Cần xây dựng một hệ thống trích rút và lưu trữ các thực thể và các quan hệ từ các bài báo mạng. Các thực thể trích rút được bao gồm người (PERSON), tổ chức (ORGANIZATION), địa danh (LOCATION), đất nước (COUNTRY), sự kiện (EVENT), thời gian (TIME), Hiệp định/hiệp ước/thỏa tuận (AGGREMENT)

Mỗi thực thể có chung các thông tin tối thiểu như sau:

1. Định danh
2. Nhãn-tên hiển thị
3. Mô tả

Ngoài ra, mỗi loại có thể có thêm một vài thuộc tính khác (tùy sinh viên thiết kế thêm, sao cho logic và hợp lý)

Mỗi bài báo (ARTICLE) trích xuất ra các FACT, là một bộ ba subject-quan hệ-object (subject, object là các thực thể trích rút được, ứng với các loại thực thể kể trên). SV tự xác định các thuộc tính cần thiết để lưu trữ FACT và ARTICLE (ví dụ với ARTICLE cần có link bài báo, với FACT cần có ngày tháng trích rút, subject, object, relationship)

Một số ví dụ sẽ được trình bày dưới đây. Lưu ý là các ví dụ chỉ đưa ra thông tin chi tiết về các thuộc tính cho một số thực thể.

Ví dụ 1: Trong một bài tin tức có đoạn "Hôm qua, 12/9/2017 Apple đã ra mắt Iphone 8". Thông tin trích xuất được có thể là:

- Thực thể ORGANIZATION: **Apple.**
  - **Định danh**: Apple00001 (có thể có nhiều thực thể có nhãn hiển thị là Apple, nếu khác nhau cần có định danh riêng)
  - **Nhãn hiển thị**: Apple
  - **Mô tả**: Apple là tập đoàn công nghệ máy tính của Mỹ có trụ sở chính đặt tại Cupertino, California. Apple được thành lập ngày 1 tháng 4 năm 1976 dưới tên Apple Computer, Inc., và đổi tên thành Apple Inc. vào đầu năm 2007.
  - **Trụ sở**: Mỹ (ví dụ một thuộc tính thêm có thể có của thực thể ORGANIZATION)
- Thực thể EVENT: **Ra_mắt_Iphone_8**
- Thực thể TIME: **12/9/2017**
- FACT thể hiện quan hệ **tổ_chức** giữa ORGANIZATION và EVENT: **Apple → tổ_chức → Ra_mắt_Iphone_8**
- FACT thể hiện quan hệ **diễn_ra_lúc** giữa EVENT và TIME: **Ra_mắt_Iphone_8 → diễn_ra_lúc → 12/9/2017**

Ví dụ 2: Với đoạn tin "Tổng thống Obama, thủ tướng Nguyễn Xuân Phúc và tập đoàn Apple đã đến thăm Văn Miếu tại Hà Nội ngày 20/7/1917", thông tin trích xuất được có thể là:

- Thực thể PERSON: **Tổng_Thống_Obama**
- Thực thể PERSON: **Thủ_Tướng_Nguyễn_Xuân_Phúc**
  - **Định danh**: Thu_Tuong_Nguyen_Xuan_Phuc
  - **Nhãn hiển thị**: Thủ tướng Nguyễn Xuân Phúc

    - **Mô tả**: Nguyễn Xuân Phúc là thủ tướng thứ 7 của nước Cộng hòa xã hội chủ nghĩa Việt Nam
    - **Chức vụ**: thủ tướng
- Thực thể ORGANIZATION: **Apple**
- Thực thể LOCATION: **Văn_Miếu_Hà_Nội**
- Thực thể TIME: **20/7/1997**
- FACT thể hiện quan hệ **đến_thăm** giữa PERSON và LOCATION:
    - **Tổng_Thống_Obama → đến_thăm → Văn_Miếu_Hà_Nội**
    - **Thủ_Tướng_Nguyễn_Xuân_Phúc → đến_thăm → Văn_Miếu_Hà_Nội**
- FACT thể hiện quan hệ **đến_thăm** giữa ORGANIZATION và LOCATION:
    - **Apple → đến_thăm → Văn_Miếu_Hà_Nội**

