CUỘC THI KHOA HỌC DỮ LIỆU UIT 2023

Khoa Khoa học và Kỹ thuật Thông tin

Trường Đại học Công nghệ Thông tin, ĐHQG-HCM

Hồ Chí Minh, Ngày 31 tháng 08 năm 2023

Bảng A: Giải pháp Khoa học Dữ liệu cho Thành phố Thông minh

Nội dung

- Dựa trên thành tựu của **Khoa học Dữ liệu/Trí tuệ Nhân tạo.**
- **Có khả năng** giải quyết một hoặc nhiều vấn đề trong cuộc sống.
- Hình thức sản phẩm: Desktop Apps, Mobile Apps, Web Apps, Robotics...
- **Ví dụ:**
  - Ứng dụng điểm danh tự động dựa trên việc nhận diện gương mặt.
  - Ứng dụng phát hiện và thông báo vị trí ùn tắc giao thông đến cho đội CSGT.
  - Ứng dụng phát hiện và báo cháy.
  - Ứng dụng phát hiện hành vi bạo lực trong trường học.
  - Ứng dụng nhận diện cá nhân xả rác ra đường phố.
  - Ứng dụng hỗ trợ học tiếng Anh như Elsa, Cake, ...
  - ...

Tiêu chí đánh giá

- Có khả năng giải quyết một hoặc nhiều vấn đề hiện có trong thực tế **(hệ số 0.4)**.
- Tính mới **(hệ số 0.4)**.
- Phù hợp với truyền thống, văn hóa của dân tộc Việt Nam và đạo đức con người **(hệ số 0.2)**.

Tiêu chí đánh giá

- **Riêng giải đặc biệt**, các ứng dụng phải thỏa mãn các tiêu chí sau:
  - Ứng dụng phải được đánh giá từ 0.9 trở lên theo 03 tiêu chí nêu trên.
  - Ứng dụng phải được chấp nhận đăng trên **một trong hai** nền tảng ứng dụng phổ biến: **CH Play** và **App Store**.
  - Ứng dụng phải đạt được **tối thiểu 1000 lượt tải về**
  - Không vi phạm bản điều lệ cho nhà phát triển của nền tảng tương ứng đối với Android và IOS/MacOS.

Nộp sản phẩm

- Nộp sản phẩm về Ban Tổ chức qua dạng files:
  - **01 file thuyết minh** (theo mẫu đính kèm) trình bày cụ thể về sản phẩm.
  - **01 file poster kích thước 0,8x1,3m (**theo mẫu đính kèm) khái quát về sản phẩm.
  - **01 video thuyết trình** sản phẩm, thời lượng tối đa 05 phút.
  - **01 video demo** các tính năng của sản phẩm, thời lượng tối đa 05 phút.

