

Khai thác văn bản và ứng dụng 2023
CH K33

Tổng quan

Nguyễn Trường Sơn – Nguyễn Tiên Huy
ntson@fit.hcmus.edu.vn – ntienhuy@fit.hcmus.edu.vn



KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Nội dung

- Giới thiệu khai thác văn bản
- Ứng dụng của khai thác văn bản
- Kiến trúc chung của một hệ thống khai thác văn bản
- Các bài toán chính trong khai thác văn bản
- Cách tiếp cận cho các bài toán khai thác văn bản
- Khai thác văn bản và các bài toán xử lý ngôn ngữ tự nhiên.
- Phạm vi môn học
- Công cụ / môi trường / ngôn ngữ lập trình
- Thi & Đánh giá
- Bài tập mở đầu
- ..

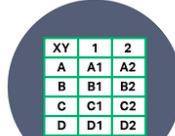
Giới thiệu

- **Thời kỳ của dữ liệu lớn:** Lượng thông tin khổng lồ sinh ra mỗi ngày và tiếp tục tăng lên → quá tải thông tin: “information overload”
 - Dữ liệu đa phương tiện (multimedia): hình ảnh, âm thanh, video, văn bản, ...
 - Vd: Dữ liệu các giao dịch thương mại điện tử, ...
- **Phân loại:**
 - Dữ liệu có cấu trúc: các thông tin đã được chọn lọc và lưu trữ dạng cấu trúc:
 - Sẵn sàng để cho để khai thác
 - Sử dụng các thuật toán khai thác dữ liệu (data mining)
 - Dữ liệu không cấu trúc: thông tin trình bày dạng tự nhiên, không theo chuẩn mực, định dạng không nhất quán
 - Chưa sẵn sàng để khai thác

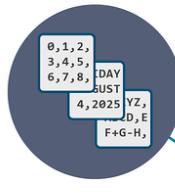
Dữ liệu có cấu trúc và không cấu trúc

Structured Data vs Unstructured Data

Can be displayed in rows, columns and relational databases



Numbers, dates and strings



Estimated 20% of enterprise data (Gartner)



Requires less storage



vs

Unstructured Data

Cannot be displayed in rows, columns and relational databases



Images, audio, video, word processing files, e-mails, spreadsheets



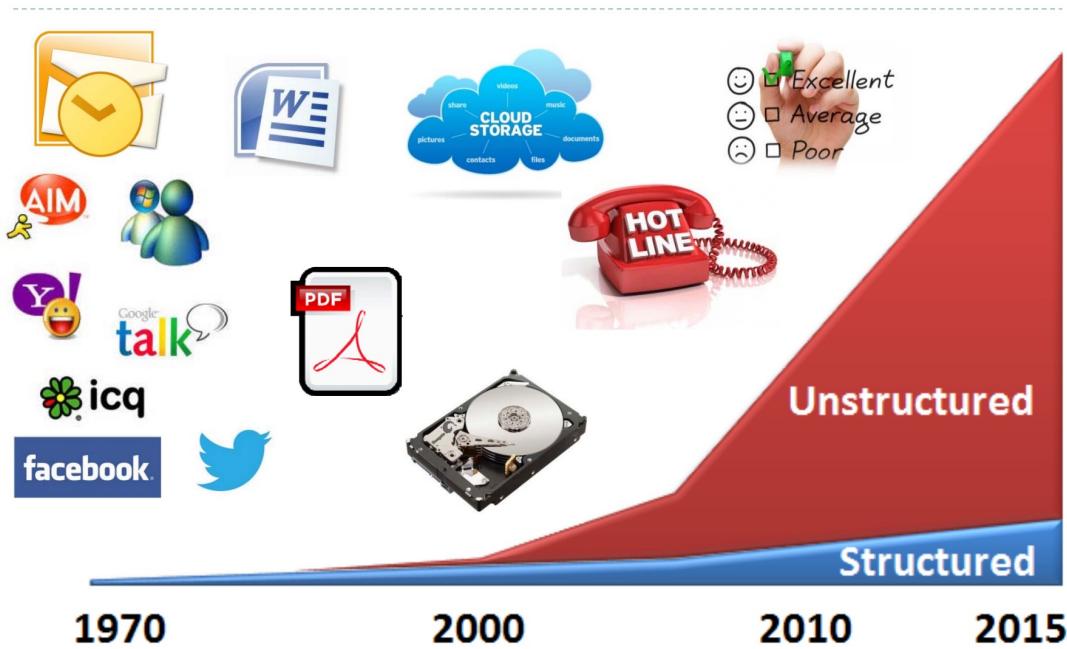
Estimated 80% of enterprise data (Gartner)



Requires more storage



Dữ liệu có cấu trúc và không cấu trúc



Dữ liệu văn bản

□ Dữ liệu văn bản:

- Là hình thức trao đổi phổ biến nhất trong giao tiếp giữa con người và con người.
- Dữ liệu văn bản có mặt ở mọi lúc mọi nơi và phát triển rất nhanh: Internet, blogs, news, email, literature, twitter, ...
 - Vd thống kê: khoảng **6,000 tweets/giây, 350,000 tweets/phút, 500 triệu tweets / ngày, 200 tỉ tweets /năm**
- Đặc điểm: Không có cấu trúc trình bày cụ thể

Làm thế nào để khai thác các thông tin hữu ích trong lượng dữ liệu khổng lồ đó ?
- Ví dụ: “Sản phẩm X mới ra mắt có thành công hay không”



Khai thác văn bản



Khai thác văn bản là gì ?

- Khai thác văn bản = Text-mining / Text-analytics
- Là quá trình phát hiện những thông tin tri thức mới một cách tự động bằng việc rút trích các thông tin từ các nguồn dữ liệu văn bản khác nhau
- Phát hiện ra những mối quan hệ từ những thông tin rút trích trong văn bản từ đó phát hiện ra những tri thức mới (new facts or new hypotheses)
- Mục tiêu cao nhất của text-mining: phát hiện những tri thức ẩn trong văn bản



Khai thác văn bản không phải là khai thác dữ liệu

- Khai thác dữ liệu (Data mining)
 - = Phát hiện tri thức từ dữ liệu có cấu trúc: số, chuỗi, ngày tháng, ...
 - Ví dụ:
 - Dựa trên dữ liệu mua hàng của khách hàng để xác định sản phẩm nào nên bán tiếp theo
 - Dựa trên thông tin các giao dịch trên thẻ để phát hiện các giao dịch gian lận
- Khai thác văn bản:
 - = Phát hiện tri thức từ dữ liệu văn bản có không có trúc
 - Ví dụ:
 - Dựa trên phân tích các bình luận của khách hàng về sản phẩm từ đó đưa ra quyết định có nên tiếp tục / ngừng kinh doanh, ...



Ứng dụng của khai thác văn bản

Các lĩnh vực ứng dụng:

- Sản xuất hàng hoá
- Thương mại điện tử
- Chính phủ
- Viễn thông
- Tài chính
- Bảo hiểm
- Sức khoẻ
- Pháp luật
- Các lĩnh vực khoa học
- ...

