



# Đồ án tốt nghiệp Data Science

## Topic 3: *Recommender System*

Phòng LT & Mạng

[https://csc.edu.vn/data-science-machine-learning/Data-Science-Capstone-Project-Hinh-thuc-2\\_225](https://csc.edu.vn/data-science-machine-learning/Data-Science-Capstone-Project-Hinh-thuc-2_225)

2021



# 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](https://en.wikipedia.org/wiki/Recommender_system)

# Giới thiệu project

- Recommender systems được sử dụng trong nhiều lĩnh vực: tạo danh sách phát nhạc/video cho các dịch vụ như Netflix, YouTube & Spotify, đề xuất sản phẩm cho các dịch vụ như Amazon, đề xuất nội dung cho các nền tảng truyền thông xã hội (social media platform) Facebook & Twitter. Những system có thể hoạt động bằng cách sử dụng một single input (như music), hay multiple input trong và trên các nền tảng như news, books,... và truy vấn tìm kiếm (search query).

# Giới thiệu project

---

- Ngoài ra, còn có các recommender system phổ biến cho các chủ đề cụ thể như nhà hàng, hẹn hò trực tuyến...
- Recommender system cũng được phát triển để khám phá các bài báo, tác giả nổi tiếng, các nhóm cộng tác và các dịch vụ tài chính.

# Giới thiệu project

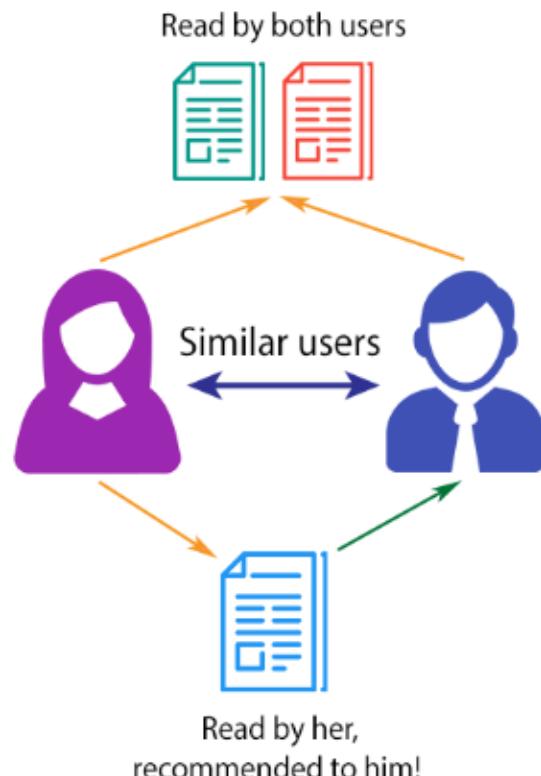
## □ Một cách tổng quát

- 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.

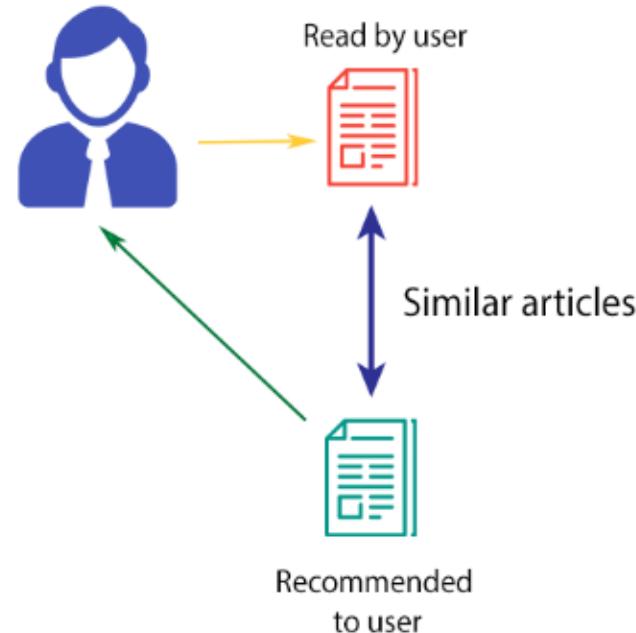
# Giới thiệu project

□ Có hai recommender system phổ biến nhất là Collaborative Filtering (CF) và Content-Based

COLLABORATIVE FILTERING



CONTENT-BASED FILTERING



# Giới thiệu project

## ❑ Business Objective/Problem

- Tiki là một hệ sinh thái thương mại “all in one”, trong đó có **tiki.vn**, là một website thương mại điện tử đứng top 2 của Việt Nam, top 6 khu vực Đông Nam Á.
- Trên trang này đã triển khai nhiều tiện ích hỗ trợ nâng cao trải nghiệm người dùng và họ muốn xây dựng nhiều tiện ích hơn nữa.
- Giả sử công ty này chưa triển khai Recommender System và bạn được yêu cầu triển khai hệ thống này, 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*)
  - 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 cho bài toán recommendation system
- 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
- Gensim, sklearn.metrics.pairwise cosine\_similarity  
....: Content-Based Filtering
- pyspark.ml.recommendation.ALS : Collaborative Filtering (CF)

## □ Triển khai dự án

### ● Bước 1: Business Understanding

- Dựa vào yêu cầu nói trên => xác định vấn đề:
  - Chưa có hệ thống Recommendation System
  - => Mục tiêu/ vấn đề: Xây dựng Recommendation System cho một hoặc một số nhóm hàng hóa trên tiki.vn giúp đề xuất và gợi ý cho người dùng/ khách hàng. => Xây dựng các mô hình đề xuất:
    - Content-based filtering
    - Collaborative filtering

# 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 các tập tin: ProductRaw.csv, ReviewRaw.csv chứa thông tin sản phẩm, review và rating cho các sản phẩm thuộc các nhóm hàng hóa như Mobile\_Tablet, TV\_Audio, Laptop, Camera, Accessory.