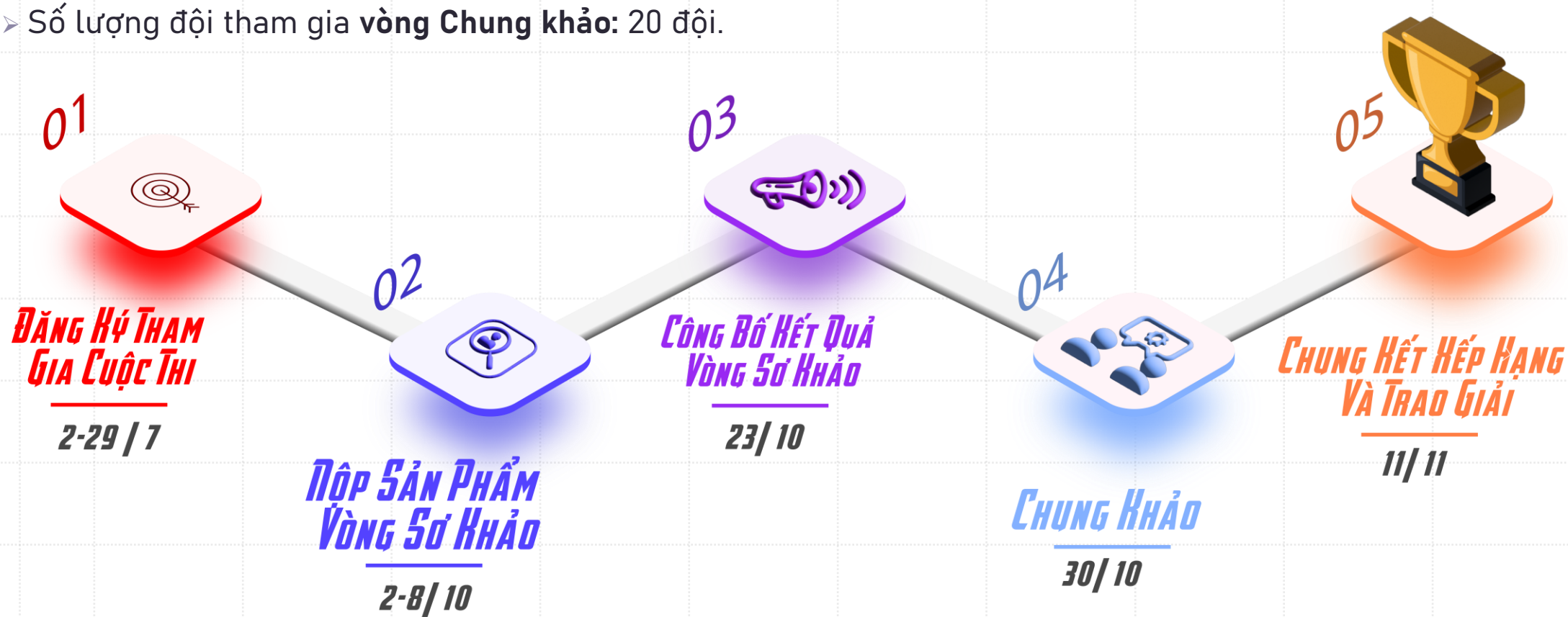
Nộp sản phẩm

- Tất cả các files được nén lại thành **01 file zip** duy nhất.
- Đặt tên file zip theo cú pháp **DSC2023_<tên nhóm>.zip**, trong đó <tên nhóm> là tên của nhóm đã đăng ký tham gia. Ví dụ: nhóm top1dsc sẽ đặt tên file zip là DSC2023_top1dsc.zip.
- Trường hợp tên nhóm **có khoảng trắng** " " thì thay khoảng trắng bằng **ký hiệu gạch nối** "-". Ví dụ: nhóm AI Warriors sẽ đặt tên file zip là DSC2023_AI-Warriors.zip.

