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

Đồ án tốt nghiệp Data Science

Topic: ***Content-Based Company Similarity Recommendation and 'Recommend or Not' Classification for Candidates***

Phòng LT & Mạng

https://csc.edu.vn/data-science-machine-learning/do-an-tot-nghiep-data-science_310

2025

Nội dung

1. **Giới thiệu project**

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

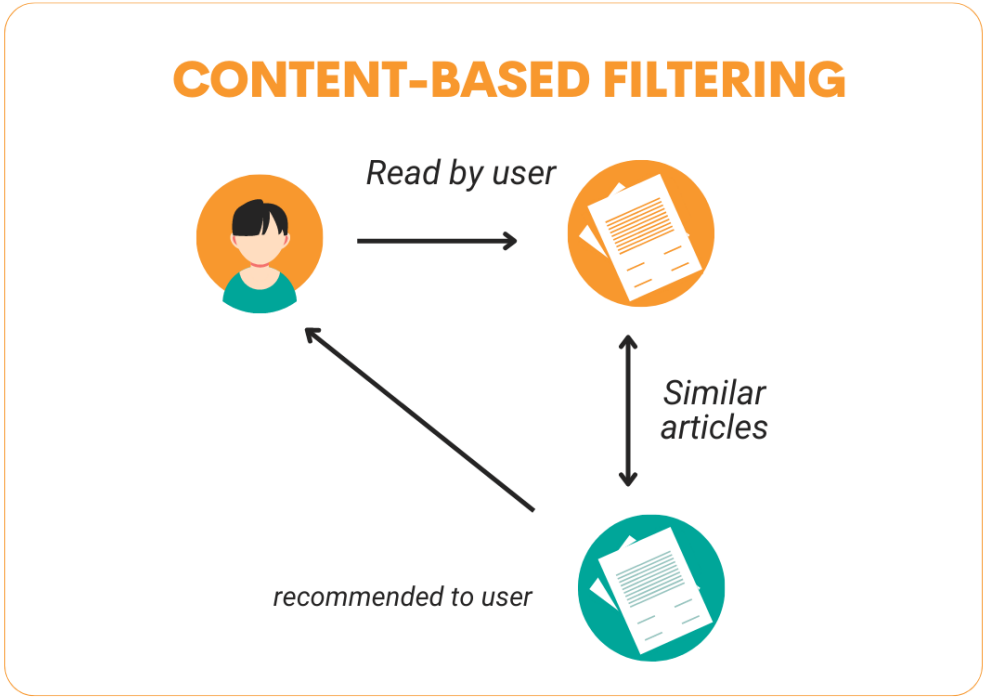
Giới thiệu project

Content-Based Filtering (Lọc dựa trên nội dung)

- Là một trong những phương pháp nền tảng và phổ biến nhất trong lĩnh vực hệ thống đề xuất (Recommendation Systems).
- Ý tưởng cốt lõi của phương pháp này là đề xuất các mục (items) cho người dùng dựa trên sự tương đồng giữa các đặc điểm của mục đó và sở thích đã được biết của người dùng.

Giới thiệu project

- Content-Based Filtering được ứng dụng rộng rãi trong nhiều lĩnh vực như đề xuất tin tức, bài viết, sản phẩm trên các trang thương mại điện tử, phim ảnh, âm nhạc, và gần đây là đề xuất công việc hoặc công ty dựa trên mô tả.



Giới thiệu project

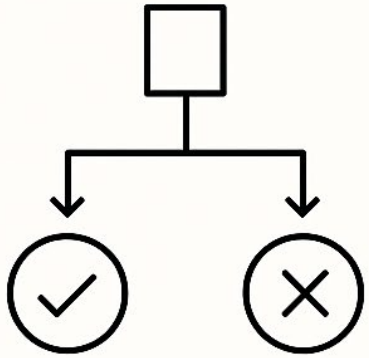
Classification

- Là một bài toán trong Machine Learning mà mục tiêu là **dự đoán nhãn (label) phân loại** cho một đầu vào. Hệ thống sẽ học từ dữ liệu có gán nhãn (supervised learning) để xây dựng mô hình có khả năng **phân loại đối tượng vào các nhóm cụ thể** (ví dụ: “spam” hoặc “not spam”, “recommend” hoặc “not recommend”).

