



Introduction to
Machine Learning and Data Mining
(Học máy và Khai phá dữ liệu)

Khoat Than

Le Minh Hoa, Nguyen Van Son

School of Information and Communication Technology

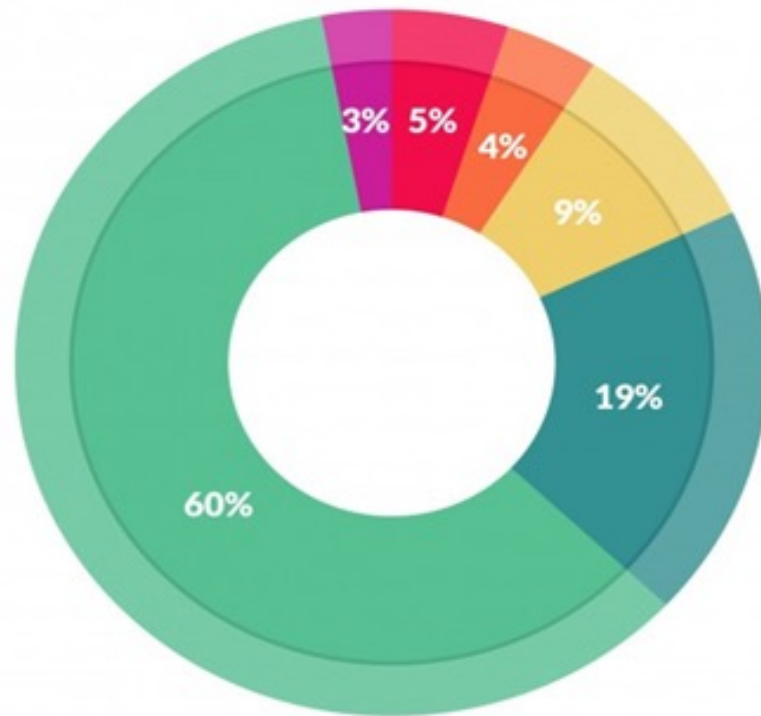
Hanoi University of Science and Technology

2021

Content

- Introduction to Machine Learning & Data Mining
- **Data crawling and pre-processing**
- Supervised learning
- Unsupervised learning
- Practical advice

Quỹ thời gian



CrowdFlower Inc., 2016

- Thời gian dành cho phân tích dữ liệu ra sao?
 - Thu thập dữ liệu: 19%
 - Thu xếp và làm sạch dữ liệu: 60%
 - Tạo tập dữ liệu huấn luyện: 3%
 - Khai phá: 9%
 - Cải thiện thuật toán: 4%
 - Khác: 5%

Why?

■ Tiền xử lý để làm gì

- Thuận tiện trong lưu trữ, truy vấn
- Các mô hình học máy thường làm việc với dữ liệu có cấu trúc: ma trận, vectơ, chuỗi,...
- Học máy thường làm việc hiệu quả nếu có **biểu diễn dữ liệu phù hợp**

