Trường ĐH Khoa Học Tự Nhiên Tp. Hồ Chí Minh
TRUNG TÂM TIN HỌC

Đồ án tốt nghiệp Data Science

Topic: ***Đề xuất xe máy tương tự, phân khúc xe máy***

https://csc.edu.vn/data-science-machine-learning/Do-An-Tot-Nghiep-Data-Science---Machine-Learning_229

2025

Nội dung

1. **Giới thiệu project**

2. **Triển khai project theo Data Science Process**

Giới thiệu project

Recommender/recommendation system

- Là một subclass của information filtering system tìm cách dự đoán "xếp hạng" hoặc "ưu tiên" mà người dùng sẽ dành cho một mục. Chúng chủ yếu được sử dụng trong các ứng dụng thương mại.

https://en.wikipedia.org/wiki/Recommender_system

Giới thiệu project

Ứng dụng của hệ thống gợi ý - Recommender System

Dịch vụ phát nhạc

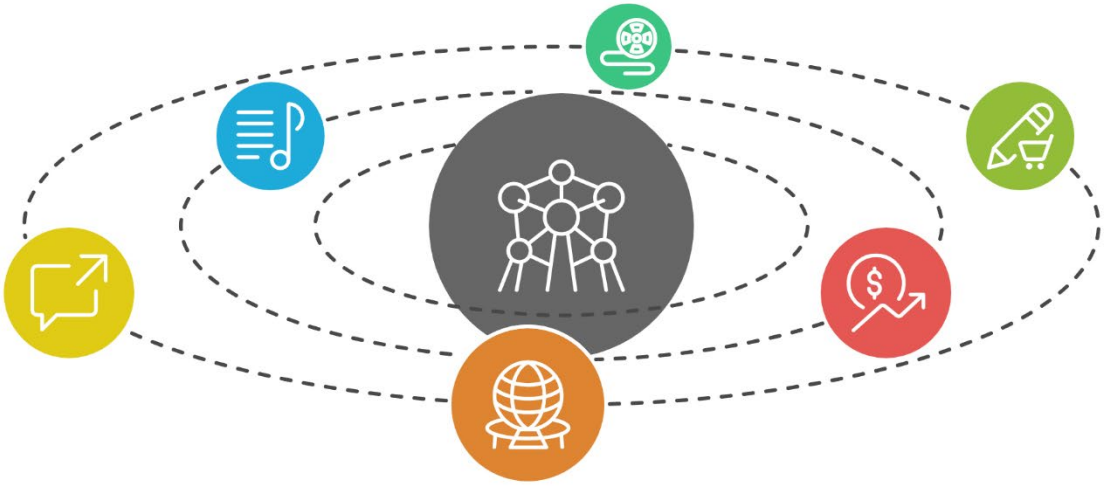
Đề xuất danh sách phát cho các dịch vụ như Spotify