Text Analytics Use Cases

| | | |
|--|--|---|
| Manufacturers | Government | Financial Institutions |
| <ul style="list-style-type: none"> • Identify root causes of product issues quicker • Identify trends in market segments • Understand competitors' products | <ul style="list-style-type: none"> • Identify fraud • Understand public sentiments about unmet needs • Find emerging concerns that can shape policy | <ul style="list-style-type: none"> • Use contact center transcriptions understand customers • Identify money laundering or other fraudulent situations |
| Retail | Legal | Healthcare |
| <ul style="list-style-type: none"> • Identify profitable customers and understand the reasons for their loyalty • Manage the brand on social media | <ul style="list-style-type: none"> • Identify topics and keywords in discovery documents • Find patterns in defendant's communications | <ul style="list-style-type: none"> • Find similar patterns in doctor's reports • Use social media to detect disease outbreaks earlier • Identify patterns in patient claims data |
| Telecommunications | Life Sciences | Insurance |
| <ul style="list-style-type: none"> • Prevent customer churn • Suggest up-sell/cross-sell opportunities by understanding customer comments | <ul style="list-style-type: none"> • Identify adverse events in medicines or vaccines • Recommend appropriate research materials | <ul style="list-style-type: none"> • Identify fraudulent claims • Track competitive intelligence • Manage the brand on social media |

<https://www.zencos.com/blog/text-mining-examples-advanced-analytics/>



Một số ứng dụng của khai thác văn bản

Ứng dụng trong mọi lĩnh vực đời sống xã hội, xây dựng các ứng dụng thông minh:

- Phân tích cảm xúc người dùng (Sentiment Analysis)
- Phát hiện thư rác (Spam filtering)
- Phát hiện tin giả / tin vịt (Fake news detections)
- Tóm tắt tài liệu (Document summarization)
- Dịch vụ chăm sóc khách hàng – trả lời tự động(Customer care service)
- Phân tích dữ liệu trên mạng xã hội (Social Media Analysis)
- Phân tích hồ sơ người dùng (Resume, CV)
- Phát hiện gian lận trong các giao dịch (Fraud Detection)
- Phát hiện thông tin dịch bệnh và cảnh báo sớm
- Xây dựng các hệ hội thoại người máy (chatbot) , hỏi đáp
- Phát hiện đạo văn, ...

Một số ứng dụng của khai thác văn bản

Ví dụ 1: Phân tích cảm xúc người dùng

- | | | |
|---|---|--|
| <ul style="list-style-type: none">- The battery life of this camera is too short- your product is so bad- your customer support is killing me- This product is amazing |  | <p>Very positive</p> <p>Positive</p> <p>Neutral</p> <p>Negative</p> <p>Very negative</p> |
|---|---|--|

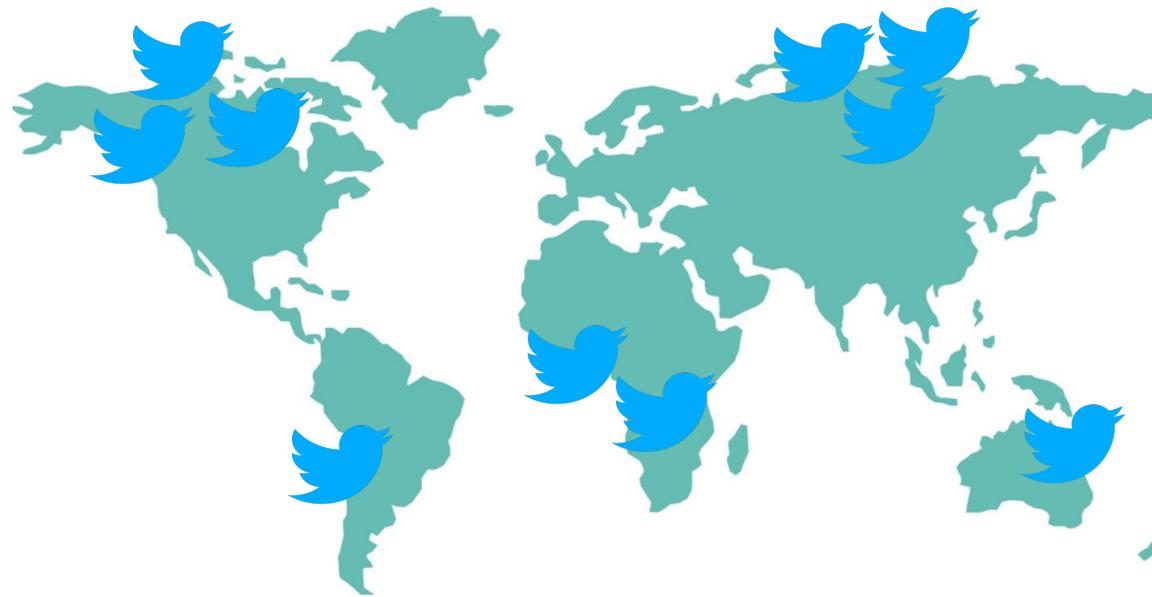
Phân tích các nhận xét của người dùng đánh giá về chất lượng của sản phẩm, dịch vụ từ đó cải tiến quy trình quản lý cho hiệu quả hơn hoặc đưa ra chiến lược kinh doanh phù hợp.

Keywords: sentiment analysis, opinion mining



Một số ứng dụng của khai thác văn bản

Ví dụ 2: Ứng dụng của khai thác văn bản trong phát hiện sớm các dịch bệnh



Phân tích các tin nhắn trên các mạng xã hội, cùng với toạ độ của người dùng từ đó phát hiện sớm các dấu hiệu bất thường và cảnh báo
Keywords: early warning outbreak, outbreak detection, ...

Một số ứng dụng của khai thác văn bản – COVID 19

Google image processing for covid 19

[https://www.frontiersin.org › articles › full › Dịch trang này](https://www.frontiersin.org/articles/full/Dich trang này)
AI-Based Image Processing for COVID-19 Detection in Chest ...
 9 thg 8, 2021 — The basic phase in **image processing** and interpretation to detect and assess **COVID-19** is segmentation. It defines the key factor for the AI ...
[Related Work](#) · [The COVID-19 Detection System](#) · [Model Simulation and Results](#)

[https://www.sciencedirect.com › pii › Dịch trang này](https://www.sciencedirect.com/pii/Dich trang này)
Deep learning and medical image processing for coronavirus
 viết bởi S Bhattacharya · 2021 · Trích dẫn 88 bài viết (e.g., X-ray, CT, and MRI) make deep learning a gre

[https://www.hindawi.com › journals › jhe › Dịch trang này](https://www.hindawi.com/journals/jhe/Dich trang này)
Computed Tomography Image Processing for COVID-19
 viết bởi S Tello-Mijares · Trích dẫn 1 bài viết — The pandemic has infected patients around the world in ...
[Introduction](#) · [Materials and Methods](#) · [Experimental](#)

Google sound processing for covid 19

[https://www.researchgate.net › publication › Dịch trang này](https://www.researchgate.net/publication/Dich trang này)
Audio, Speech, Language, & Signal Processing for COVID-19
 17 thg 7, 2021 — PDF | The **Coronavirus (COVID-19)** pandemic has been the research focus world-wide in the year 2020. Several efforts, from collection of ...

[https://signalprocessingsociety.org › app... › Dịch trang này](https://signalprocessingsociety.org/app.../Dich trang này)
This App Could Help Detect COVID-19 By Analyzing A ...
 Home » This App Could Help Detect COVID-19 By Analyzing A Person's Speech. Top Reasons to Join SPS Today! 1. IEEE Signal Processing Magazine

[.../Dich trang này](#)
processing for COVID-19 - PubMed
 ed human **audio processing for COVID-19**: A comprehensive ...
 2 Feb;122:108289. doi: 10.1016/j.patog.2021.108289.