Quy trình đánh giá – xếp hạng

1. Vòng Sơ khảo

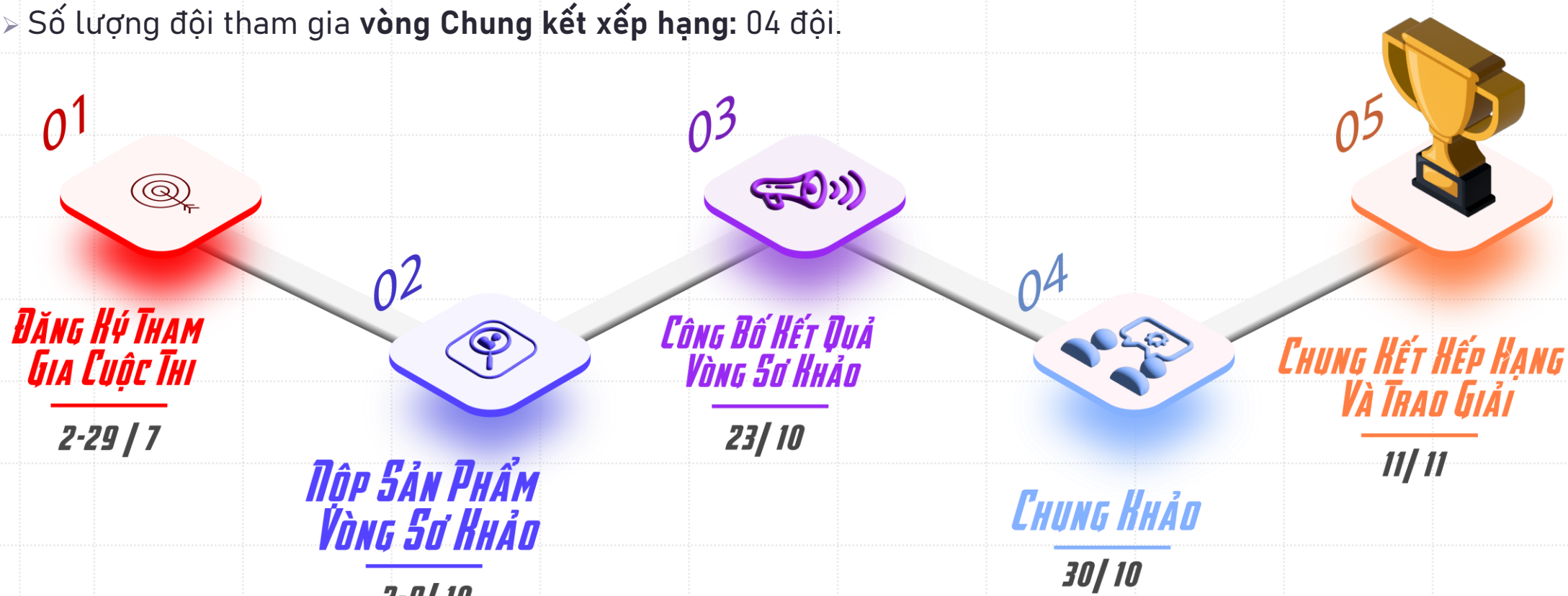
- Các đội thi **thực hiện đề tài** và **nộp sản phẩm** chuyển về cho Ban tổ chức **trước ngày 8/10/2023.**
- Ban tổ chức sẽ tiến hành **xét duyệt, đánh giá và chọn lọc** các đề tài **đủ điều kiện** tham gia vòng Chung khảo.
- Số lượng đội tham gia **vòng Chung khảo:** 20 đội.



Quy trình đánh giá – xếp hạng

2. Vòng Chung khảo

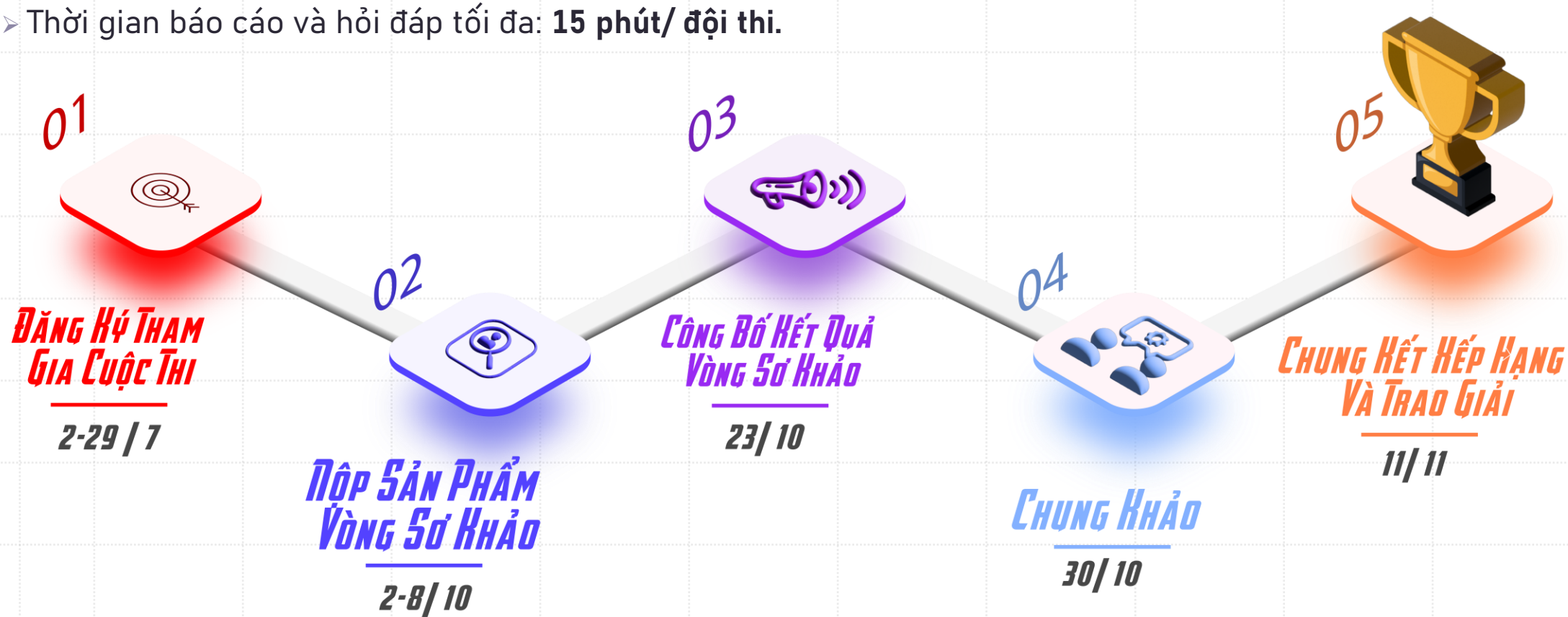
- Các đội thi **thực hiện báo cáo đề tài** theo **poster** tại **Trường ĐH Công nghệ Thông tin – ĐHQG TP. HCM**.
- Ban tổ chức lập hội đồng khoa học **chấm điểm và chọn lọc** các đề tài tham gia vòng Chung kết xếp hạng.
- Số lượng đội tham gia **vòng Chung kết xếp hạng:** 04 đội.