Input

Vấn đề cần giải quyết của
lĩnh vực



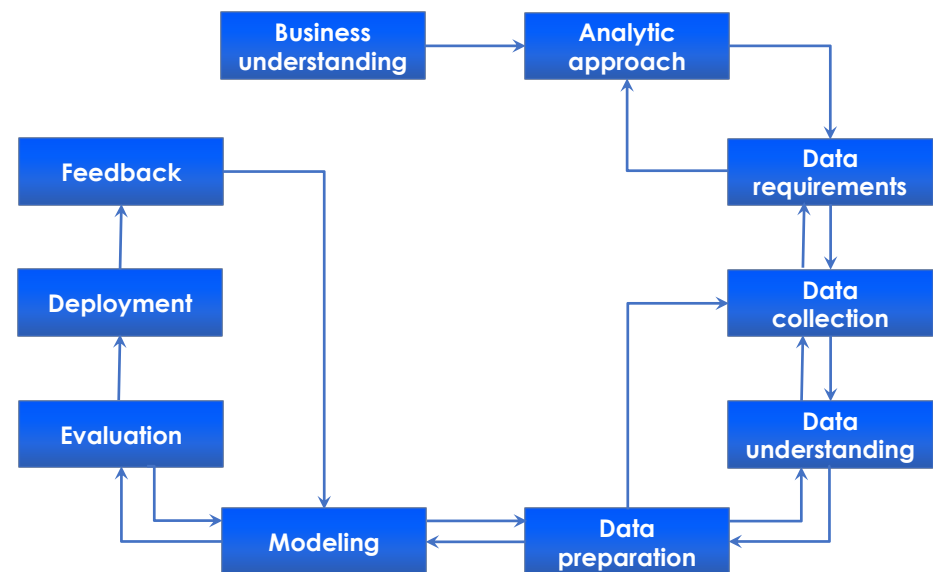
Output

Dữ liệu số - ma trận vector

$$x^{(n)} = \begin{bmatrix} -0.0920 \\ 3.4931 \\ -1.8493 \\ \dots \\ \dots \\ -0.2010 \\ -1.3079 \end{bmatrix} \quad \mathcal{D} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \dots \\ x^{(n)} \end{bmatrix}$$

How?

- Thu thập dữ liệu
 - Lấy mẫu (sampling)
 - Kỹ thuật: crawling, logging, scraping
- Xử lý dữ liệu
 - Lọc nhiễu, làm sạch, số hoá,...



Data collection

Input

Vấn đề cần giải quyết



Output

Mẫu dữ liệu

	A	B	C	D	E	F	G
1	Country	Region	Population	Under15	Over60	Ferti	LifeExp
2	Zimbabwe	Africa	13724	40.24	5.68	3.64	54
3	Zambia	Africa	14075	46.73	3.95	5.77	55
4	Yemen	Eastern M	23852	40.72	4.54	4.35	64
5	Viet Nam	Western P	90796	22.87	9.32	1.79	75
6	Venezuela (Bo	Americas	29955	28.84	9.17	2.44	75
7	Vanuatu	Western P	247				
8	Uzbekistan	Europe	29544				
9	Uruguay	Americas					

Fundamentals :: Sampling

- **WHAT** – lấy tập mẫu nhỏ, phổ biến để đại diện cho lĩnh vực cần học.
- **WHY** – không thể học toàn bộ. Giới hạn về thời gian và khả năng tính toán
- **HOW** – thu thập các mẫu từ thực tế, hoặc các nguồn chứa dữ liệu web, database,...

"One or more small spoon(s) can be enough to assess whether the soup is good or not."



<https://www.coursera.org/learn/inferential-statistics-intro>

Fundamentals :: Sampling :: How

- **Variety** – tập mẫu thu được đủ đa dạng để phủ hết các ngữ cảnh của lĩnh vực.
- **Bias** – dữ liệu cần tổng quát, không bị sai lệch, thiên vị về 1 bộ phận nhỏ nào đó của lĩnh vực.

"One or more small spoon(s) can be enough to assess whether the soup is good or not."

Remember to stir to avoid tasting biases.



<https://www.coursera.org/learn/inferential-statistics-intro>

Techniques

- **Crowd-sourcing:** Survey – *thực hiện các khảo sát*
- **Logging:** lưu lại lịch sử tương tác của người dùng, truy cập sản phẩm,...
- **Scrapping:** tìm kiếm nguồn dữ liệu trên các website, tải về, bóc tách, lọc,...

Techniques :: Scrapping :: DEMO

- **Mục tiêu:** Dữ liệu cho bài toán phân loại văn bản – miền báo chí.
- **DEMO:** Hệ thống crawl dữ liệu báo

DEMO

Input

Vấn đề: phân loại văn bản
báo chí



Output

Mẫu dữ liệu: báo chí và
nhãn tương ứng

A screenshot showing a file explorer window on the left and a JSON file viewer on the right. The file explorer shows a directory structure with folders like 'Dẫn trí', 'Ban đọc', 'Đời sống', etc. The JSON file viewer shows the following content:

```
1 {
2   "date": "2018-05-20, 07:44:1
3   "code": "651ab2f45f0305220d
4   "labels": "D\u00e2n tr\u00e0
5   "content": "\nD\u00e2n tr\u00
6   "image_url": "https://dantr
7   "url": "http://dantri.com.v
8   "domain": "dantri.com.vn",
9   "title": "B\u00e0 Giang: \\"/>A screenshot of a file explorer window showing a directory structure. The selected file is a JSON file named "651ab2f45f0305220d1f57bb21913620f75d128d.json". The JSON content is displayed in a text editor on the right, showing fields like "date", "code", "labels", "content", "image_url", "url", "domain", and "title".
```