[articles › Dịch trang này](#)
19 from speech signal using bio-inspired ...
 h dẫn 3 bài viết — The cepstral analysis is one the oldest and ...
 ds used in various applications like speech signal...

- Tìm hiểu các bài toán khai thác văn bản hỗ trợ cho COVID

Một số ứng dụng của khai thác văn bản – Social Listening

- Ứng dụng:
 - Đo lường Sức khoẻ thương hiệu
 - Xử lý khủng hoảng Truyền thông
 - Thăm dò đối thủ cạnh tranh
 - Theo dõi hoạt động truyền thông
 - Hỗ trợ Chăm sóc khách hàng



- Phương pháp thực hiện:
 - Thu thập dữ liệu
 - Phân loại dữ liệu
 - Phân tích dữ liệu
 - Xuất dữ liệu
 - Trình bày báo cáo nghiên cứu

Một kết quả của “Social Listening”

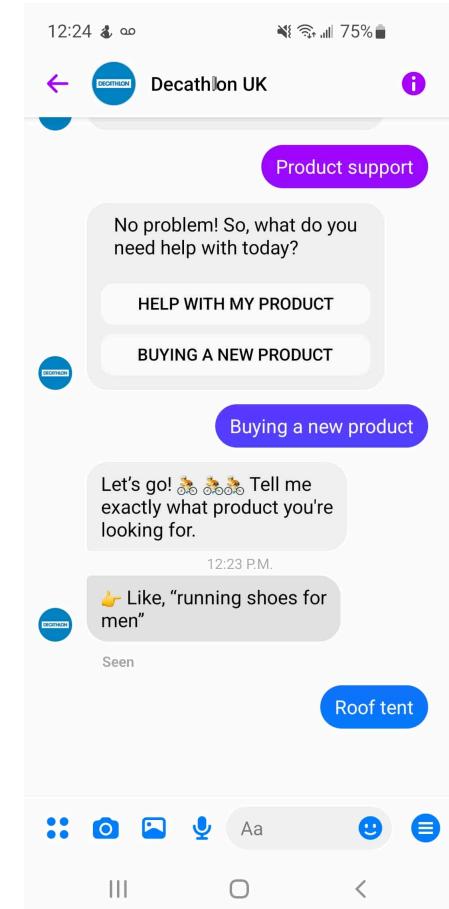
Bảng xếp hạng Doanh nghiệp ngành Thương mại điện tử trong tháng 06/2023

| | | | | | |
|----|------------------|---------------|------------------|----|------------------|
| 1 | Shopee VN | Shopee | 129.06 -13.33 | 11 | Mediamart |
| 2 | Lazada VN | Lazada | 50.6 +14.23 | 12 | CellphoneS |
| 3 | Thế Giới Di Động | thegioididong | 32.54 +7.94 | 13 | Viettel Store |
| 4 | Tiki | TIKI | 22.35 -1.24 | 14 | Điện Máy Chợ Lớn |
| 5 | FPT Shop | FPT Shop | 22.3 +4.8 | 15 | Di Động Việt |
| 6 | Điện Máy Xanh | Điện máy XANH | 22.02 -0.14 | | |
| 7 | Nguyễn Kim | NGUYENKIM | 20.56 +13.2 | | |
| 8 | Meta | META.vn | 10.46 -4.93 | | |
| 9 | Sendo | Sendo | 10.39 -12.89 | | |
| 10 | Bách Hóa XANH | Bách hóa XANH | 8.03 -0.21 | | |

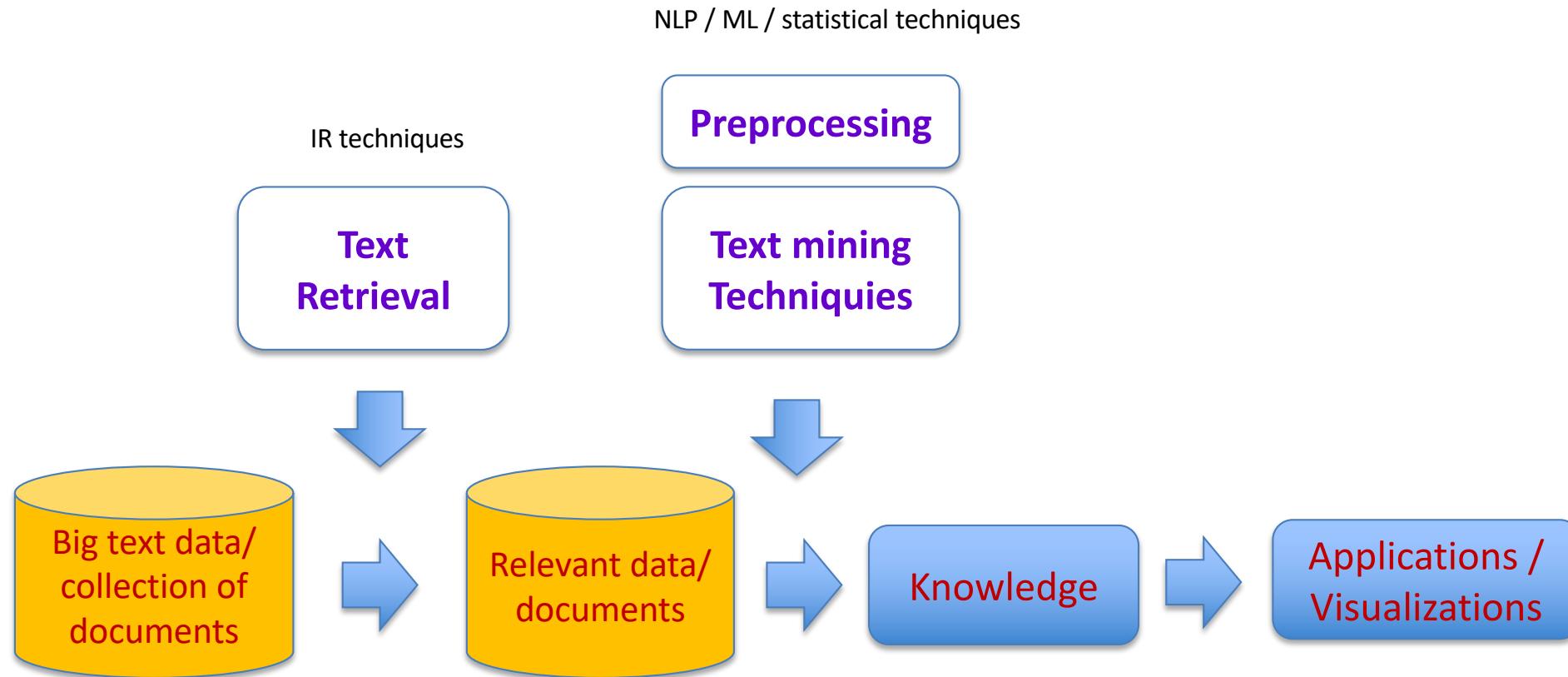
Thời gian thu thập dữ liệu: từ 00h00 ngày 01/06 đến 23h59 ngày 30/06
 Tổng điểm = Điểm xác thực⁽¹⁾ + Điểm thị phần thảo luận⁽²⁾ + Điểm tương tác⁽³⁾ + Điểm độ lan tỏa⁽⁴⁾ + Điểm lưu lượng truy cập Web⁽⁵⁾