Quy trình đánh giá – xếp hạng

3. Vòng Chung kết xếp hạng

- Ban tổ chức lập **01** hội đồng khoa học để **chất vấn** và **chấm giải** cho các đội thi.
- Các đội thi **trình bày, bảo vệ** sản phẩm trực tiếp trước hội đồng tại **Trường ĐH Công nghệ Thông tin**.
- Thời gian báo cáo và hỏi đáp tối đa: **15 phút/ đội thi.**



Sản phẩm công nghệ được yêu thích nhất

- Tham gia bình chọn cho đề tài được yêu thích nhất theo các bước:
  - **Bước 1:** Like fanpage **UIT Data Science Challenge** (https://www.facebook.com/uit.dsc)
  - **Bước 2:** Reactions/thả cảm xúc và chia sẻ cho **poster** đề tài mà bạn bình chọn.
    - 1 Reactions/thả cảm xúc: **3 điểm** (Không giới hạn số điểm tối đa)
    - 1 chia sẻ: **5 điểm** (Tối đa 10.115 điểm)

- **Lưu ý:**
  - Thực hiện **ĐẦY ĐỦ** 2 bước thì mới hợp lệ.
  - Tinh thần Khoa học và tuyệt đối Trung thực bình chọn cho đề tài được yêu thích.

  (BTC sẽ **KHÔNG** tính điểm những thí sinh/nhóm thí sinh có **hành vi gian lận**).

Sản phẩm công nghệ được yêu thích nhất

- **01** đề tài có điểm bình chọn **cao nhất** sẽ nhận được giải **"Sản phẩm công nghệ được yêu thích nhất".**
- Đề tài sẽ nhận được Quà tặng, Giấy khen từ Trường ĐH Công nghệ Thông tin – ĐHQG TP. HCM.
- Đề tài đó sẽ được Ban tổ chức **mời 01 đại diện** tham dự **Vòng chung kết Cuộc thi Khoa học Dữ liệu UIT 2023** (trong trường hợp đề tài không được vào Vòng chung kết).
- **04** đề tài có điểm bình chọn cao tiếp theo sẽ được nhận GCN **"Top 5 đề tài công nghệ được yêu thích nhất".**

Lưu ý: kết quả giải ***Sản phẩm công nghệ được yêu thích nhất*** sẽ được công bố ở lễ trao giải của cuộc thi.

Cơ cấu giải thưởng

- Giải đặc biệt: 25,000,000 VNĐ.
- Giải nhất: 15,000,000 VNĐ.
- Giải nhì: 5,000,000 VNĐ.
- Giải ba: 3,000,000 VNĐ.
- Giải khuyến khích: 1,000,000 VNĐ.
- Giải sản phẩm được yêu thích nhất: 1,000,000 VNĐ.

