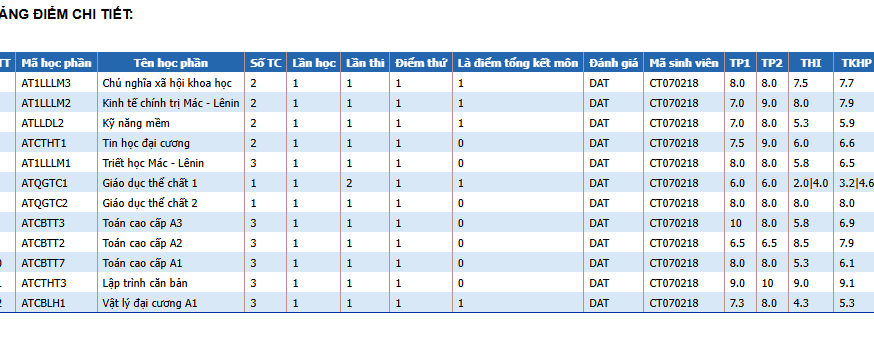
1.4: Khảo sát

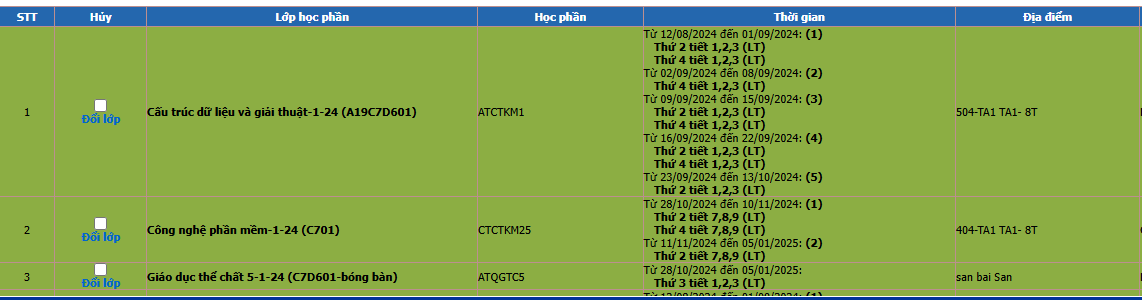
1.4.1: Phương pháp khảo sát

\* Khảo sát thực tế

- Dựa trên tình hình thực tế và khảo sát tại chính Học Viện Kỹ Thuật Mật Mã, nhóm chúng tôi đã khảo sát 10 bạn sinh viên ngẫu nhiên của từng khóa, từng lớp và kết quả chúng tôi nhận được là có tới 8/10 bạn sinh viên nói rằng việc không thể theo dõi sát sao điểm số thi của từng môn, các môn đã hoàn thành, các môn còn nợ, GPA, CPA. Việc sinh viên tới mùa đăng ký tín chỉ thì khá là khó khăn theo dõi các lớp dự định đăng ký. Bên cạnh đó nhóm chúng tôi có phỏng vấn một vài giảng viên về việc theo dõi phổ điểm hiện tại là còn đang gặp một chút khó khăn, việc theo dõi phổ điểm đối với giảng viên rất quan trọng trong việc điều chỉnh phương pháp dạy học mới.

+, Bạn Nguyễn Ngọc Hà – CT7B: Theo như mình thấy thì việc theo dõi điểm GPA hiện tại của mình trên website qldt edu vn của trường mình đang chưa cập nhật điểm của sinh viên cho tới hiện tại bây giờ. Bây giờ sắp sang năm 2025-2026 rồi nhưng điểm chỉ còn đang được update ở năm 2022-2023 mà thôi. Mình thấy khá là bất tiện về việc theo dõi điểm như vậy, mỗi lúc như vậy thì mình cần phải lên gặp cô giáo viên chủ nhiệm để xin cô phiếu điểm mới hiện tại của mình. Việc đó làm mình cảm thấy khá là bất tiện vì thời gian biểu của sinh viên và giáo viên là không hề giống nhau chút nào, cô có thể có nhiều lúc bận công chuyện hoặc đi họp và trở ngại lớn nhất là số lượng sinh viên dồn lên phòng cô khoảng vào thời gian xét học bổng khá nhiều vì họ muốn xem bảng điểm và tính toán xem liệu khả năng mình có nằm trong danh sách xét học bổng kỳ vừa rồi đó không.