Một số ứng dụng của khai thác văn bản – Chatbot & QA

- Hệ thống tự động trả lời câu hỏi / chăm sóc khách hàng
- Lợi ích:
 - 1. Nâng cao trải nghiệm khách hàng
 - 2. Hỗ trợ mở rộng quy mô doanh nghiệp
 - 3. Phân phối nội dung
 - 4. Cải thiện khả năng nhận diện thương hiệu
 - 5. Hỗ trợ khách hàng đặt hàng
 - 6. Tự động hóa chốt đơn hàng
 - 7. Nghiên cứu thị trường
 - 8. Lợi ích của chatbot trong hoạt động tuyển dụng
 - 9. Lợi ích của chatbot trong việc giảm thiểu chi phí
 - 10. Lợi ích của chatbot trong việc tăng doanh thu
- Phương pháp thực hiện:
 - Hiểu câu hỏi người dùng
 - Tìm kiếm các thông tin từ kho dữ liệu
 - Đưa ra trả lời cho người dùng



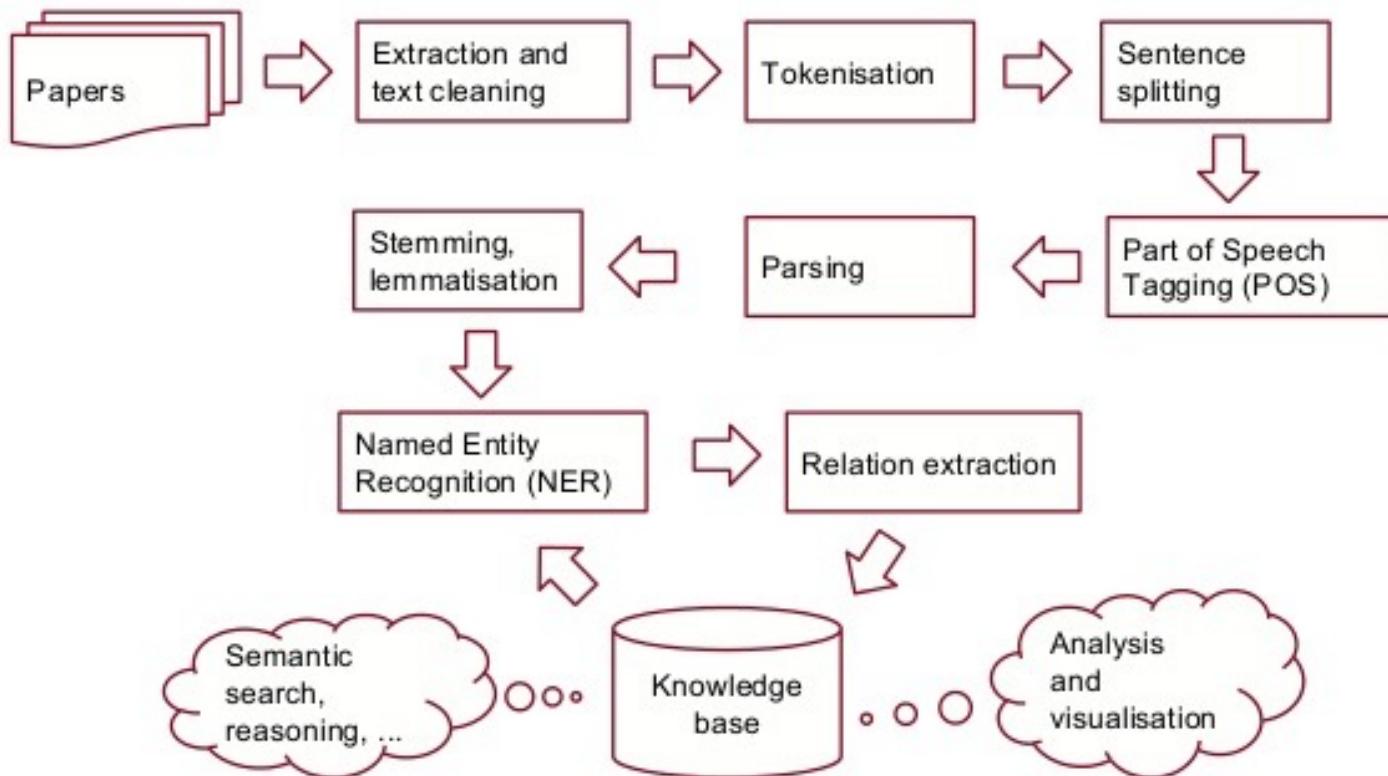
Kiến trúc chung của một hệ thống khai thác văn bản



Kiến trúc chung của một hệ thống khai thác văn bản



Generic text mining workflow



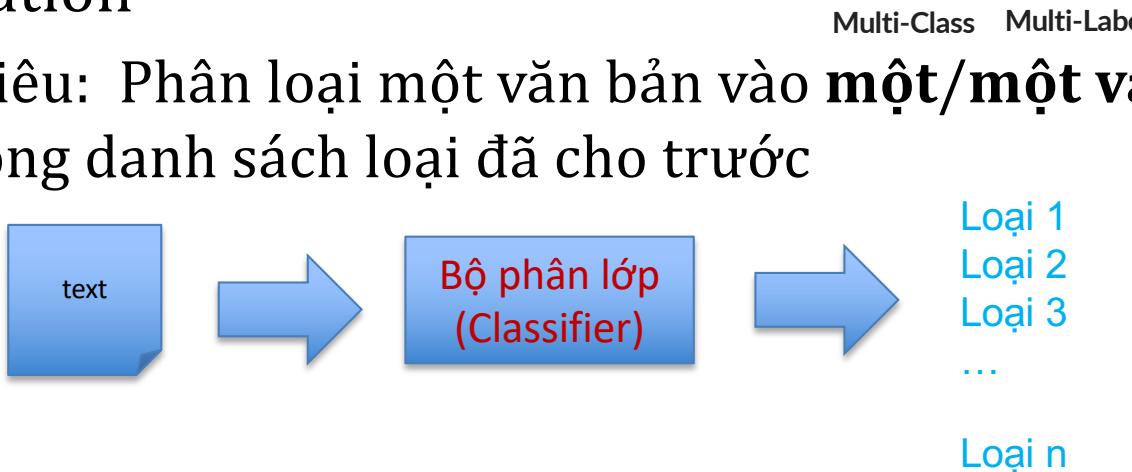
Phân loại các bài toán của hệ thống khai thác văn bản

- Một số bài toán chính:
 - Phân lớp văn bản
 - Rút trích thông tin (Information Extraction)
 - Rút trích thực thể có tên (Named entity extraction)
 - Rút trích quan hệ (Relation Extraction)
 - Phân nhóm tài liệu (Document clustering)
 - Tìm kiếm thông tin (Information Retrieval)
 - ...
- Hầu hết các ứng dụng khai thác văn bản có thể quy về những bài toán này hoặc là sự kết hợp của những bài toán này.
 - Ví dụ: Phân loại đánh giá người dùng về sản phẩm → Phân lớp văn bản