DEMO :: Sample

JSON

- date : "2018-05-20, 07:44:00-07:00"
- code : "651ab2f45f0305220d1f57bb21913620f75d128d"
- labels : "Dân trí/Bạn đọc"
- content : " Dân trí Sau khi Bí thư Tỉnh ủy Bắc Giang yêu cầu dẹp tan nạn xe quá tải trong năm 2018, Phòng CSGT Công an tỉnh Bắc Giang đã tổ chức ra quân"
- image_url : "https://dantrcdn.com/zoom/80_50/2018/5/20/7-1526776517717498023080.png"
- url : "http://dantri.com.vn/ban-doc/bac-giang-doan-xe-coi-noi-thung-ram-rap-chay-qua-mat-can-sat-giao-thong-20180520074415778.htm"
- domain : "dantri.com.vn"
- title : "Bắc Giang: Đoàn xe coi nói thủng rầm rập chạy qua mặt cảnh sát giao thông?"


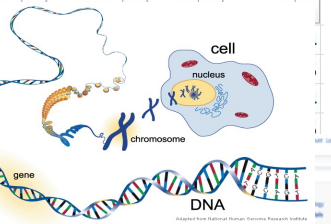
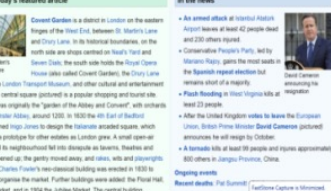
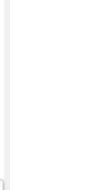

Data preprocessing

Input

Mẫu dữ liệu thô
(text, ảnh, audio, ...)

Output

Dữ liệu số theo từng ML/AI
model(s)

	A	B	C	D	E	F	G
1	Country	Region					
2	Zimbabwe	Africa					
3	Zambia	Africa					
4	Yemen	Eastern M					
5	Viet Nam	Western P					
6	Venezuela (Bo	Americas					
7	Vanuatu	Wester					
8	Uzbekistan	Europe					
9	Uruguay	America					

$$x^{(n)} = \begin{bmatrix} -0.0920 \\ 3.4931 \\ -1.8493 \\ \dots \\ \dots \\ -0.2010 \\ -1.3079 \end{bmatrix}$$

$$D = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \dots \\ x^{(n)} \end{bmatrix}$$

Fundamentals :: Data “rawness”

Completeness (đầy đủ)

Từng mẫu thu thập nên đầy đủ thông tin các trường thuộc tính cần thiết

Integrity (trung thực)

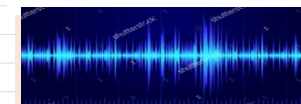
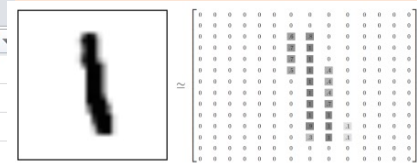
- Nguồn thu thập chính thống, đảm bảo mẫu thu được chứa giá trị chính xác trên thực tế.
- Jan. 1 as *everyone's* birthday? – *intentional (systematic) noises*

Homogeneity (đồng nhất)

- Rating “1, 2, 3” & “A, B, C”; or Age = “42” & Birthday = “03/07/2010” (*inconsistency*)
- Heterogenous data sources / schemas

Structures (cấu trúc)

C	D	E	F
Population	Under15	Over60	Fertil
13724	40.24	5.68	3.64
14075	46.73	3.95	5.77
23852	40.72	4.54	4.35
90796	22.87	9.32	1.79
29955	28.84	9.17	2.44
247	37.37	6.02	3.46
28541	28.9	6.38	2.38
3395	22.05	18.59	2.07