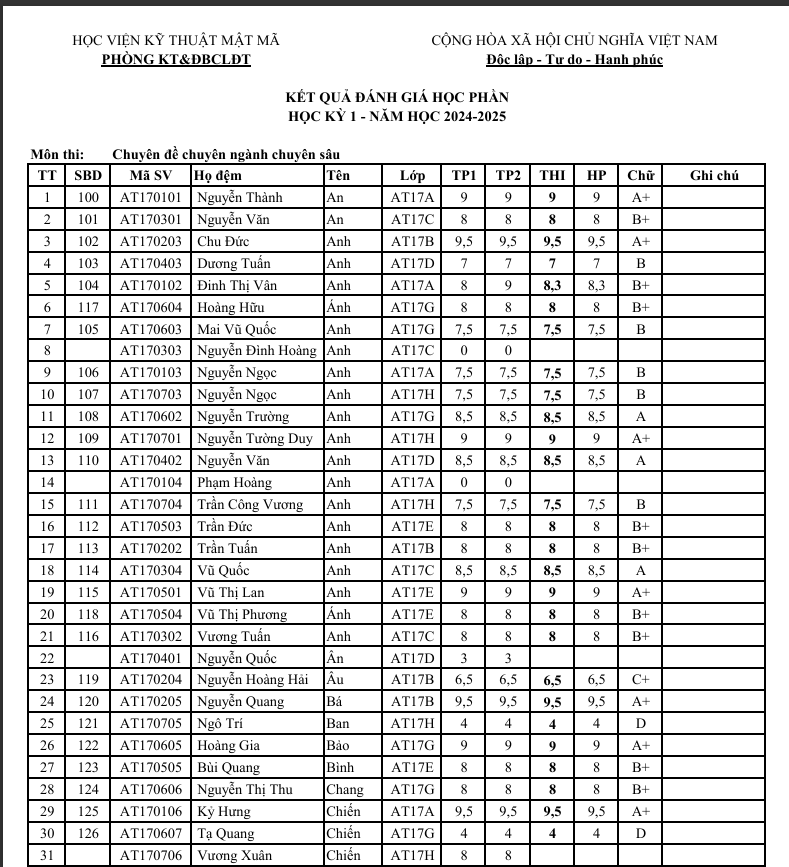


+, Bạn Nguyễn Huy Hoàng – CT7B: Mình vừa bị đi nhầm lớp vào phòng ở tầng 402-TA1 nhưng thực tế hôm nay mình học Cấu trúc dữ liệu và giải thuật ở tầng 502-TA1 cơ, việc theo dõi giao diện thời gian biểu trên website qldt edu vn của trường mình chưa được khoa học lắm nên việc sinh viên bị nhầm phòng hoặc nhầm tiết rất là nhiều.

+, Bạn Đỗ Minh Đăng – CT7B: Mình vừa bị trượt môn kì vừa rồi mà mình chẳng nhớ là môn gì, giờ mình chẳng biết phải làm sao để có thể biết được đó là môn gì và để chuẩn bị tâm lý, kiến thức để bắt đầu thi lại môn đó đây. Khéo chắc lát mình phải lên trên phòng cô chủ nhiệm xin lại bảng điểm thôi, việc không theo dõi được những môn học còn nợ nó có thể dẫn tới việc mình bỏ lỡ thông tin bị nhận cảnh cáo của trường.

