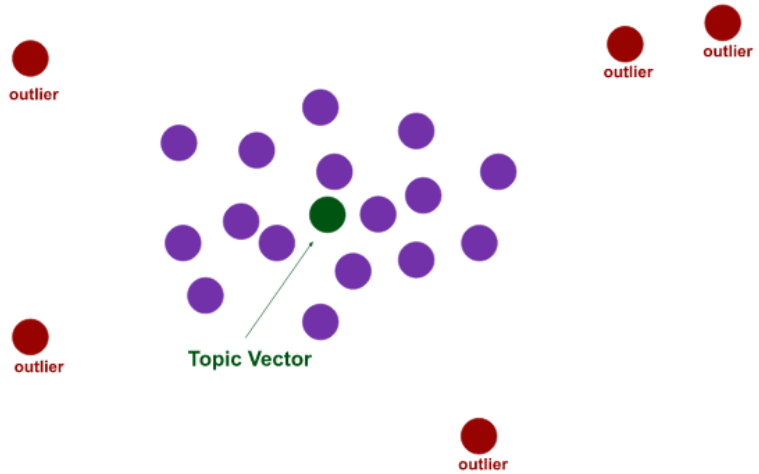
Thuật toán

3, Tìm những khu vực tập trung nhiều tài liệu (sử dụng HDBSCAN)



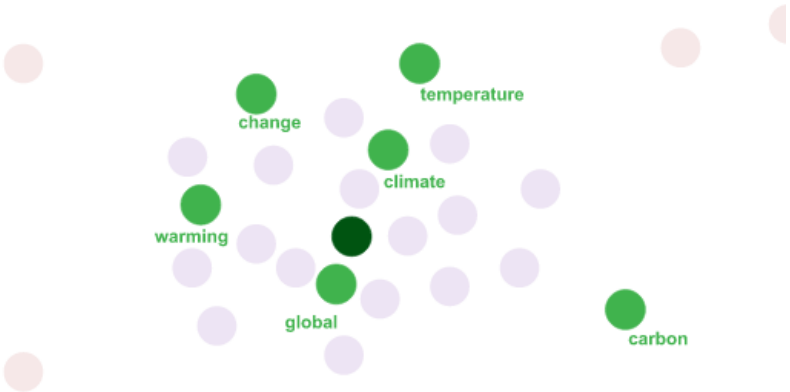
Thuật toán

4, Tính trọng tâm tại các cụm dày đặc của các vector tài liệu trong chiều không gian ban đầu



Thuật toán

5, Tìm các từ mô tả chủ đề



Một số chức năng

1, Xác định số lượng chủ đề

2, Tìm kiếm các tài liệu liên quan đến chủ đề

3, Tìm kiếm các tài liệu tương đương nhau

4, Tìm kiếm các tài liệu tương tự về mặt ngữ nghĩa theo từ khóa

DỰ ĐỊNH

REALTIME

DATABASE

FETCH API