Trao thưởng

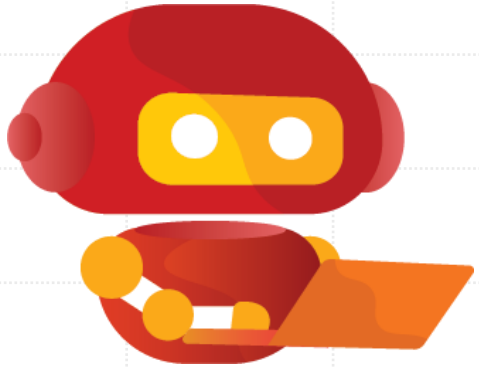
- Các đội tham gia Cuộc thi ở **Vòng Sơ loại** sẽ nhận **01 Giấy chứng nhận tham gia cuộc thi UIT Data Science Challenge 2023** do Trường ĐH Công nghệ Thông tin – ĐHQG TP. HCM cấp.
- Các đội thi xuất sắc **vào vòng chung khảo** nhận **01 Giấy chứng nhận đã tham gia vòng Chung khảo Bảng A** của Cuộc thi UIT Data Science Challenge 2023.

Trao thưởng

- Các đội đạt giải tại vòng chung kết:
  - Hiện kim.
  - 01 kỷ niệm chương.
- Vòng chung kết Bảng A sẽ được đồng tổ chức với lễ bế mạc cuộc thi UIT Data Science Challenge 2023 tại hội trường E **ngày 11/11/2023**.

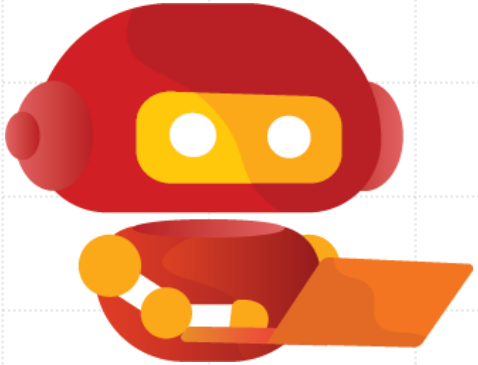
Bảng B: Cuộc khi Khoa học Dữ liệu

Chủ đề: Kiểm tra thông tin dựa trên một văn bản



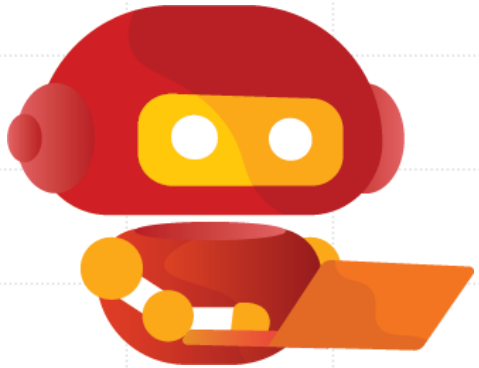
Nội dung

- Các nhóm tham gia bảng B sẽ được cung cấp **01 bộ dữ liệu** thông qua đường link dẫn đến trang **Codalab** của Cuộc thi Khoa học Dữ liệu UIT 2023.
- Các nhóm dựa trên bộ dữ liệu này **đề xuất và huấn luyện** phương pháp để giải quyết bài toán mà bộ dữ liệu đặt ra.