+, Anh Nguyễn Văn Minh – CT5A: Anh đang không biết là tổng số môn/tín chỉ anh hoàn thành đã đủ điều kiện để ra trường chưa. Việc theo dõi số tín chỉ đã hoàn thành rất quan trọng nhưng web trường vẫn chưa cập nhật được. Nên đa số sinh viên khóa cuối bọn anh phải nhờ tới cô chủ nhiệm để theo dõi việc đó, hơi bất tiện cho cô và trò em ạ.

+, Thầy Lê Đức Thuận: Thầy đang gặp vấn đề về việc xem phổ điểm của môn Nguyên lý hệ điều hành năm ngoái, không biết rằng sinh viên có làm được bài không. Thầy đang muốn thay đổi một chút phương pháp dạy học để nâng cao, cải thiện chất lượng của sinh viên.



**1.4.2: Thu thập dữ liệu**

**1. Nội dung**

Thu thập dữ liệu là bước khởi đầu quan trọng và mang tính nền tảng trong dự án, nhằm mục tiêu tổng hợp, quản lý thông tin điểm số của sinh viên từ nhiều nguồn khác nhau một cách có hệ thống và hiệu quả.

Các nguồn dữ liệu chính bao gồm:

* **Hệ thống quản lý đào tạo (QLĐT)** của Học viện Kỹ thuật Mật mã.
* **Các nhóm chia sẻ thông tin khảo thí** trên các nền tảng mạng xã hội.
* **Các file PDF điểm số** được cung cấp trực tiếp bởi giảng viên hoặc cán bộ nhà trường.

Đặc điểm nổi bật của dữ liệu thu thập được là sự đa dạng về cấu trúc. Các file PDF thường được thiết kế không đồng nhất, tùy thuộc vào nguồn phát hành, bộ môn, hoặc kỳ học. Điều này tạo ra thách thức lớn trong việc nhận diện, chuẩn hóa và tích hợp dữ liệu.

Để vượt qua những thách thức này, dự án đã tập trung phát triển một **service tổng quát** với khả năng xử lý mọi định dạng file PDF và các form dữ liệu khác nhau. Cụ thể, service được xây dựng trên các công nghệ hiện đại, bao gồm:

* **Thư viện phân tích PDF chuyên dụng** (như Java PDFReader), hỗ trợ xử lý văn bản và trích xuất dữ liệu từ các form phức tạp.
* **Thuật toán chuẩn hóa dữ liệu**, giúp đồng bộ hóa các thông tin như mã sinh viên, điểm số, tên môn học, và kỳ học.

Sau khi hoàn tất quá trình xử lý, dữ liệu được chuẩn hóa sẽ được lưu trữ vào hệ thống cơ sở dữ liệu. Việc xây dựng một quy trình tự động hóa này không chỉ giúp tiết kiệm thời gian mà còn giảm thiểu tối đa sai sót trong quá trình nhập liệu thủ công. Đồng thời, hệ thống lưu trữ được thiết kế để đảm bảo tính nhất quán, dễ dàng truy vấn, và phục vụ tốt cho việc quản lý và báo cáo dữ liệu học tập.

**2. Phạm vi**

**Nguồn dữ liệu**

Dữ liệu được thu thập từ:

* Hệ thống khảo thí trực tuyến.
* Các nhóm trao đổi thông tin khảo thí.
* Các file điểm dạng PDF do giảng viên và cán bộ cung cấp.

**Định dạng và cấu trúc file**

* Dữ liệu được lưu trữ dưới dạng file PDF.
* Cấu trúc file đa dạng, thay đổi theo môn học, lớp học, và kỳ học.

**Xử lý dữ liệu**

* Sử dụng **thư viện Java PDFReader** để đọc và trích xuất dữ liệu quan trọng như mã sinh viên, điểm số, tên môn học, và kỳ học.
* Xây dựng các thuật toán nhận diện form, đảm bảo nhận diện chính xác **3 định dạng form khác nhau**.

