Text Mining and Social Media Mining

THS. LÊ NHẬT TÙNG

Biểu diễn Tri thức
Knowledge Representation

THS. LÊ NHẬT TÙNG

Biểu diễn Tri thức (Knowledge Representation)

• Là một **lĩnh vực con của Trí tuệ Nhân tạo (AI)** nghiên cứu cách biểu diễn kiến thức về thế giới và cách thực hiện suy luận từ kiến thức đó

• Cần thiết để máy tính có thể "hiểu" và xử lý thông tin giống như con người

• Bao gồm nhiều phương pháp như: mạng ngữ nghĩa, khung, quy tắc và bản thể học (ontologies)

Đặc điểm của Biểu diễn Tri thức

- Tất cả các biểu diễn đều không hoàn hảo:
  - Chỉ là hình ảnh xấp xỉ của thực tế
  - Mỗi cách biểu diễn chú trọng một số khía cạnh và bỏ qua các khía cạnh khác
- **Ví dụ:** Khi biểu diễn tri thức về "ô tô", ta có thể chỉ tập trung vào đặc điểm kỹ thuật (động cơ, tốc độ) mà bỏ qua khía cạnh thẩm mỹ (thiết kế, màu sắc).

Trích xuất Thông tin (Information Extraction)

Trích xuất Thông tin là gì?

- Là **quá trình tự động lấy thông tin có cấu trúc** từ dữ liệu phi cấu trúc (như văn bản)
- Là một trong những nhiệm vụ quan trọng nhất trong khai thác văn bản
- Được ứng dụng rộng rãi trong y học, kinh doanh, phân tích mạng xã hội...

Ví dụ về Trích xuất Thông tin

Xem xét câu sau:

"Công ty VinFast được thành lập bởi Tập đoàn Vingroup vào năm 2017 tại Hải Phòng, Việt Nam."

Thông tin trích xuất được:

- Tên công ty: VinFast
- Người sáng lập: Tập đoàn Vingroup
- Năm thành lập: 2017
- Địa điểm: Hải Phòng, Việt Nam
  - → Máy tính có thể hiểu: ThànhLập(Vingroup, VinFast, 2017, "Hải Phòng, Việt Nam")

Nhận dạng Thực thể Có Tên (Named Entity Recognition - NER)

Nhận dạng Thực thể Có Tên là gì?

- Là quá trình **xác định và phân loại các thực thể** trong văn bản như:
  - Tên người
  - Tên tổ chức/công ty
  - Địa điểm
  - Thời gian
  - Số tiền
  - ...
- Là bước quan trọng trong xử lý ngôn ngữ tự nhiên
- Giúp máy tính "hiểu" văn bản tốt hơn

Ví dụ về Nhận dạng Thực thể

Câu gốc: "Chủ tịch Phạm Nhật Vượng của Vingroup đã đầu tư 4 tỷ USD vào nhà máy ở Chicago vào tháng 3 năm 2023."

Kết quả NER:

- [Phạm Nhật Vượng] → NGƯỜI
- [Vingroup] → TỔ CHỨC
- [4 tỷ USD] → TIỀN TỆ
- [Chicago] → ĐỊA ĐIỂM
- [tháng 3 năm 2023] → THỜI GIAN

Nhận dạng Thực thể Có Tên (Named Entity Recognition - NER)

Thách thức trong NER

Các thực thể phụ thuộc vào ngữ cảnh

Ví dụ:

- "Apple đang tuyển nhân viên mới" → Apple là CÔNG TY
- "Tôi vừa mua một quả apple" → apple là TRÁI CÂY
- "Big Apple thu hút nhiều du khách" → Big Apple là biệt danh của NEW YORK

Trích xuất Quan hệ (Relation Extraction)

Trích xuất Quan hệ là gì?

- Là quá trình **tìm và xác định mối quan hệ** giữa các thực thể trong văn bản
- Thường được thực hiện sau bước Nhận dạng Thực thể (NER)
- Mục tiêu: xác định loại quan hệ cụ thể giữa hai hoặc nhiều thực thể

Ví dụ về Trích xuất Quan hệ

Câu gốc: “Nguyễn Văn A là Chủ tịch nước ABC từ năm 2021 đến năm 2023."

Trích xuất quan hệ:

- GiữChứcVụ(Nguyễn Văn A, Chủ tịch nước, ABC)
- ThờiGianTạiVị(Nguyễn Văn A, 2021, 2023)

Các phương pháp Trích xuất Quan hệ

- **Phương pháp đồng xuất hiện (Co-occurrence)**
  - Dựa trên tần suất xuất hiện cùng nhau của các thực thể
  - Kết quả: Độ phủ (Recall) cao, Độ chính xác (Precision) thấp
- **Phương pháp dựa trên quy tắc (Rule-based)**
  - Sử dụng các quy tắc ngôn ngữ được định nghĩa trước
  - Ví dụ: "X được thành lập bởi Y" → ThànhLập(Y, X)
- **Phương pháp học máy (Machine Learning)**
  - Sử dụng các kỹ thuật phân loại để nhận diện quan hệ
  - Cần dữ liệu huấn luyện đã được gán nhãn

Trực quan hóa trong Khai thác Văn bản

Đám mây từ (Word Cloud)

- Là cách **hiển thị trực quan các từ quan trọng** trong văn bản
- Kích thước của từ thể hiện tần suất xuất hiện hoặc mức độ quan trọng
- Ưu điểm:
  - Đơn giản, dễ hiểu
  - Cung cấp cái nhìn tổng quan nhanh chóng
  - Công cụ truyền thông hiệu quả


TEAMWORK
INNOVATION
RESEARCH
COLLABORATION
COOPERATION
DEFINITION
COMMUNITY
VISION
PROFESSIONAL
TOGETHER
SOLVING
CRISIS
WORK
CAREER
UNION
SUCCESS
SUPPORT
HELP
VISION
SKILLED
TEAMWORK
INNOVATION
DETERMINATION
HAND
TRUST
DIVERSITY
MAN
GROUP
SKILLS
INVESTIGATION
BUSINESS
DOCUMENT
UNITE
PUZZLE
PLANNING
SUCCESSFUL
UNITY
LIFT
TEAM
PROJECT
ARROW
SOLUTION

Các loại trực quan hóa khác

- **Biểu đồ cột/thanh**: So sánh dữ liệu theo nhóm
- **Biểu đồ đường**: Thể hiện xu hướng theo thời gian
- **Biểu đồ mạng lưới (Network Visualization)**: Hiển thị mối quan hệ giữa các thực thể
- **Công cụ phân tích tình cảm (Sentiment Analysis)**: Đánh giá sắc thái tình cảm trong văn bản