Phân lớp văn bản

- Phân lớp văn bản = text categorization / document classification
 - Mục tiêu: Phân loại một văn bản vào **một/một vài loại** nào đó trong danh sách loại đã cho trước



- Một số ứng dụng phổ biến:
 - Phân loại nội dung email thành 1 trong 2 loại thư rác / không
 - Phân loại tin tức vào trong các chủ đề tin tức: tin thể thao, tin thời sự...
 - Phân loại ý định người dùng trong các hệ hội thoại (chatbot)
 - “Tôi muốn đặt 2 phòng đơn” → book_a_room
 - “Tôi muốn huỷ đặt phòng”. → cancel_booking
 - Phân loại cảm xúc: tích cực /tiêu cực về sản phẩm, dịch vụ
 - Phân loại bệnh án điện tử:
 - E.g. phân loại thù bệnh án trong quá trình tiếp nhận bệnh

Phân loại cặp văn bản

□ Phân lớp cặp văn bản:

- Mục tiêu: Phân loại một cặp văn bản vào **một/một vài loại** nào đó trong danh sách loại đã cho trước, tìm quan hệ giữa 2 văn bản



□ Một số ứng dụng phổ biến:

- RTE (Recognize Textual Entailment),
NLI (Natural Language Inference): Xác định suy diễn văn bản
- Paraphrase detection: Phát hiện văn bản đồng nghĩa
- QA detection: Phát hiện đoạn văn có trả lời cho câu hỏi không



Rút trích thông tin

- Rút trích thông tin = Information Extraction
 - Xác định/rút trích các thành phần quan tâm trong văn bản



- Các bài toán đặc trưng:
 - Rút trích thực thể (tên người, địa điểm, tên tổ chức, tên sản phẩm, tên thương hiệu, ...): Named entity recognition
 - Rút trích các quan hệ (các sự kiện): Tìm các thực thể và quan hệ giữa các thực thể
- Ứng dụng:
 - Web mining: trending analysis (phân tích xu hướng), social listening:

Rút trích thông tin

- Rút trích các thực thể tên riêng:

Stanford Named Entity Tagger

Classifier: english.muc.7class.distsim.crf.ser.gz

Output Format: highlighted

Preserve Spacing: yes

Please enter your text here:

Partial invoice (€100,000, so roughly 40%) for the consignment C27655 we shipped on 15th August to London from the Make Believe Town depot. INV2345 is for the balance.. Customer contact (Sigourney) says they will pay this on the usual credit terms (30 days).

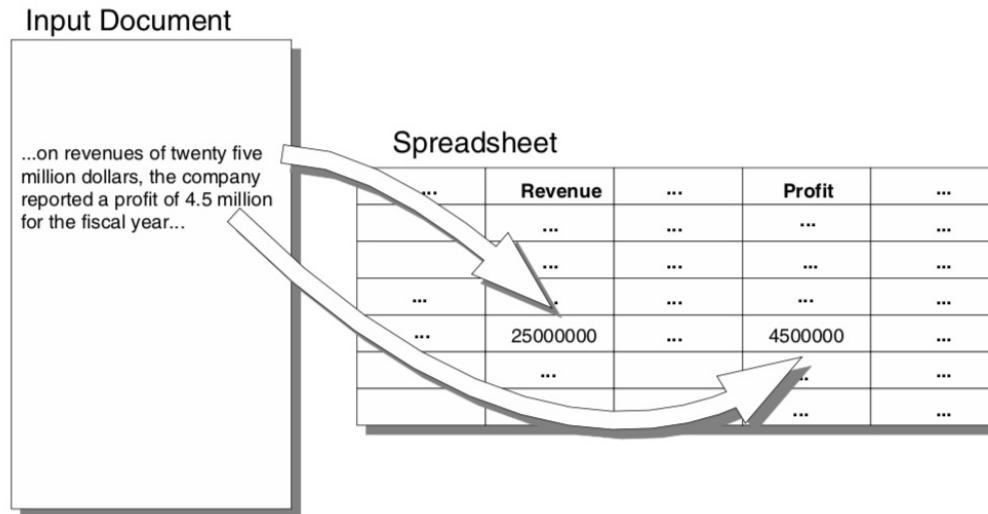
Partial invoice (€100,000, so roughly 40%) for the consignment C27655 we shipped on 15th August to London from the Make Believe Town depot. INV2345 is for the balance.. Customer contact (Sigourney) says they will pay this on the usual credit terms (30 days).

Potential tags:

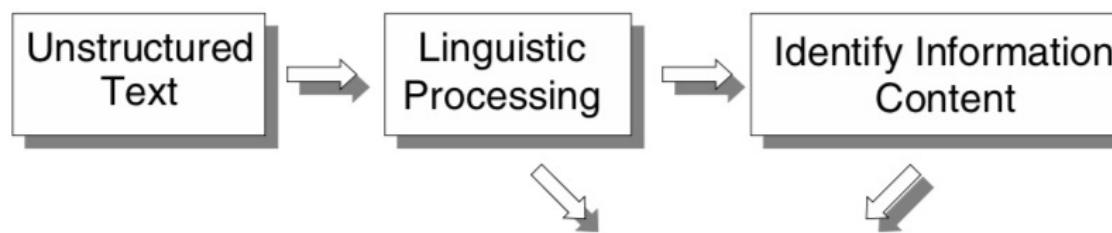
LOCATION
TIME
PERSON
ORGANIZATION
MONEY
PERCENT
DATE

Rút trích thông tin

Rút trích thông tin = Chuyển từ dạng tài liệu không có cấu trúc



Các bước cho bài toán rút trích thông tin



Fill Templates,
Create Structured Knowledge Base

Report Generation

Question Answering

Analysis, Inferencing

Rút trích thông tin

Jack Welch will retire as CEO of General Electric tomorrow. The top role at the Connecticut company will be filled by Jeffrey Immelt.