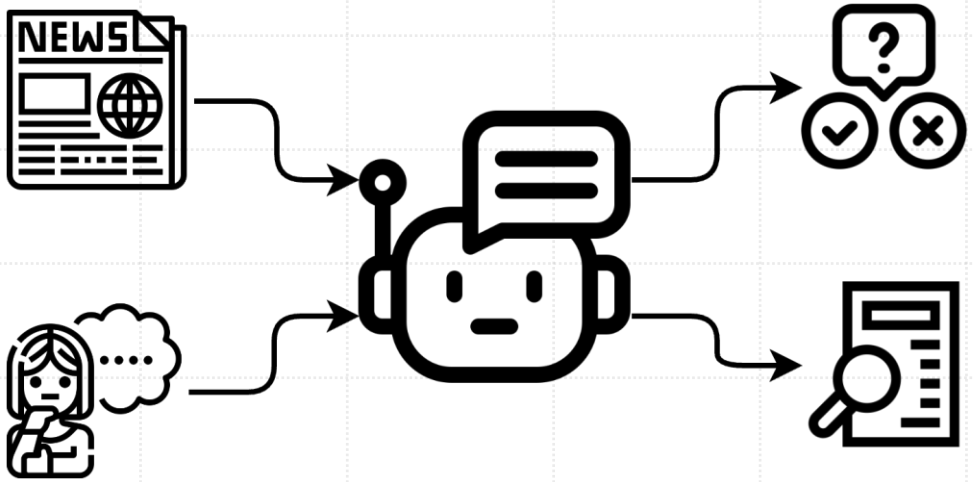
Nội dung

- **Chỉ được sử dụng** bộ dữ liệu do Ban Tổ chức cung cấp để huấn luyện phương pháp đề xuất.
- **Không sử dụng** các kỹ thuật tăng cường dữ liệu và gán nhãn dữ liệu thủ công.
- **Chỉ sử dụng** các mô hình ngôn ngữ huấn luyện trước được Ban Tổ chức công bố.
- Danh sách các mô hình ngôn ngữ huấn luyện trước được xây dựng từ danh sách các các mô hình ngôn ngữ mà mỗi nhóm tham gia khai báo với Ban Tổ chức sẽ sử dụng trong cuộc thi.
- Form đăng ký mô hình ngôn ngữ huấn luyện trước sẽ được gửi đến cho các nhóm đăng ký tham gia qua email và thông tin trên Fanpage cũng như Website của cuộc thi. Các đội cần hoàn tất đăng ký **trước ngày 15/09/2023**.



Nội dung

- Bài toán: Kiểm tra tính chính xác của thông tin dựa trên một văn bản.
- Bộ dữ liệu: **ISE-DSC01**.
- Bộ dữ liệu gồm 3 phần: training set, public test set, private test set.
- Yêu cầu: Với input là **một văn bản** và một câu **claim**, hãy kiểm tra câu claim đó là **SUPPORTED**, **REFUTED** hay **NEI** (Not Enough Information). Nếu là SUPPORTED hay REFUTED, cần **chỉ ra dẫn chứng bằng** cách trích dẫn 01 câu từ văn bản.



Nội dung

Cấu trúc file json:

{

"id": {

"context": "văn phòng quận yongsan, trung tâm thủ đô ... ",

"claim": "xấp xỉ một triệu khách du lịch đã đến Hàn Quốc mỗi năm vì lễ hội ẩm thực toàn cầu",

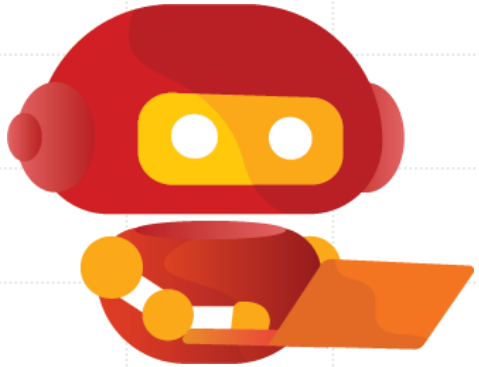
"verdict": "SUPPORTED",

"evidence": "văn phòng quận cũng sẽ thực hiện ..."

},

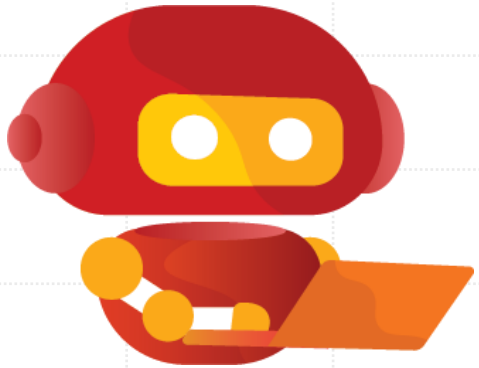
...