Dịch vụ phát video

Gợi ý nội dung cho nền tảng như Netflix, Youtube

Thương mại điện tử

Đề xuất sản phẩm cho các trang web như Amazon, Shopee, Tiki, Lazada



Mạng xã hội

Đề xuất nội dung cho Facebook và Twitter

Dịch vụ & Du lịch

Gợi ý nhà hàng, khách sạn, nơi lưu trú như Agoda, Mytour

Tài chính

Gợi ý cho dịch vụ tài chính

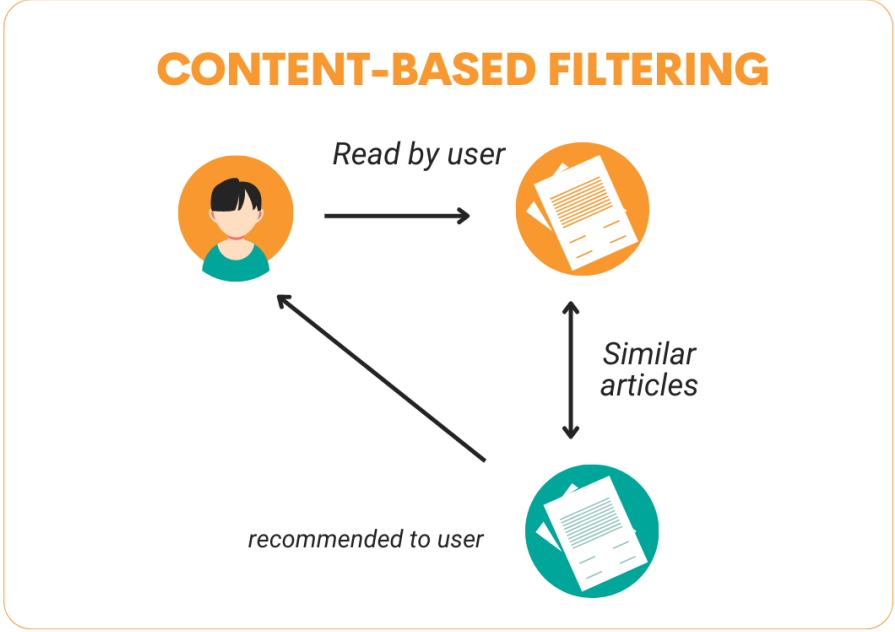
Giới thiệu project

- Recommender system là các thuật toán nhằm đề xuất các item có liên quan cho người dùng (Item có thể là phim để xem, văn bản để đọc, sản phẩm cần mua hoặc bất kỳ thứ gì khác tùy thuộc vào ngành dịch vụ).
- Recommender system thực sự quan trọng trong một số lĩnh vực vì chúng có thể tạo ra một khoản thu nhập khổng lồ hoặc cũng là một cách để nổi bật đáng kể so với các đối thủ cạnh tranh.

Giới thiệu project

❑ Content-based Recommendation

- Là một trong những ứng dụng thuộc nhóm Recommender system dùng để đề xuất cho người dùng dựa trên nội dung mà họ đang quan tâm.



Giới thiệu project

❑ Market Segmentation

- Là việc phân chia một thị trường rộng lớn thành các nhóm nhỏ hơn dựa trên đặc điểm hoặc hành vi tương tự.
- Việc này giúp doanh nghiệp có thể xây dựng **chiến lược tiếp thị và sản phẩm tập trung, hiệu quả hơn** cho từng nhóm cụ thể.

Giới thiệu project

- Trong ngữ cảnh thị trường xe máy cũ, chúng ta sẽ áp dụng các kỹ thuật **phân cụm** để **phân loại các mẫu xe** dựa trên các đặc điểm như **giá cả, thương hiệu, độ phổ biến, năm sản xuất, và tình trạng sử dụng...**, nhằm mục đích tìm ra các phân khúc riêng biệt để định giá và tiếp cận phù hợp.

Giới thiệu project

Business Objective/Problem

- **Chợ Tốt** là thị trường mua bán trực tuyến hàng đầu tại Việt Nam cung cấp đa dạng các hạng mục như mua bán nhà cửa, ô tô, xe máy, dịch vụ gia đình.
- Trong project này tập trung vào đối tượng là xe máy cũ.

Giới thiệu project

- Giả sử Chợ Tốt chưa triển khai các chức năng như ***gợi ý xe máy tương tự*** và ***phân khúc cho xe máy*** và bạn được yêu cầu triển khai, bạn sẽ làm gì?

Giới thiệu project

- **Các kiến thức/ kỹ năng cần để giải quyết vấn đề này:**
  - Hiểu vấn đề
  - Import các thư viện cần thiết và hiểu cách sử dụng
  - Đọc dữ liệu (dữ liệu project này được cung cấp)
  - Thực hiện EDA cơ bản (sử dụng *Pandas Profiling Report, dataprep...)*
  - Tiền xử lý dữ liệu: làm sạch, tạo tính năng mới, lựa chọn tính năng cần thiết…

Giới thiệu project

- Trực quan hóa dữ liệu
- Phân tích dữ liệu
- Lựa chọn thuật toán phù hợp cho bài toán recommendation system, và bài toán clutering
- Xây dựng model
- Đánh giá model
- Báo cáo kết quả