# Triển khai project theo Data Science Process

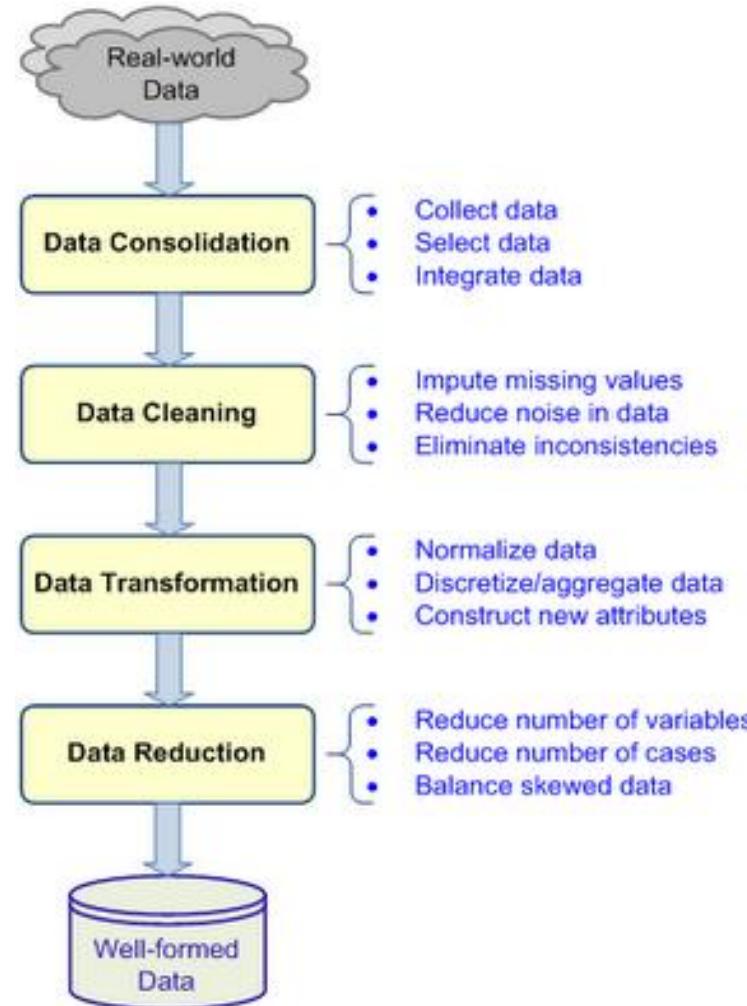
Product	
Item_id	Mã sản phẩm
Name	Tên sản phẩm
Description	Mô tả tính năng sản phẩm
Rating	Điểm đánh giá chung sản phẩm
Price	Giá bán sản phẩm
List_Price	Giá sản phẩm (chưa giảm giá hoặc khuyến mãi)
Brand	Thương hiệu
Group	Nhóm phân loại sản phẩm

Review	
Customer_id	Mã khách hàng
Product_id	Mã sản phẩm
Name	Tên tài khoản
Full_name	Tên khách hàng
Created_time	Thời điểm phản hồi được tạo
Rating	Điểm đánh giá sản phẩm của khách hàng
Title	Tiêu đề đánh giá
Content	Nội dung đánh giá

# 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

=> Tập trung giải quyết hai bài toán

- Bài toán 1: Đề xuất người dùng với Content-based filtering
- Bài toán 2: Đề xuất người dùng với Collaborative filtering

# 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

## □ 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

## □ 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.

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

- Tham khảo:

- [https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine\\_similarity.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html)
- [https://en.wikipedia.org/wiki/Cosine\\_similarity](https://en.wikipedia.org/wiki/Cosine_similarity) demo

# Triển khai project theo Data Science Process



## ■ Với bài toán 2:

- Xây dựng Collaborative Filtering  
(pyspark.ml.recommendation.ALS)
- ...
- Thực hiện/ đánh giá kết quả
  - RMSE
  - Kết luận

# Triển khai project theo Data Science Process

---



## ● Bước 6: Deployment & Feedback/ Act

- Triển khai Recommender System lên website thương mại điện tử và theo dõi kết quả.

# 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 gensim (content-based filtering)
- Áp dụng pyspark.ml.recommendation.ALS (Collaborative filtering)
- Đánh giá và report các kết quả



## ❑ Gợi ý

- Thực hiện việc tìm hiểu các thuộc tính trong dữ liệu, các tiền xử lý, khám phá dữ liệu cần thiết
  - Dựa trên ProductRaw.csv => xử lý => Product.csv, ReviewRaw.csv => xử lý => Review.csv (ví dụ: kiểm tra và xử lý dữ liệu trùng, dữ liệu null, loại cột không cần thiết...)
  - EDA: Từ Product.csv và Review.csv vừa tạo ở trên: thực hiện các công việc liên quan đến khám phá dữ liệu (ví dụ: thống kê cơ bản, pandas profiling; xem xét về giá, thương hiệu, rating, các sản phẩm được đánh giá nhiều nhất, khách hàng đánh giá nhiều nhất...)

# Triển khai project theo Data Science Process



- Dựa trên các thư viện được gợi ý thực hiện recommendation system.
- Ngoài thư viện được gợi ý và đã thực hiện ở trên, có thuật toán nào khác cho kết quả tốt hơn không? Thực hiện với thuật toán đó (*điểm cộng*)
- Tổng hợp các kết quả

