**BÁO CÁO XỬ LÝ DỮ LIỆU TUYỂN DỤNG VIỆC LÀM**

**1. Mục tiêu**

* Chuẩn hóa dữ liệu tuyển dụng, gom các tiêu đề công việc (job titles) khác nhau nhưng cùng một nghề thành tên nghề chuẩn (ví dụ: "English teacher").
* Chuẩn hóa danh sách kỹ năng (skills) để đảm bảo tính thống nhất.

**2. Các phương pháp đã thử và đánh giá**

Trước khi áp dụng phương pháp cuối cùng, nhóm đã thử nhiều cách tiếp cận để chuẩn hóa dữ liệu, nhưng gặp các vấn đề sau:

* **Dịch văn bản trước khi chuẩn hóa (sử dụng API dịch)**:
  + **Thử nghiệm**: Dịch tiêu đề công việc và kỹ năng từ tiếng Việt sang tiếng Anh (ví dụ: "giáo viên tiếng Anh" thành "English teacher") bằng các API như Google Translate, Microsoft Translator, hoặc DeepL trước khi chuẩn hóa.
  + **Vấn đề**:
    - **Giới hạn số lượng**: Các API dịch miễn phí có giới hạn yêu cầu, không đáp ứng được khối lượng dữ liệu lớn. API trả phí (như Google Cloud Translation) có chi phí cao, không khả thi với ngân sách dự án.
    - **Kết quả dịch không tối ưu**: Các cụm từ chuyên ngành nghề nghiệp thường bị dịch sai. Ví dụ: "kế toán trưởng" thành "near math leader" hoặc "chief mathematician" thay vì "chief accountant".
    - **Thiếu độ chính xác với biến thể ngôn ngữ tự nhiên**: Tiếng Việt trong tuyển dụng có nhiều biến thể (như "giáo viên Anh văn", "dạy tiếng Anh lương cao"), nhưng API dịch không xử lý tốt các cụm từ địa phương hoặc không chuẩn.
* **Dịch văn bản trước khi chuẩn hóa (sử dụng mô hình dịch khác)**:
  + **Thử nghiệm**: Sử dụng các mô hình dịch máy mã nguồn mở (như MarianMT, OPUS-MT) hoặc tự xây dựng mô hình dựa trên Transformer/LSTM, huấn luyện trên dữ liệu song ngữ tiếng Việt - tiếng Anh.
  + **Vấn đề**:
    - **Mô hình mã nguồn mở**: Thiếu dữ liệu huấn luyện đặc thù cho tuyển dụng tiếng Việt, dẫn đến kết quả dịch không chính xác. Ví dụ: "giáo viên tiếng Anh" thành "language teacher" thay vì "English teacher".
    - **Mô hình tự xây dựng**: Yêu cầu dữ liệu song ngữ lớn, thời gian và tài nguyên huấn luyện vượt khả năng dự án. Kết quả vẫn sai, như "kế toán trưởng" thành "accounting leader" hoặc "math supervisor".
    - **Hạn chế chung**: Các mô hình dịch thiếu khả năng hiểu ngữ cảnh nghề nghiệp và xử lý biến thể ngôn ngữ tự nhiên tiếng Việt.
* **Sử dụng mô hình AI có sẵn để chuẩn hóa trực tiếp**:
  + **Thử nghiệm**: Áp dụng các mô hình ngôn ngữ lớn (như BERT, GPT) để chuẩn hóa hoặc hỗ trợ dịch.
  + **Vấn đề**: Thiếu hỗ trợ tốt cho tiếng Việt, kết quả chuẩn hóa tiếng Anh không đồng nhất, sai với các cụm từ phức tạp hoặc biến thể địa phương.
* **Dùng chatbot để chuẩn hóa**:
  + **Thử nghiệm**: Sử dụng chatbot (như Grok hoặc tương tự) để chuẩn hóa từng tiêu đề công việc và kỹ năng.
  + **Vấn đề**: Khối lượng dữ liệu lớn khiến chatbot không xử lý hết. Kết quả không ổn định: lúc đầu đúng, sau đó sai do thiếu ngữ cảnh hoặc thay đổi logic.
* **Các phương pháp khác**:
  + **Từ điển thủ công**: Tạo từ điển ánh xạ tiêu đề công việc tiếng Việt sang tên chuẩn tiếng Anh.
    - **Vấn đề**: Tốn thời gian, khó bao quát biến thể ngôn ngữ.
  + **Biểu thức chính quy (regex)**: Nhận diện mẫu tiêu đề công việc qua regex.
    - **Vấn đề**: Regex thiếu linh hoạt với ngôn ngữ tự nhiên, dễ bỏ sót hoặc nhầm lẫn.
  + **Kiểm tra thủ công (crowdsourcing)**: Thuê nhân sự kiểm tra và chuẩn hóa.
    - **Vấn đề**: Chi phí cao, chậm, dễ xảy ra lỗi con người.
  + **Phân tích ngữ nghĩa đơn giản**: Dùng TF-IDF để tìm từ khóa chính.
    - **Vấn đề**: Không đủ mạnh để xử lý biến thể phức tạp hoặc ngữ cảnh nghề nghiệp.