Nội dung

1. Giới thiệu project

2. **Triển khai project theo Data Science Process**

Triển khai project theo Data Science Process

- Thư viện sử dụng
  - numpy, pandas, matplotlib, seaborn, wordcloud
  - pandas_profiling / dataprep
  - Gensim, sklearn.metrics.pairwise cosine_similarity: Content-Based Filtering
  - Sklearn; các thuật toán phân cụm như Kmeans, GMM, AgglomerativeClustering
  - PySpark: các thuật toán phân cụm như Kmeans, GMM, Bisecting K-Means
  - Dùng PCA hoặc t-SNE để giảm chiều khi trực quan kết quả phân cụm

Triển khai project theo Data Science Process

❑ Triển khai dự án

- Bước 1: Business Understanding
  - Dựa vào yêu cầu nói trên → xác định vấn đề:

    → Xây dựng hệ thống gợi ý xe tương tự cho người mua:
    - Content-based filtering

    →Phân nhóm thị trường xe máy TP.HCM theo hành vi và đặc trưng kỹ thuật.
    - Clustering

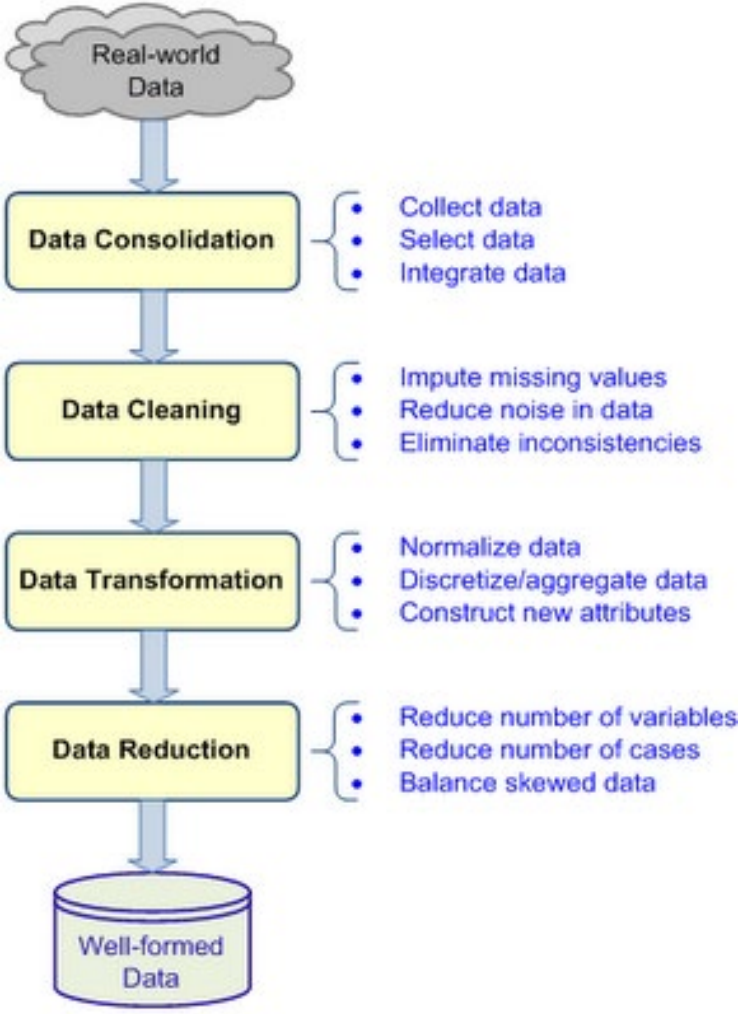
Triển khai project theo Data Science Process

- Bước 2: Data Understanding/ Acquire
  - Từ mục tiêu/ vấn đề đã xác định: xem xét các dữ liệu cần thiết:
    - Dữ liệu được cung cấp sẵn gồm có tập tin: data_motorbikes.xlsx với hơn 7000 mẫu
    - Thông tin chi tiết: tập tin **Mô tả bộ dữ liệu Chợ Tốt.pdf**