**Lưu trữ dữ liệu**

* Dữ liệu sau khi xử lý được lưu trữ trong cơ sở dữ liệu MySQL.
* Quản lý cơ sở dữ liệu thông qua **HeidiSQL** và **DB Schema** để tối ưu hóa việc tổ chức và truy vấn thông tin.

**3. Kết luận**

Dự án không chỉ dừng lại ở việc thu thập dữ liệu điểm số từ nhiều nguồn mà còn hướng đến việc thiết kế và triển khai một hệ thống mạnh mẽ, linh hoạt để xử lý, lưu trữ và quản lý dữ liệu hiệu quả. Hệ thống này tạo ra một nền tảng vững chắc, hỗ trợ việc:

* **Tự động hóa báo cáo và phân tích dữ liệu học tập.**
* **Tối ưu hóa công tác quản lý thông tin học tập và khảo thí.**
* **Đáp ứng nhu cầu tra cứu nhanh chóng, chính xác của giảng viên và sinh viên.**

Thông qua việc sử dụng công nghệ tiên tiến và quy trình bài bản, dự án đã tạo ra một công cụ mạnh mẽ, giúp nâng cao chất lượng quản lý dữ liệu học tập và đóng góp tích cực vào sự phát triển của Học viện.

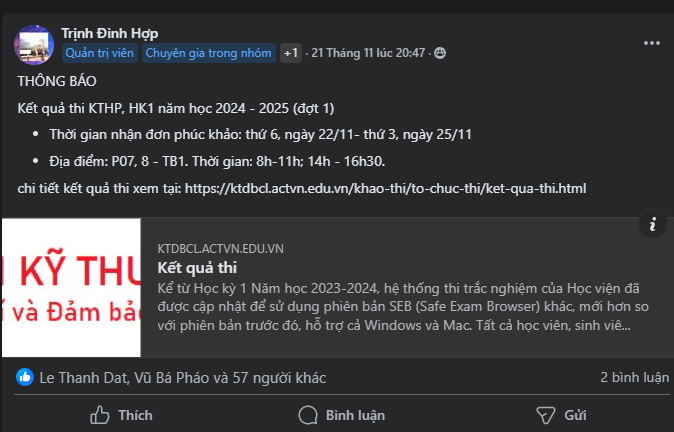
**1.4.3: Báo cáo dữ liệu**

**1. Tổng quan về quá trình thu thập dữ liệu**

Từ năm 2020 đến năm 2024 có rất nhiều khóa và nhiều file điểm đã được public lên trên mạng xã hội sinh viên có thể xem nhưng rất khó để gộp lại tổng quát, nhóm chúng tôi đã triển khai một kế hoạch thu thập và xử lý dữ liệu điểm số một cách có hệ thống từ nhiều nguồn khác nhau. Nguồn dữ liệu bao gồm:

* **Website khảo thí chính thức** của Học viện Kỹ thuật Mật mã.
* **Các nhóm chia sẻ thông tin khảo thí** trên các nền tảng mạng xã hội.