Ví dụ 3: Với đoạn tin "Vượt qua 6 đội tại vòng bảng, Mỹ và Italy đã xuất sắc vào vòng chung kết Lễ hội pháo hoa quốc tế 2018 tại Đà Nẵng", thông tin trích rút được có thể là:

- Thực thể COUNTRY: **Mỹ**
- Thực thể COUNTRY: **Italty**
- Thực thể LOCATION: **Đà_Nẵng**
    - **Định danh**: Đà_Nẵng00001
    - **Nhãn hiển thị**: Đà Nẵng
    - **Mô tả**: Đà Nẵng là một thành phố thuộc trung ương, nằm trong vùng Nam Trung Bộ, Việt Nam, là trung tâm kinh tế, tài chính, chính trị, văn hoá, du lịch, xã hội, giáo dục, đào tạo, khoa học và công nghệ, y tế chuyên sâu của khu vực miền Trung - Tây Nguyên và cả nước
    - **Quốc gia**: Việt Nam
- Thực thể EVENT: **Chung_kết_Lễ_hội_pháo_hoa_quốc_tế_2018**
- FACT thể hiện quan hệ **diễn_ra_tại** giữa EVENT và LOCATION
    - **Chung_kết_Lễ_hội_pháo_hoa_quốc_tế_2018 → diễn_ra_tại → Đà_Nẵng**
- FACT thể hiện quan hệ **tham_gia** giữa COUNTRY và EVENT:
    - **Mỹ → tham_gia → Chung_kết_Lễ_hội_pháo_hoa_quốc_tế_2018**
    - **Italy → tham_gia → Chung_kết_Lễ_hội_pháo_hoa_quốc_tế_2018**

Ví dụ 4: Với đoạn tin "Đà Nẵng tổ chức thi trình diễn pháo hoa quốc tế lần thứ 7", thông tin trích xuất được có thể là:

- Thực thể ORGANIZATION: **Đà_Nẵng** (lưu ý, cùng là Đà_Nẵng, nhưng ví dụ 3 là LOCATION, còn ví dụ này là ORGANIZATION)
    - **Định danh**: Đà_Nẵng00002
    - **Nhãn hiển thị**: Thành phố và chính quyền Đà Nẵng
    - **Mô tả**: Hội đồng nhân dân thành phố Đà Nẵng, Ủy ban nhân dân thành phố Đà Nẵng, Ban chấp hành đảng bộ Đà Nẵng, và Ủy ban Mặt trận Tổ quốc Thành phố
    - **Trụ sở**: Việt Nam
- Thực thể EVENT: **Thi_trình_diễn_pháo_hoa_quốc_tế_lần_thứ_7**
- FACT thể hiện quan hệ tổ_chức giữa ORGANIZATION và EVENT
    - **Đà_Nẵng→tổ_chức→Thi_trình_diễn_pháo_hoa_quốc_tế_lần_thứ_7**

Giả sử nhiệm vụ trích rút tự động các thực thể và quan hệ từ các bản tin đã được giải quyết. Sinh viên cần xây dựng hệ thống lưu trữ số lượng lớn các thực thể và quan hệ trích rút được, rồi thực hiện kiểm thử hiệu năng trên hệ thống xây dựng được. Cụ thể các công việc cần thực hiện như sau:

(1) Tìm hiểu các công nghệ lưu trữ và truy vấn dữ liệu theo đề tài được phân công

(2) Thiết kế mô hình lưu trữ dữ liệu tối ưu theo công nghệ được phân công

(3) Tạo module cho phép sinh giả lập lượng lớn dữ liệu về các thực thể và quan hệ theo mô tả ở trên, rồi lưu trữ lại, sử dụng công nghệ tìm hiểu ở (1). Tuy giả lập nhưng cần lập trình sao cho dữ liệu gần với thực tế. Ví dụ: sinh viên xây dựng tập các thực thể, tập các trường thông tin của thực thể, tập các quan hệ, … rồi chọn ngẫu nhiên các kết hợp của các thành phần trong các tập đó để sinh dữ liệu giả lập. Việc tạo tập dữ liệu và chọn ngẫu nhiên phải đảm bảo CSDL có ý nghĩa, để sau đó có thể truy vấn được.