CHÚ Ý: TẤT CẢ DỮ LIỆU ĐƯỢC CUNG CẤP CHỈ DÀNH CHO VIỆC HỌC TẬP, KHÔNG CHIA SẺ, SỬ DỤNG VỚI MỤC ĐÍCH KHÁC. Vì việc chia sẻ, sử dụng cho các mục đích khác đều có thể vi phạm Bản quyền và quyền sở hữu trí tuệ, Điều khoản sử dụng của trang web, Bảo vệ dữ liệu cá nhân... → vấn đề pháp lý. Các bạn nhớ lưu ý!

Triển khai project theo Data Science Process

- Bước 3: Data preparation/ Prepare



Triển khai project theo Data Science Process

- Bước 4&5: Modeling & Evaluation/ Analyze & Report
  - Thực hiện việc đánh giá các model ở từng bài toán → báo cáo kết quả thu được → đưa ra các lựa chọn phù hợp.

Triển khai project theo Data Science Process

- Với bài toán 1:
  - Xây dựng model Content-based filtering
    - cosine_similarity
    - Gensim
  - Thực hiện/ đánh giá kết quả
    - Kết luận

Triển khai project theo Data Science Process

❑ Giới thiệu Gensim - *"Generate Similar"*

- Là một thư viện Python chuyên xác định sự tương tự về ngữ nghĩa giữa hai tài liệu thông qua mô hình không gian vector và bộ công cụ mô hình hóa chủ đề.
- Có thể xử lý kho dữ liệu văn bản lớn với sự trợ giúp của việc truyền dữ liệu hiệu quả và các thuật toán tăng cường
- Tốc độ xử lý và tối ưu hóa việc sử dụng bộ nhớ tốt
- Tuy nhiên, Gensim có ít tùy chọn tùy biến cho các function
- Tham khảo:
  https://www.tutorialspoint.com/gensim/index.htm
  https://www.machinelearningplus.com/nlp/gensim-tutorial/

demo

Triển khai project theo Data Science Process

❑ Giới thiệu cosine_similarity

- Ý tưởng chính của phương pháp này là đưa ra gợi ý dựa vào sự tương đồng với nhau giữa các sản phẩm.

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

- Tham khảo:
  - https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html
  - https://en.wikipedia.org/wiki/Cosine_similarity

demo

Triển khai project theo Data Science Process

- Với bài toán 2:
  - Xây dựng model Clustering
    - Dùng Machine Learning truyền thống: Kmeans, GMM, AgglomerativeClustering
    - Dùng PySpark: Kmeans, GMM, Bisecting K-Means
  - Thực hiện/ đánh giá kết quả
    - Kết luận
  - Chú ý: Cần phải thực hiện trên cả 2 môi trường ML truyền thống và PySpark, mỗi môi trường ít nhất 2 thuật toán.

Triển khai project theo Data Science Process

- Bước 6: Deployment & Feedback/ Act
  - Triển khai lên website và theo dõi kết quả. (Sẽ thực hiện ở tuần sau)

Triển khai project theo Data Science Process

Các công việc cần thực hiện:

- Hãy triển khai project trên với các bước theo Data Science Process
- Áp dụng **cosine_similarity** và **genism** (content-based filtering)
- Áp dụng các thuật toán phân cụm trong cả sklearn (+ các gói cung cấp khác) và PySpark
- Với mỗi bài toán cần có report và đưa ra kết luận.

Triển khai project theo Data Science Process

Gợi ý

- Thực hiện việc tìm hiểu các thuộc tính trong dữ liệu, các tiền xử lý, khám phá dữ liệu, kèm theo các trực quan cần thiết
- Xây dựng và đánh giá các model của cả hai bài toán.
- Ngoài các thuật toán được gợi ý và đã thực hiện, có thuật toán nào khác cho kết quả tốt hơn không? Thực hiện với thuật toán đó *(điểm cộng)*
- Tổng hợp các kết quả