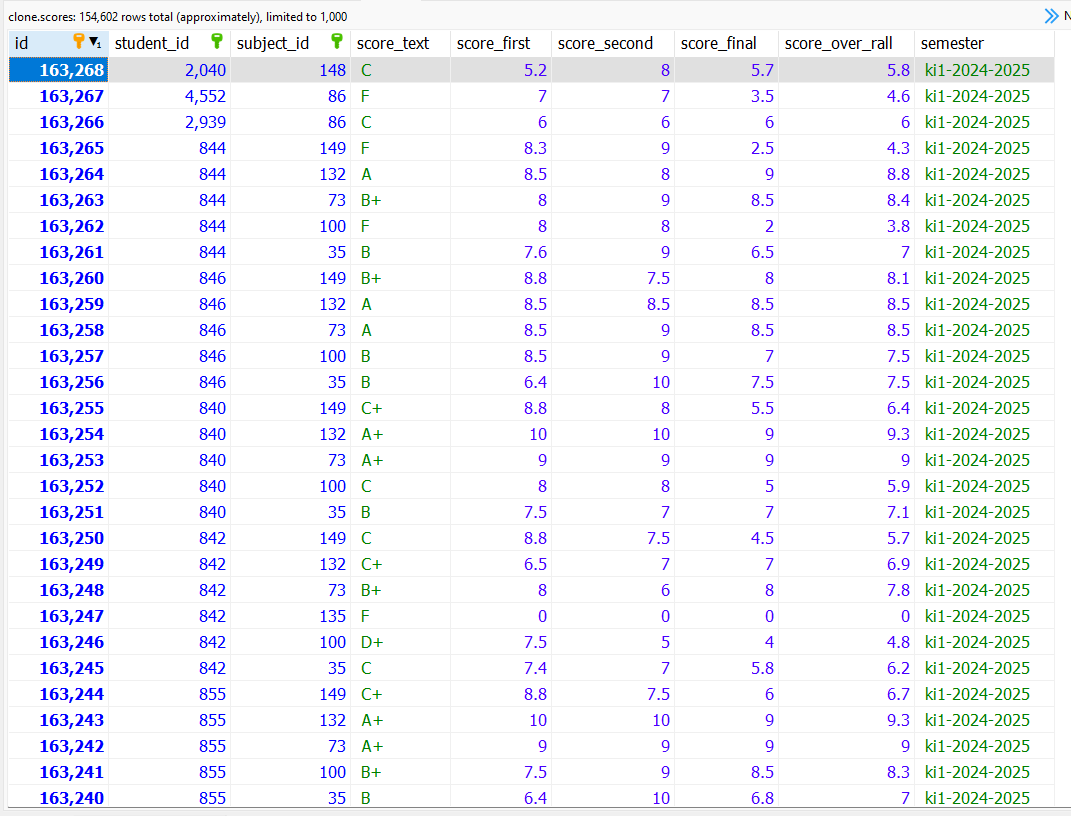
Dữ liệu chủ yếu được cung cấp dưới dạng file PDF, với tổng dung lượng đạt **200 MB**, bao gồm **35 file** chứa thông tin chi tiết về điểm số của sinh viên. Các file này có cấu trúc không đồng nhất, phụ thuộc vào ngành học, môn học, và từng kỳ học cụ thể.



**2. Kết quả thu thập và xử lý dữ liệu**

Sau khi hoàn tất quá trình xử lý, chúng tôi đã đạt được các kết quả đáng chú ý như sau:

* **Tổng số bản ghi thu thập được**: 154.602, tương ứng với điểm số của sinh viên cho từng môn học.



* **Số lượng sinh viên**: 4.814 sinh viên, thuộc ba ngành đào tạo chính:
  + **An toàn thông tin**.
  + **Công nghệ thông tin**.
  + **Điện tử viễn thông**.



* **Số lượng môn học**: 181 môn, bao gồm tất cả các môn học của nhà trường trong giai đoạn 2020–2024.



* **Tỷ lệ trích xuất dữ liệu thành công**: Đạt 80%, với 20% lỗi phát sinh từ sự đa dạng định dạng file PDF và các vấn đề về chất lượng dữ liệu.

**3. Phân tích chi tiết**

**3.1. Độ chính xác của dữ liệu**  
Quá trình thu thập và xử lý đạt độ chính xác **96%**, với phần lớn dữ liệu được chuẩn hóa và lưu trữ chính xác trong cơ sở dữ liệu. Tuy nhiên, vẫn có một số trường hợp cần điều chỉnh do sự thay đổi cấu trúc file PDF theo thời gian.