Danh sách các loại quan hệ (tối thiểu phải có, SV có thể mở rộng thêm)

	Thực thể A	**Quan hệ**	**Thực thể B**
1	PER	*gặp gỡ*	PER
2	ORG|PER	*tổ chức*	EVENT
3	CTY	*ký thỏa thuận*	CTY
4	PER|ORG	*tham gia*	ORG|EVENT|ARG
5	EVENT	*diễn ra tại*	LOC|CTY
6	PER|CTY	*ủng hộ*	CTY|ARG|EVENT
7	PER|CTY	*phản đối*	CTY|ARG|EVENT
8	PER	*phát biểu tại*	EVENT
9	CTY	*căng thẳng với*	CTY
10	PER|CTY	*hủy bỏ*	ARG|EVENT
11	CTY	*đám phán với*	CTY

(4) Xây dựng tối thiểu 10 truy vấn **cơ bản** (dạng truy vấn đơn giản) trên cơ sở dữ liệu giả lập. Ví dụ:

- Lấy thông tin mô tả của Đà_Nẵng00001
- Chung_kết_Lễ_hội_pháo_hoa_quốc_tế_2018 diễn ra tại đâu

(5) Xây dựng tối thiểu 10 truy vấn mang tính chất thống kê (dạng truy vấn phức tạp hơn) trên cơ sở dữ liệu giả lập. Ví dụ:

- Những EVENT nào diễn_ra_tại Đà_Nẵng vào tháng 2/2018
- Thủ_tướng_Nguyễn_Xuân_Phúc đến_thăm LOCATION nào năm 2017

(6) Làm báo cáo

- Giới thiệu về công nghệ tìm hiểu, hướng dẫn chi tiết các bước cài đặt, cách sử dụng API tương tác với cơ sở dữ liệu
- Vẽ thiết kế biểu đồ lớp của chương trình/hệ thống xây dựng được
- Xây dựng chương trình kiểm thử tự động, xác định thông số hiệu năng của hệ thống. Sinh viên code và tạo giả lập ra N thực thể, M quan hệ và thử nghiệm với lần lượt 10 câu truy vấn cơ bản và 10 câu truy vấn thống kê, rồi lập bảng kết quả thời gian phản hồi (Lưu ý: giá trị N, M tùy sinh viên, tăng đến mức tối đa có thể. Giá trị càng lớn, càng được đánh giá cao). Khi sinh dữ liệu, cần dùng kỹ thuật batch processing (xử lý theo lô), sinh một lượng xác định dữ liệu rồi add vào model và submit, thay vì sinh được một thành phần dữ liệu và submit ngay.

(N, M)	Thời gian sinh DL	TVCB1	TVCB2	…	TVCB9	TVCB10
100, 200		1ms	0.7ms		0.8ms	0.7ms
5000, 7000						
60k, 80k						
1M, 2M						
15M, 17M						
...						
...						

Các công nghệ lưu trữ và truy vấn dữ liệu:

Đề tài 1: OrientDB
- Server: https://orientdb.com/
- Java API: http://www.orientdb.com/docs/last/Java-API.html

Đề tài 2: ArangoDB
- Server: https://www.arangodb.com/
- Java API: https://www.arangodb.com/tutorials/

Đề tài 3: GraphDB
- Server: http://graphdb.ontotext.com/
- Java API: http://graphdb.ontotext.com/free/devhub/programming.html

Đề tài 4: Neo4J
- Server: https://neo4j.com/
- Java API: https://neo4j.com/developer/java/
- Doc: https://neo4j.com/docs/developer-manual/current/introduction/graphdb-concepts/

Đề tài 5: MySQL