Giới thiệu project

CLASSIFICATION IN MACHINE LEARNING

Classification is a machine learning task of predicting the label of an input.



POPULAR ALGORITHMS:

- Logistic Regression
- Decision Trees
- Random Forest
- Support Vector Machine
- Neural Networks

Applications: email filtering, disease diagnosis, sentiment analysis, customer satisfaction prediction

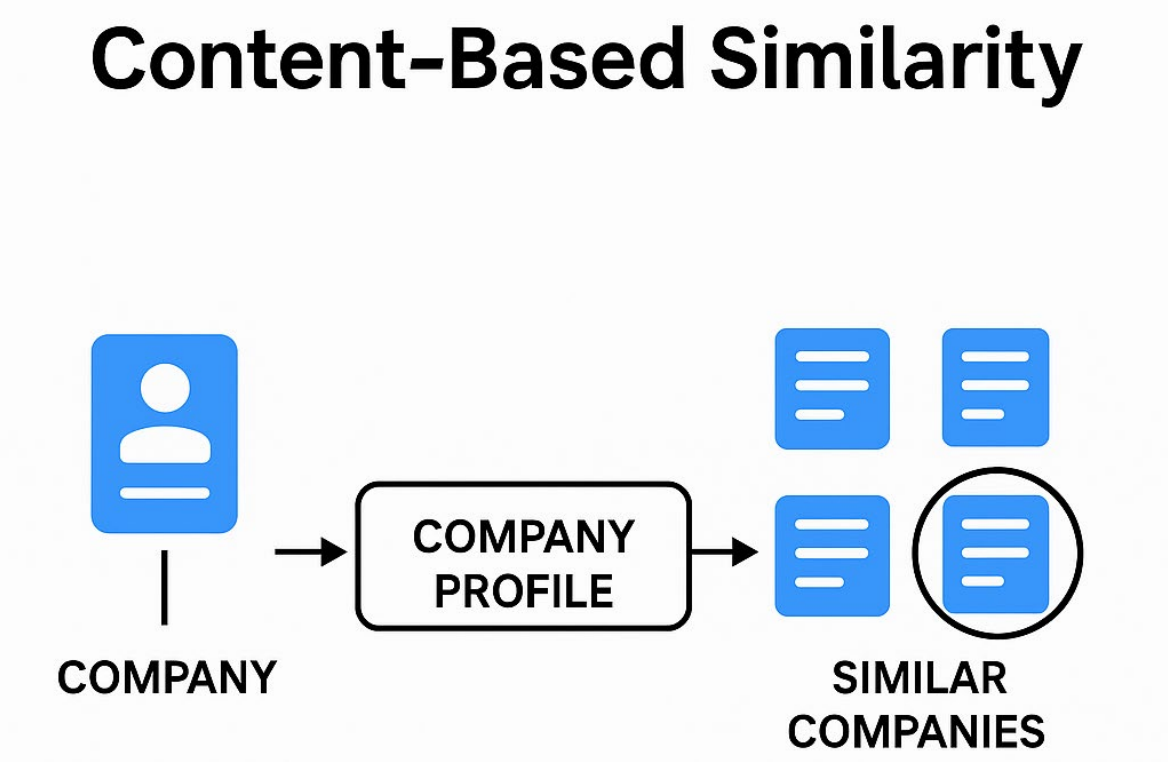
Giới thiệu project

- **Business Objective/Problem**
  - https://itviec.com/

Xem thông tin giới thiệu tại: Mô tả bộ dữ liệu ITViec.pdf

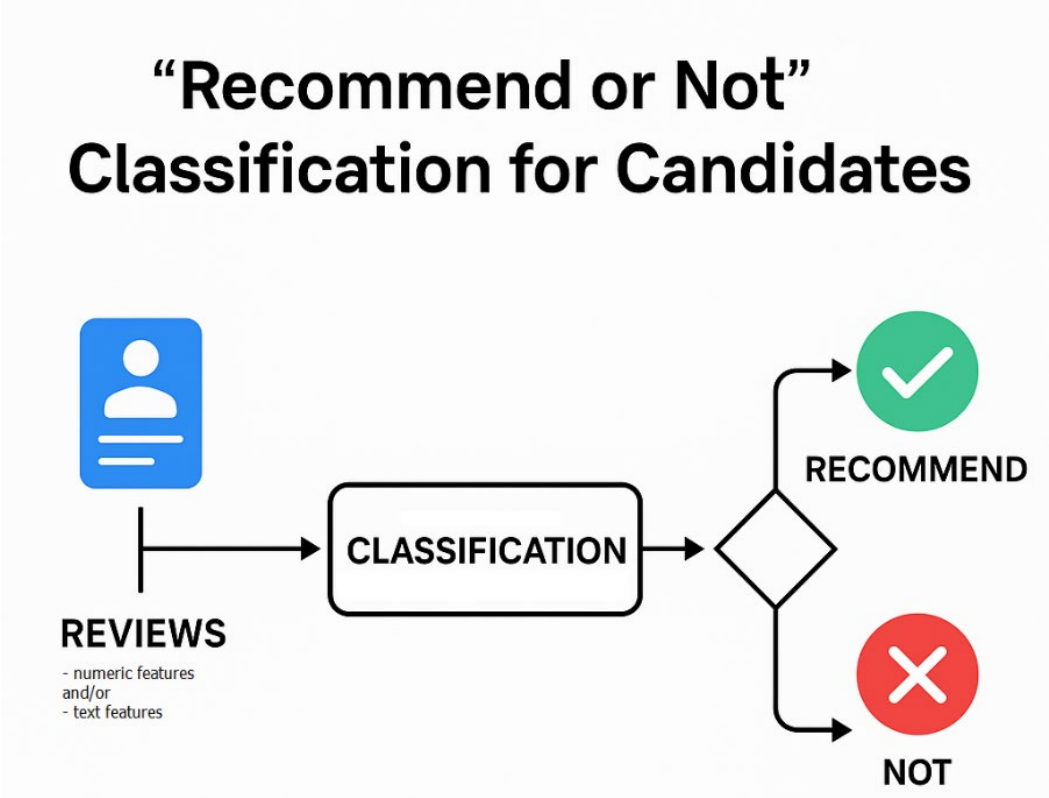
Giới thiệu project

- Yêu cầu 1: Dựa trên những thông tin từ các công ty đăng trên ITViec để gợi ý các công ty tương tự dựa trên nội dung mô tả.



Giới thiệu project

- Yêu cầu 2: Dựa trên những thông tin từ review của ứng viên/ nhân viên đăng trên ITViec để dự đoán khả năng "Recommend" công ty.



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
  - 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
- Lựa chọn thuật toán phù hợp cho bài toán
- 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
  - underthesea
  - wordcloud
  - scikit-learn (sklearn)
  - Gensim
  - Pyspark
  - …

Triển khai project theo Data Science Process

❑ Triển khai dự án

- Bước 1: Business Understanding

  - Dựa vào yêu cầu 1 và yêu cầu 2 → xác định vấn đề:
    - Mục tiêu/ Vấn đề 1: Xây dựng mô hình dự đoán giúp đề xuất các công ty tương tự với một công ty → Giúp một công ty có thể tìm thấy các đối thủ cạnh tranh, đối tác tiềm năng, hoặc các công ty có cấu trúc tương đồng để học hỏi.
    - Mục tiêu/ Vấn đề 2: Xây dựng mô hình phân loại "Recommend or Not" → Giúp ứng viên tìm thấy những cơ hội việc làm phù hợp nhất.

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 trong các tập tin:
      - Overview_Companies.xls
      - Overview_Reviews.xls
      - Reviews.xls
    - Kèm theo đó là file mô tả: Mô tả bộ dữ liệu ITViec.pdf
    - Chú ý: Các nhóm chọn lựa các thông tin cần thiết để đưa vào giải quyết từng vấn đề.

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!