Single entity

Person: Jack Welch

Person: Jeffrey Immelt

Location: Connecticut

Binary relationship

Relation: Person-Title

Person: Jack Welch

Title: CEO

Relation: Company-Location
Company: General Electric
Location: Connecticut

N-ary record

Relation: Succession

Company: General Electric

Title: CEO

Out: Jack Welsh

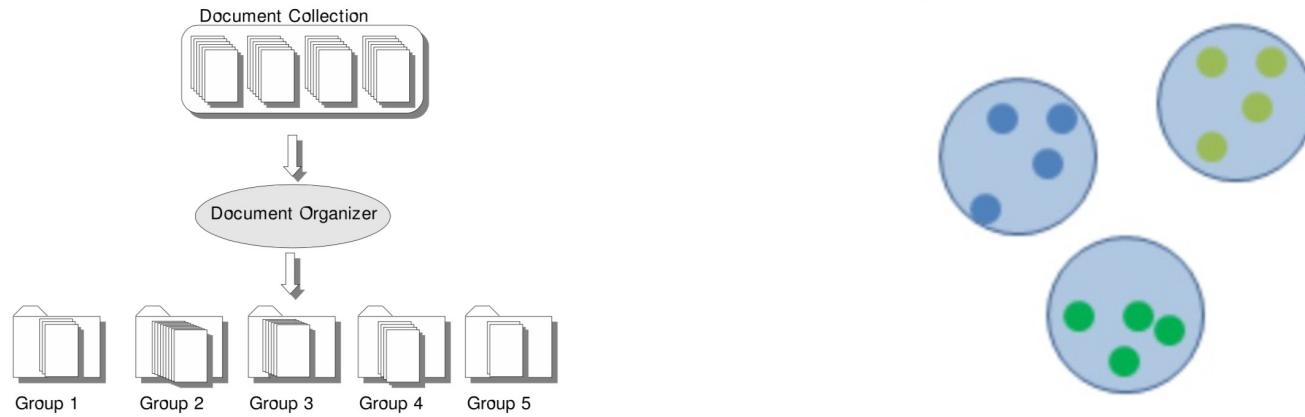
In: Jeffrey Immelt

“Named entity” extraction

Relation extraction

Phân nhóm văn bản

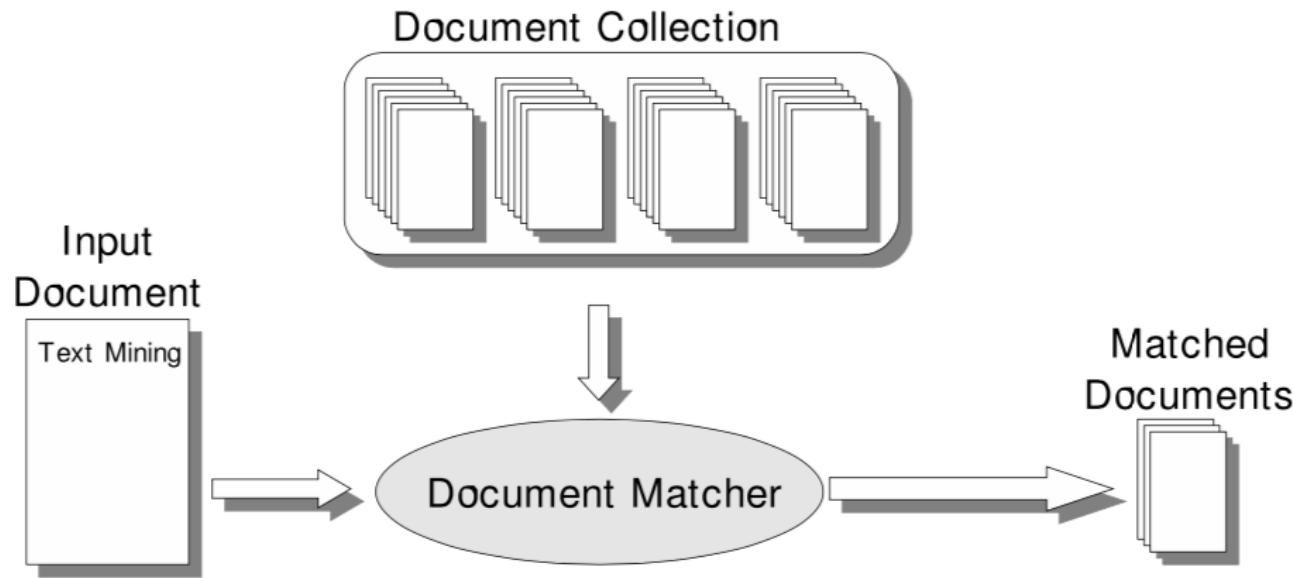
- Phân nhóm văn bản: text clustering / document clustering
- Tự động chia một tập hợp các văn bản thành các nhóm



- Phân loại văn bản và phân nhóm văn bản:
 - Phân loại văn bản: phân một văn bản vào một trong những nhóm xác định trước (có ý nghĩa xác định trước)
 - Phân nhóm văn bản: tự động chia nhóm các văn bản sao cho các văn bản trong 1 nhóm thì có ý nghĩa “tương đồng” nhau

Tìm kiếm thông tin

- Tìm kiếm thông tin = Information Retrieval: Chọn lọc những thông tin liên quan trước khi thực hiện các kỹ thuật mining.



- Phương pháp cơ bản: So sánh sự tương đồng giữa các tài liệu với câu truy vấn
 - Cần mô hình lưu trữ tài liệu và mô hình tính tương đồng

Tìm kiếm thông tin

- Tìm kiếm thông tin = Search engine



[en.wikipedia.org › wiki › Text_mining](#) ▾ [Dịch trang này](#)

[Text mining - Wikipedia](#)

Text mining, also referred to as **text data mining**, similar to **text analytics**, is the process of ... As **text mining** is transformative, **meaning** that it does not supplant the original work, it is viewed as being lawful under fair use. For example, as part of ...

[Category:Text mining](#) · [Biomedical text mining](#) · [List of text mining software](#)

[www.linguamatics.com › what-text-mini...](#) ▾ [Dịch trang này](#)

[What is Text Mining, Text Analytics and Natural Language ...](#)

Text mining (also referred to as **text analytics**) is an artificial intelligence (AI) technology that uses natural language processing (NLP) to transform the free (unstructured) **text** in documents and databases into normalized, structured data suitable for analysis or to drive machine learning (ML) algorithms.

Bạn đã truy cập trang này vào ngày 01/10/2020.

[monkeylearn.com › text-mining](#) ▾ [Dịch trang này](#)

[Text Mining: The Beginner's Guide - MonkeyLearn](#)

What is Text Mining? Text mining, also known as text analysis, is the process of transforming unstructured text data into meaningful and actionable information.



Cách tiếp cận cho các bài toán khai thác văn bản

- **Phương pháp sử dụng luật / heuristics** (Rule-based methods): Sử dụng một tập hợp các luật để rút trích hoặc phân loại văn bản
 - If-Else
 - Regular expression,
 - Heuristics
 - Rút trích tên riêng:
 - Tên người thường đi sau các từ Dr.
 - Tên địa điểm thường đi sau các giới từ “at” và viết hoa
 - Ưu điểm: Đơn giản
 - Hạn chế:
 - Tốn công xây dựng / khó xây dựng bộ luật
- Thường được dùng với những bài toán đơn giản:
- Rút trích số điện thoại, ngày tháng, email

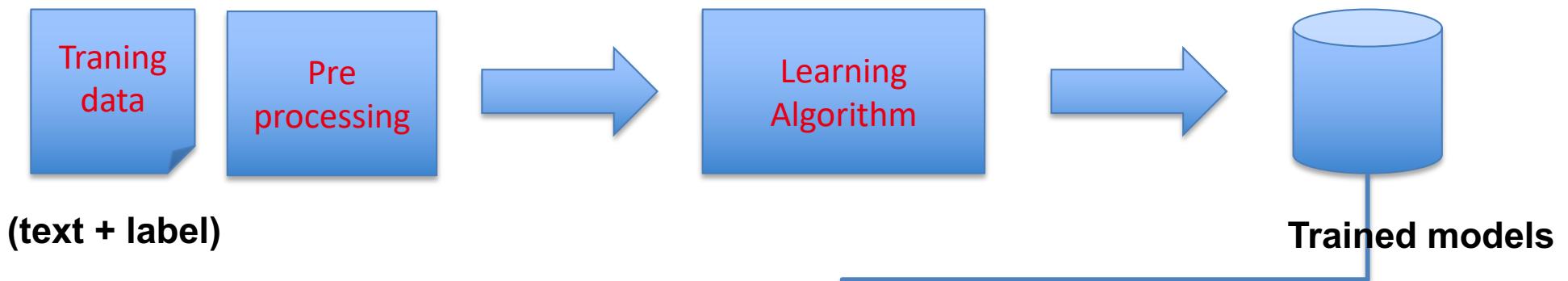
Phân loại kỹ thuật của text mining

- **Phương pháp học máy**: Hệ thống tự động học cách giải quyết các bài toán
 - Học máy có giám sát (Supervised learning)
 - Thuật toán học máy sẽ được học dựa trên bộ dữ liệu gán nhãn của bài toán.
 - Thuật toán học máy sẽ học một bộ tham số dựa trên việc quan sát dữ liệu huấn luyện
 - Học máy không giám sát (Unsupervised learning)
 - Không cần dữ liệu gán nhãn
 - Các phương pháp chính: clustering
- Phương pháp “prompt engineering” ?
 - Sử dụng các mô hình ngôn ngữ lớn

