1. **Giới thiệu**
   1. **Nền tảng**
   2. **Hướng nghiên cứu**
2. **Thu thập và xử lý dữ liệu**
   1. **Nguồn dữ liệu**

Toàn bộ dữ liệu bao gồm thông tin về game kèm theo một trăm bình luận mỗi game được lấy từ Google Play trong phân mục “Trò Chơi” (https://play.google.com/store/apps/category/GAME)

Thư viện đã dùng:

* scrapy
* google-play-scraper

Dữ liệu đã thu thập được:

* 3509 trò chơi với 17 cột
  + Id
  + Title (tên game)
  + url (đường link của game)
  + category (thể loại)
  + avg\_rating (số điểm đánh giá trung bình)
  + rating\_count (số lượt đánh giá)
  + price (giá)
  + developer\_name (tên nhà phát triển)
  + developer\_url (trang Google Play của nhà phát triển)
  + description (mô tả về game)
  + summary (mô tả vắng tắt)
  + min\_installs (số lượt tải tối thiểu)
  + editors\_choice (game được biên tập viên lựa chọn)
  + size (kích cỡ)
  + android\_version (yêu cầu về phiên bản android tối thiểu)
  + content\_rating (phân loại nội dung theo độ tuổi)
  + ad\_supported (game có chứa quảng cáo)
  + released (ngày phát hành)
* Gần 250,000 dòng đánh giá game với 9 cột
  + id
  + username (tên người dùng)
  + score (số điểm đánh giá)
  + content (nội dung đánh giá)
  + like\_count (số lượt thích của đánh giá)
  + date (ngày đánh giá)
  + game\_title (tên game)
  + game\_id (id game)
  + game\_category (thể loại của game)
  1. **Làm sạch**
     1. **Games**
* Chuyển đổi cột “price” từ VNĐ sang USD.
* Tạo cột “is\_free” dựa trên cột “price”.
* Thay thế các giá trị NaN của cột “size” thành “Varies with device”.
* Bỏ các dòng có “content\_rating” NaN.
* Chuyển đổi cột size thành dạng biến phân loại.
  + Tiny: Nhỏ hơn 20 Mb.
  + Small: Từ 20 Mb đến 100 Mb.
  + Medium: Từ 100 Mb đến 500 Mb.
  + Large: Trên 500 Mb.
  + Varies with device: Không xác định được kích thước cụ thể.
* Do các trò chơi có min\_installs nhỏ hơn 5000 khá ít nên chúng ta sẽ gom các game này vào nhóm 5000.

Tương tự với các game có min\_installs trên 100 triệu, sẽ gom chung vào nhóm 100 triệu

.

1. **Phân tích Khám phá Dữ liệu – EDA**
2. **Xây dựng mô hình**
   1. **Content-based filtering**
      1. **Tf-Idf and Cosine similarity**

Ý tưởng: Dựa trên cột “description” và các cột khác trong mỗi game để tìm xem game nào có nội dung gần giống nhau để tiến hành gợi ý do người dùng.

Sau khi tạo lập ma trận TF-IDF của description, chúng ta có ma trận có kích thước [3509 x 28991] kết hợp thêm các thuộc tính “category”, “content\_rating”, “size”, “avg\_rating”, “rating\_count”, ma trận mở rộng có kích thước [3509 x 29019]. Tiếp theo, áp dụng cosine similarity đế tính ra ma trận độ tương đồng. Kết quả là chúng ta có thể chọn ra TOP các game dựa trên số điểm tương đồng.



Vấn đề:

* Các nhà phát triển trên Google Play thường dùng “Description” để mô tả nội dụng bản cập nhật của họ chứ không phải hoàn toàn là nội dụng trò chơi. Vì vậy, dẫn đến việc gợi ý nội dung không thật sự tương đồng.
* Không áp dụng với tập dữ liệu game có ngôn ngữ khác nhau.
  1. **Collaborative filtering**
     1. **Các phương pháp đánh giá mô hình**
     2. **Item-Based Collaborative Filtering**
     3. **User-Based Collaborative Filtering**

1. **Kết luận**
2. **Định hướng cải tiến**