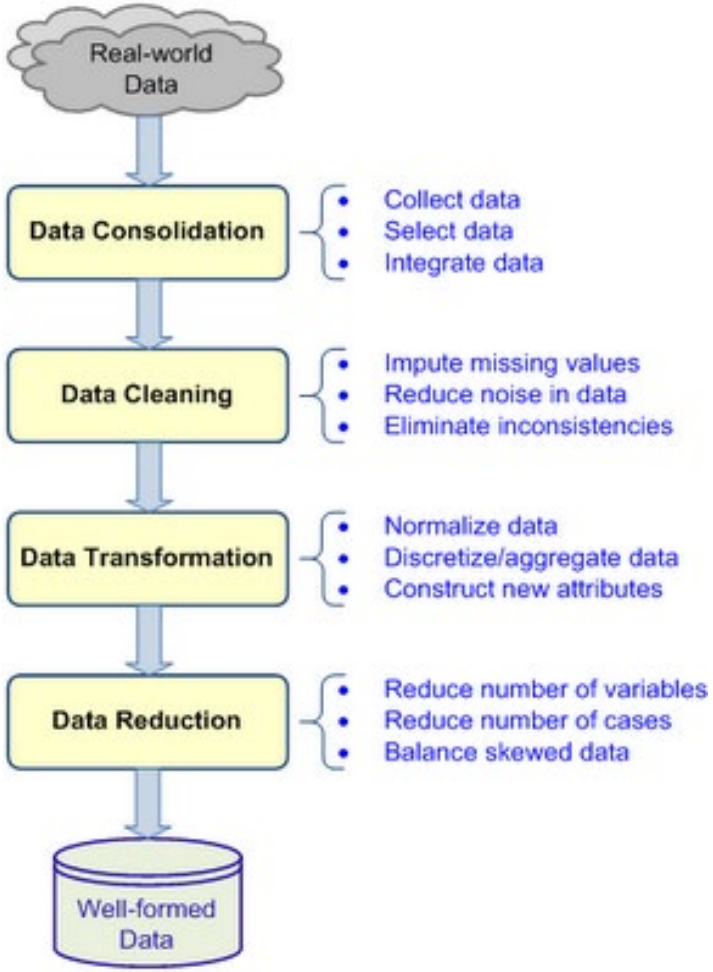
Triển khai project theo Data Science Process

→ Có thể tập trung giải quyết bài toán

- Content-Based Filtering với Gensim, sklearn.metrics.pairwise cosine_similarity…
- Classification với các thuật toán thuộc nhóm như: các thuật toán như Logistic Regression, Tree Models, SVM...

Triển khai project theo Data Science Process

- Bước 3: Data preparation/ Prepare
  - Thực hiện các công việc:
    - Tiền xử lý dữ liệu text
    - Tiền xử lý dữ liệu numeric



Tham khảo:

Triển khai project theo Data Science Process

- Bước 4&5: Modeling & Evaluation/ Analyze & Report
  - 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_NLP_ContentBased.ipynb

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_NLP_ContentBased.ipynb

Triển khai project theo Data Science Process

- Bước 4&5: Modeling & Evaluation/ Analyze & Report
  - Xây dựng các Classification model
    - Machine Learning with Python: KNN, Logistic Regression, Tree Algorithms, SVM
    - PySpark: Logistic Regression, Tree Algorithms
  - Thực hiện/ đánh giá kết quả các Classification model
    - R-squared, ccc, precision, recall, f1, confusion matrix, ROC curve…
    - Kết luận

Triển khai project theo Data Science Process

- Bước 6: Deployment & Feedback/ Act
  - Triển khai Recommender System lên website và theo dõi kết quả.

Triển khai project theo Data Science Process

Các công việc cần thực hiện…

Triển khai project theo Data Science Process

Với project phía trên

- Công việc 1: Thực hiện việc tiền xử lý dữ liệu text/ numeric để làm đầu vào cho việc xây dựng các model khác nhau.

Triển khai project theo Data Science Process

- Công việc 2: Với yêu cầu 1:
  - Áp dụng **cosine_similarity** và **genism** (content-based filtering)
  - So sánh các kết quả
  - Có thể đề xuất thêm các thuật toán mới (+0.25đ)

Triển khai project theo Data Science Process

- Công việc 3: Với yêu cầu 2, đề xuất các thuật toán có thể, sau đó chọn thuật toán phù hợp để giải quyết các vấn đề, report kết quả.
  - Lựa chọn thuật toán phù hợp ở cả sklearn và PySpark.ml (3 thuật toán mỗi loại)
  - Nếu dữ liệu mất cân bằng gây ảnh hưởng đến kết quả thì xem xét thêm việc xử lý mất cân bằng
  - So sánh các kết quả
  - Có thể đề xuất thêm các thuật toán mới (+0.25đ)

Triển khai project theo Data Science Process

Chú ý

- Trong slide báo cáo, cần có bảng phân công công việc (ở cuối)
- Mỗi nhóm trình bày trong 10 phút.