**3.2. Định dạng dữ liệu và số form nhận diện**  
Hệ thống đã nhận diện và xử lý thành công **3 form dữ liệu chính**, tương ứng với các cấu trúc từ các nguồn khác nhau. Điều này đòi hỏi giải pháp kỹ thuật phải linh hoạt và liên tục cải tiến để đáp ứng sự thay đổi.

**3.3. Sự phân bổ theo năm và ngành**

* Dữ liệu thu thập được phân bổ không đồng đều, phụ thuộc vào từng năm học và ngành đào tạo.
* Các ngành có lượng dữ liệu lớn nhất:
  + **An toàn thông tin**: 60 môn học.
  + **Công nghệ thông tin**: 62 môn học.
  + **Điện tử viễn thông**: 59 môn học.

**4. Công nghệ sử dụng và quy trình xử lý**

**4.1. Công nghệ chính**

* **Java PDF Reader**: Dùng để phân tích và trích xuất dữ liệu từ file PDF.
* **Cơ sở dữ liệu quan hệ (MySQL)**: Lưu trữ và quản lý dữ liệu, đảm bảo khả năng truy vấn và báo cáo chính xác.

**4.2. Quy trình thực hiện**

1. **Thu thập dữ liệu**: Tổng hợp file từ các nguồn chính như website khảo thí và nhóm chia sẻ thông tin.
2. **Tiền xử lý**: Chuẩn hóa file PDF, nhận diện cấu trúc dữ liệu và loại bỏ thông tin không cần thiết.
3. **Trích xuất dữ liệu**: Sử dụng công nghệ trích xuất văn bản để lấy thông tin như mã sinh viên, điểm số, môn học và kỳ học.
4. **Lưu trữ và kiểm thử**: Lưu trữ dữ liệu vào cơ sở dữ liệu, đồng thời kiểm thử để đảm bảo tính chính xác và toàn vẹn dữ liệu.

**5. Thách thức và cải tiến**

**5.1. Thách thức**

* **Sự thay đổi định dạng file PDF**: Định dạng file PDF khác nhau qua từng năm đòi hỏi phải thường xuyên điều chỉnh code.
* **Chất lượng file không đồng đều**: Một số file PDF bị lỗi, thiếu thông tin hoặc chứa các ký tự không thể đọc được.

**5.2. Cải tiến**

* **Áp dụng thuật toán mới**: Tăng tỷ lệ trích xuất thành công lên trên 90%, đảm bảo hiệu quả xử lý cao hơn.
* **Tự động hóa kiểm tra và sửa lỗi**: Giảm thiểu sự can thiệp thủ công, tối ưu hóa quá trình xử lý dữ liệu.
* **Tích hợp công nghệ AI**: Phân tích và dự đoán các lỗi phổ biến để tăng cường khả năng nhận diện và xử lý dữ liệu.

**6. Kết luận và định hướng tương lai**

**Kết luận**  
Dự án đã hoàn thành xuất sắc mục tiêu thu thập và xử lý dữ liệu điểm số từ năm 2020 đến năm 2024. Kết quả đạt được không chỉ phản ánh khả năng xử lý dữ liệu lớn với độ chính xác cao mà còn xây dựng nền tảng vững chắc cho các ứng dụng báo cáo và phân tích dữ liệu học tập.

**Định hướng tương lai**  
Trong thời gian tới, dự án sẽ tập trung vào:

1. **Tối ưu hóa hệ thống**: Cải tiến hiệu suất xử lý, nâng cao khả năng nhận diện và trích xuất dữ liệu.
2. **Mở rộng quy mô**: Xây dựng hệ thống phù hợp với dữ liệu đa dạng hơn, phục vụ cho các nhu cầu phát sinh mới của nhà trường.
3. **Ứng dụng phân tích dữ liệu**: Hỗ trợ phân tích học tập chuyên sâu, cung cấp thông tin hữu ích cho việc cải thiện chất lượng đào tạo.