}



Nội dung

- Phương pháp được đề xuất bởi các nhóm sẽ được đánh giá như sau:
  - Bài toán được chia thành **hai quá trình**: (1) **xác minh** câu claim được củng cố (SUPPORTED) hay bị bác bỏ (REFUTED) hay không thể xác minh (NEI) bởi văn bản, và (2) **đưa ra dẫn chứng** nếu câu claim đó được cho là SUPPORTED hoặc REFUTED bằng cách trích xuất 1 câu từ văn bản.
  - Quy trình đánh giá các phương pháp cũng được chia làm **2 bước**: (1) xác định phương pháp được đề xuất **xác minh đúng** câu claim và (2) **xác định câu dẫn chứng** mà phương pháp đưa ra là chính xác.



Nội dung

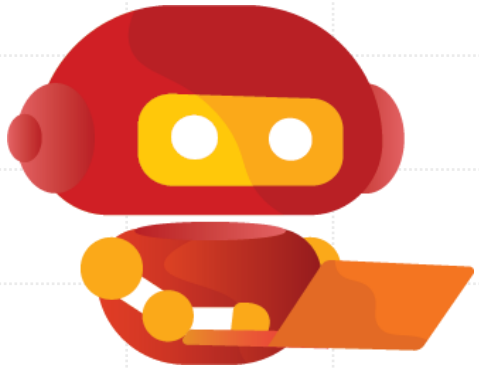
- Thứ hạng của mỗi phương pháp sẽ được xác định dựa trên **độ đo Strict Accuracy.**

  Gọi v và v' lần lượt là verdict mẫu và verdict được dự đoán $(v, v' \in \{SUPPORTED, REFUTED, NEI\})$.

  Gọi e và e' lần lượt là evidence mẫu và evidence được dự đoán.

  $$\mathrm{StrAcc} = \delta(\mathrm{v}, \mathrm{v}') \times \delta(e, e')$$

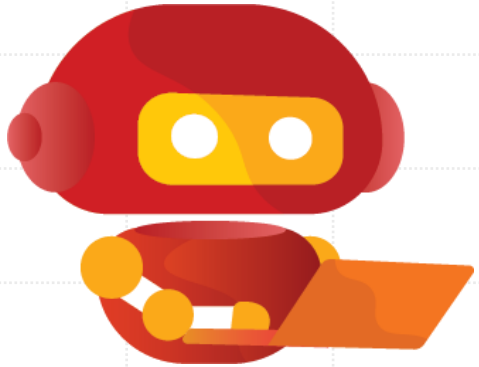
  Trong đó δ là ký hiệu Kronecker với $\delta(x, y) = 1 \Leftrightarrow x = y$ và $\delta(\mathrm{x}, \mathrm{y}) = 0 \Leftrightarrow \mathrm{x} \neq \mathrm{y}$ và StrAcc là Strict Accuracy.



- Kết quả cuối cùng sẽ được xác định trên **private test set**.

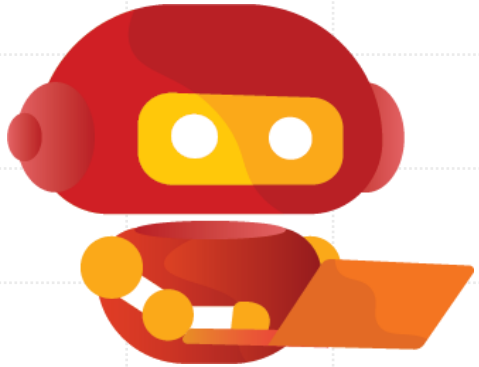
Cơ cấu giải thưởng

- Giải nhất: 15,000,000 VNĐ.
- Giải nhì: 5,000,000 VNĐ.
- Giải ba: 3,000,000 VNĐ.
- Giải khuyến khích 1: 1,000,000 VNĐ.
- Giải khuyến khích 2: 1,000,000 VNĐ.



Trao thưởng

- Các đội đạt giải tại vòng chung kết sẽ được nhận hiện kim và **01 kỷ niệm chương** từ Ban Tổ chức.
- Các đội không đạt giải sẽ được nhận **01 giấy chứng nhận** đã tham gia Bảng B cuộc thi UIT Data Science Challenge 2023.
- Vòng chung kết Bảng B sẽ được đồng tổ chức với lễ bế mạc cuộc thi UIT Data Science Challenge 2023 tại hội trường E ngày **11/11/2023**.



Chúc các nhóm đạt giải cao nhất tại Cuộc thi