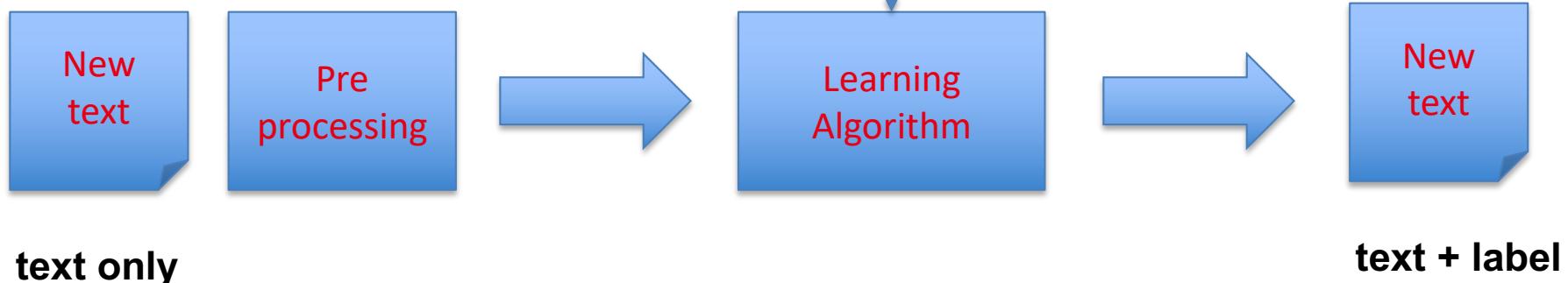


Học máy có giám sát

- Gồm 2 giai đoạn chính:
 - Giai đoạn 1: Huấn luyện mô hình:



- Giai đoạn 2: Sử dụng mô hình để đoán dữ liệu mới:



Các bước chung để giải bài toán khai thác văn bản

- Phân tích bài toán lớn và chia thành các bài toán nhỏ:
 - Document classification
 - Information Extraction
 - ..
- Với mỗi bài toán nhỏ:
 - Xác định nguồn dữ liệu
 - Tiền xử lý dữ liệu
 - Gán nhãn dữ liệu
 - Xây dựng mô hình
 - Kiểm thử / đánh giá
 - Tinh chỉnh mô hình
 - Triển khai
- **Kết hợp các phương pháp khác nhau để giải các bài toán**



Học máy và Học sâu

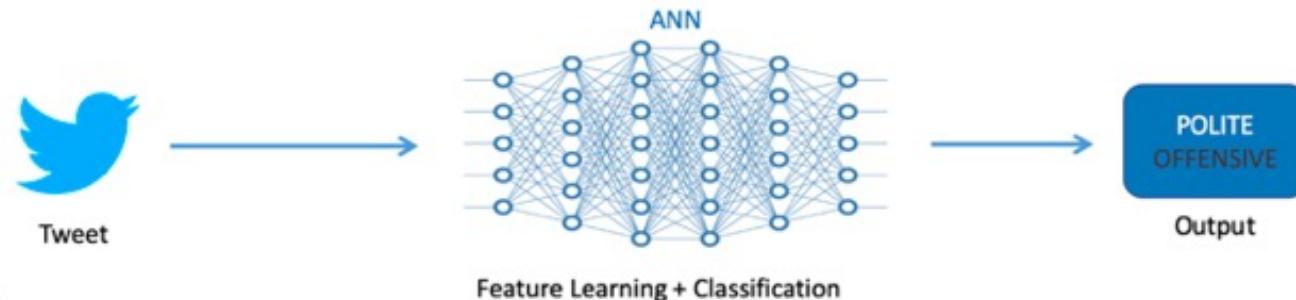
Học máy
truyền
thống

TRADITIONAL MACHINE LEARNING



Học sâu

DEEP LEARNING



Các mô hình deep-learning phổ biến: CNN, LSTM, Bert, ...

Khai thác văn bản và xử lý ngôn ngữ tự nhiên

- Xử lý ngôn ngữ tự nhiên là thành phần quan trọng trong khai thác văn bản.
 - Rút trích đặc trưng cho các bài toán khai thác văn bản:
 - Đặc trưng từ, từ loại,
- Các bài toán trong ngôn ngữ tự nhiên:
 - Tách từ (Word segmentation)
 - Gán nhãn từ loại (Part-Of-Speech tagging)
 - Phân tích quan hệ cú pháp (Syntactic analysis)
 - Phân tích quan hệ ngữ nghĩa (Semantic analysis)
 - Rút trích thực thể
 - Rút trích quan hệ
 - ...



Khai thác văn bản và xử lý ngôn ngữ tự nhiên

Stanford CoreNLP 4.0.0 (updated 2020-04-16)

- Các công cụ:
 - Tiếng Anh:
 - Standford parser
 - Tiếng Việt
 - VncoreNLP

"Ông Nguyễn Khắc Chúc đang làm việc tại Đại học Quốc gia Hà Nội. Bà Lan, vợ ông Chúc, cũng làm việc tại đây."



`[['Ông', 'Nguyễn_Khắc_Chúc', 'đang',
'làm_việc', 'tại', 'Đại_học', 'Quốc_gia',
'Hà_Nội', '.'], ['Bà', 'Lan', ',', 'vợ', 'ông',
'Chúc', ',', 'cũng', 'làm_việc', 'tại', 'đây', '.']]`

— Text to annotate —

Ha Nοi is a big city in Vietnam.

— Annotations —

parts-of-speech × named entities × dependency parse × openie ×

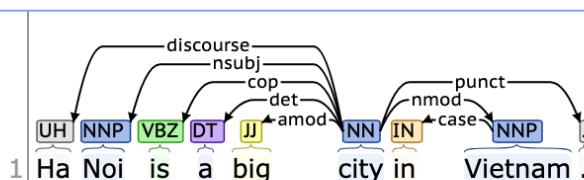
Part-of-Speech:

1 UH NNP VBZ DT JJ NN IN NNP
1 Ha Nοi is a big city in Vietnam .

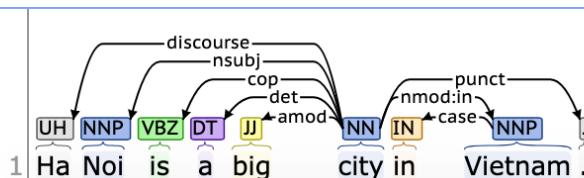
Named Entity Recognition:

PERSON COUNTRY
1 Ha Nοi is a big city in Vietnam .