**Phương pháp cuối cùng**

* **Kết hợp chatbot và chỉnh sửa thủ công**:
  + Sử dụng chatbot để tạo danh sách ban đầu các tiêu đề công việc và kỹ năng chuẩn.
  + Kiểm tra, chỉnh sửa thủ công để bổ sung, sửa lỗi, và xây dựng mảng dữ liệu từ khóa đủ lớn.
  + Dựa trên mảng từ khóa này, tiến hành chuẩn hóa tự động cho toàn bộ dữ liệu.
* **Ưu điểm**: Kết hợp khả năng xử lý ngôn ngữ của chatbot và độ chính xác từ chỉnh sửa thủ công, tạo bộ dữ liệu chuẩn hóa đáng tin cậy.

**3. Phương pháp xử lý chính**

**3.1. Chuẩn hóa tiêu đề công việc (Job Titles)**

* **Bước 1**: Xây dựng danh sách nghề chuẩn (ví dụ: "English teacher").
* **Bước 2**: Tạo mảng từ khóa cho từng nghề, ví dụ: ["giao", "vien", "tieng", "anh", "day", "anh", "van"] cho "English teacher".
* **Bước 3**: Duyệt văn bản tiêu đề, kiểm tra từ khóa và gán vào nghề chuẩn.

**3.2. Chuẩn hóa kỹ năng (Skills)**

* **Bước 1**: Xây dựng danh sách kỹ năng chuẩn và từ khóa (ví dụ: "Teaching English" với "dạy tiếng Anh").
* **Bước 2**: Tạo mảng từ khóa cho kỹ năng.
* **Bước ivit 3**: Duyệt văn bản job description, gán kỹ năng chuẩn nếu từ khóa xuất hiện.

**3.3. Quy trình kỹ thuật**

* **Công cụ**: PySpark, Python (thư viện unicodedata, re).
* **Các bước**:
  + Chuẩn hóa văn bản: loại bỏ dấu tiếng Việt, chuyển chữ thường, bỏ ký tự đặc biệt.
  + Đọc dữ liệu từ file CSV: danh sách kỹ năng chuẩn và dữ liệu tuyển dụng.
  + Xử lý kỹ năng: broadcast danh sách kỹ năng chuẩn, dùng UDF để gán kỹ năng chuẩn vào cột job\_skills.
  + Thống kê: đếm số dòng kỹ năng rỗng/null, tính tỷ lệ phần trăm.

**4. Kết quả**

* **Tổng số dòng**: 114,287.
* **Số dòng có skills\_standard rỗng**: 281.
* **Tỷ lệ phần trăm dòng có skills\_standard rỗng**: 0.25%.
* **Nhận xét**: Chuẩn hóa job titles và skills đạt hiệu quả cao, với tỷ lệ dữ liệu thiếu rất thấp (0.25%), cho thấy dữ liệu đầu vào chất lượng tốt và quy trình xử lý tối ưu.

**5. Kết luận**

Sau khi thử nhiều phương pháp, bao gồm dịch văn bản bằng API và mô hình dịch khác nhưng không hiệu quả do giới hạn số lượng, kết quả không tối ưu, và thiếu chính xác với biến thể tiếng Việt, phương pháp kết hợp chatbot và chỉnh sửa thủ công đã tạo ra bộ dữ liệu từ khóa đủ lớn để chuẩn hóa job titles và skills. Quy trình đạt tỷ lệ lỗi cực thấp (0.25%).