Basic Dependencies:



Enhanced++ Dependencies:



Nội dung môn học

Khai thác văn bản và ứng dụng



KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Mục tiêu môn học cần đạt được

- Hiểu và tóm tắt lại các bài toán chính trong lĩnh vực khai thác văn bản, cho ví dụ
- Biết vận dụng một số công cụ xử lý ngôn ngữ tự nhiên để xử lý dữ liệu văn bản
- Biết vận dụng các thuật toán / công cụ để áp dụng vào bài toán cụ thể
- Tự đặt ra bài toán, thực hiện toàn bộ các bước để giải bài toán:
 - Xác định các yêu cầu nghiệp vụ
 - Phát biểu thành bài toán có thể sử dụng các kỹ thuật của khai thác văn bản
 - Đưa ra giải pháp



Mục tiêu môn học cần đạt được

- Thực hành:
 - Sử dụng thành thạo ngôn ngữ Python
 - Sử dụng các thư viện xử lý ngôn ngữ tự nhiên phục vụ cho bài toán khai thác văn bản (nltk, vncorenlp, ...)
 - Tiền xử lý dữ liệu
 - Sử dụng / Huấn luyện word embedding
 - Sử dụng một số thư viện học để giải các bài toán phân loại văn bản, nhận diện thực thể có tên
 - Sklearn, tensorflow, torch, huggingface, ...
 - Sử dụng một số công cụ để xây dựng hệ tìm kiếm thông tin
 - Elastic search, faiss, ...



Các nội dung chính của môn học

- Giáo viên trình bày:
 - 1. Tổng quan về khai thác văn bản
 - 2. Tiền xử lý và biểu diễn văn bản, phân loại văn bản
 - 3. Rút trích thông tin
 - 4. Tìm kiếm thông tin
 - 5-6-7. Deep learning
 - Word embedding
 - Introduction to Deep learning
 - CNN, RNN
 - Attention Transformer BERT
 - 8. Mô hình ngôn ngữ lớn – GPT1, 2, 3, ChatGPT, ...
 - 9, 10. Các chủ đề khác

- Tuần 9 đến 10: Học viên trình bày seminar

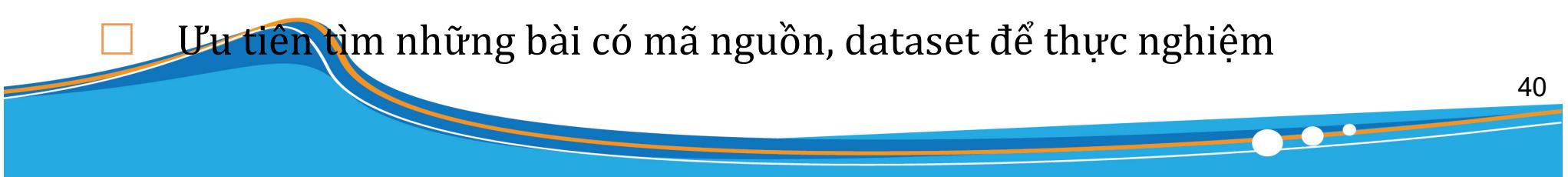
Đồ án seminar

- Các bước thực hiện:
 - Tìm một/nhóm bài báo trên các hội nghị “top” về một trong các chủ đề trên (chọn những bài có cả mã nguồn)
 - Viết báo cáo tìm hiểu và trình bày trước lớp



Một số từ khoá

- Hội nghị TOP: ACL, EMNLP, HLT-NAACL, COLING, CoNLL, IJCNLP
 - ACL 2023 all sessions:
 - https://2023.aclweb.org/program/accepted_main_conference/
 - https://2023.aclweb.org/program/best_papers/
 - EMNLP 2022
 - <https://aclanthology.org/events/emnlp-2022>
 - ACL 2022 all sessions:
 - <https://underline.io/events/284/sessions?eventSessionId=10659>
 - EMNLP 2021
 - <https://2021.emnlp.org/papers>
- Ưu tiên tìm những bài có mã nguồn, dataset để thực nghiệm



Một số từ khoá

- Danh sách chủ đề - giúp tìm bài báo phù hợp:
 - Word/document representation techniques
 - aspect sentiment analysis
 - information extraction / open IE
 - text summarization
 - question answering
 - Text generation
 - recognizing textual entailments, natural language inference
 - text mining in domains: healthcare / bioinformatics, finance, legal
 - Data augmentation
 - Discourse parsing
 - ...
- <https://paperswithcode.com/dataset/>

Tìm dataset (bài toán) sau đó tìm các bài báo hội nghị tốt.

Ví dụ

- E.g Question Answering:
 - WikiHowQA: A Comprehensive Benchmark for Multi-Document Non-Factoid Question Answering
 - MultiTabQA: Generating Tabular Answers for Multi-Table Question Answering
 - A Survey on Asking Clarification Questions Datasets in Conversational Systems
 - Single Sequence Prediction over Reasoning Graphs for Multi-hop QA
 - Elaboration-Generating Commonsense Question Answering at Scale
 - Using contradictions improves question answering systems
- Relation Extraction:
 - More than Classification: A Unified Framework for Event Temporal Relation Extraction
 - WebIE: Faithful and Robust Information Extraction on the Web
 -



Đồ án thực hành

- Yêu cầu đồ án:
 - Project proposal
 - Xây dựng / Tìm kiếm dữ liệu huấn luyện
 - Chọn lựa mô hình, thực hiện các thí nghiệm, huấn luyện và đánh giá so sánh các mô hình
 - Xây dựng ứng dụng minh họa
 - **Project report + Project presentation**



Đồ án thực hành

- Tìm hiểu và xây dựng một ứng demo cho một bài toán:
 - Phân loại văn bản (Text categorization)
 - Phân tích loại tin tức trực tuyến / social listening
 - Xây dựng hệ thống hỗ trợ phân loại tin tức tự động từ Internet
 - Phân tích reviews:
 - Phân tích review của một sản phẩm từ các trang web thương mại điện tử (tiki)
 - Rút trích các khía cạnh và review cho các khía cạnh của sản phẩm
 - Gán nhãn thực thể có tên (Named entity recognition) / Rút trích thông tin (Information Extraction)
 - Nhận diện thực thể có tên & social Listening
 - Phát hiện các sự kiện đang được diễn ra
 - Xây dựng một chatbot / hỏi đáp hỗ trợ cho một lĩnh vực nào đó:
 - Hỏi các thông tin chung: thời tiết, xổ số, giá vàng, trả lời các câu hỏi là ai, cái gì (Wikipedia)
 - Legal question answering: Hỏi đáp pháp luật
 - Legal Texts, Finance, Ecommerce ...
 -

Đánh giá môn học

- Điểm đồ án thực hành: 4 điểm
- Điểm seminar lý thuyết: 4 điểm
- Bài tập: 2 điểm



Tài liệu tham khảo / công cụ

- Tài liệu:
 - Books:
 - Fundamental of predictive text mining
 - The Text Mining handbook
 - Internet
- Ngôn ngữ lập trình/thư viện:
 - Python: Nltk, spacy, Tensorflow, Sklearn, torch, huggingface, ...
 - Java: Weka, ...
 - ...
- Các công cụ xử lý ngôn ngữ tự nhiên
 - Tiếng Anh, Tiếng Việt



Bài tập 1:

- Input: Tên bài báo
- Output:
 - Kiểm tra bài báo đó có publish code hay chưa, nếu có trả về liên kết của **github**
- Tools:
 - **python, selenium, requests, ...**