Techniques

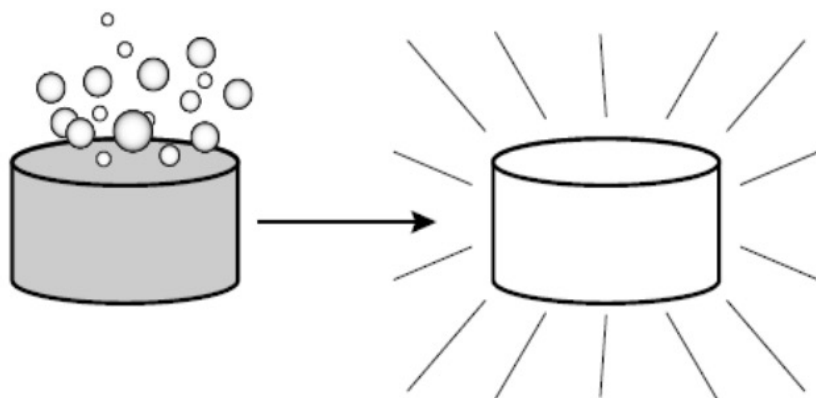
Cleaning

Integrating

Transforming

Techniques :: Cleaning

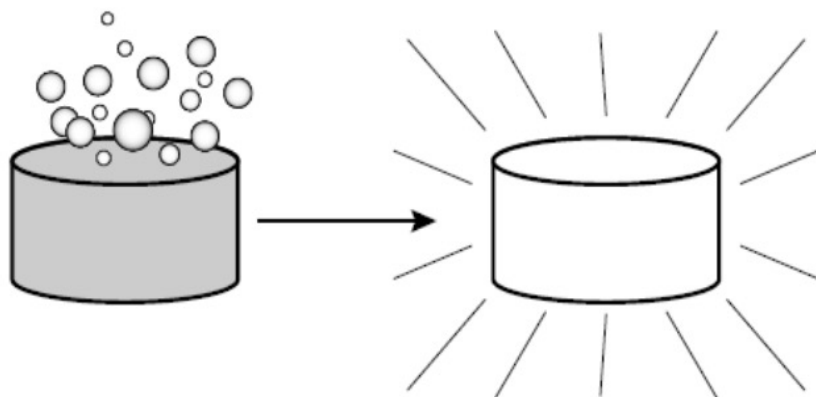
■ Tính đầy đủ + trung thực



- Mẫu dữ liệu cần được thu thập từ các **nguồn đáng tin cậy**. Phản ánh vấn đề cần giải quyết.
- Loại bỏ **những** (ngoại lai): bỏ vài mẫu dữ liệu mà có khác biệt lớn với các mẫu khác.
- Một mẫu dữ liệu có thể **bị trống** (thiếu, chưa đầy đủ), cần có chiến lược phù hợp:
 - Bỏ qua, không đưa vào phân tích?
 - Bổ sung các trường còn thiếu cho mẫu?

Techniques :: Cleaning

■ Điền giá trị thiếu

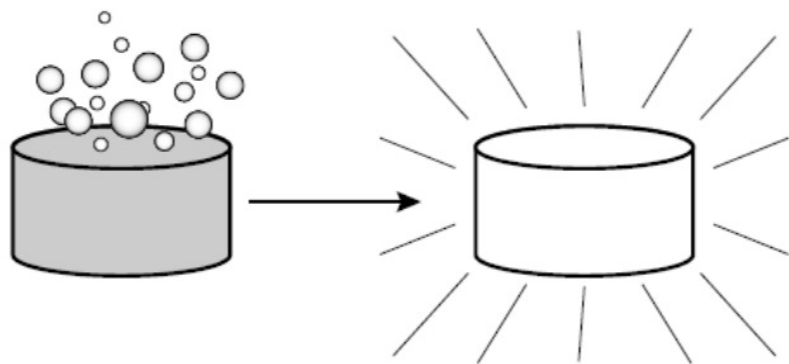


- Điền lại giá trị bằng tay
- Gán cho giá trị nhãn đặc biệt hay ngoài khoảng biểu diễn
- Gán giá trị trung bình cho nó.
- Gán giá trị trung bình của các mẫu khác thuộc cùng lớp đó.
- Tìm giá trị có xác suất lớn nhất điền vào chỗ bị mất (hồi quy, suy diễn Bayes,...)

A1	A2	A3	A4	A5	A6	A7	A8	y
?	3.683	?	-0.634	1	0.409	7	30	5
?	?	60	1.573	0	0.639	7	30	5
?	3.096	67	0.249	0	0.089	?	80	3
2.887	3.870	68	-1.347	?	1.276	?	60	5
2.731	3.945	79	1.967	1	2.487	?	100	4

Techniques :: Cleaning (cont.)

- Tính đồng nhất



Các mẫu dữ liệu cần có tính đồng nhất về cách biểu diễn, ký hiệu.

Ví dụ không đồng nhất:

Rating “1, 2, 3” & “A, B, C”;

Age = 42 & Birthday = 03/08/2020

Techniques :: Integrating w/ some Transforming

Un-structured

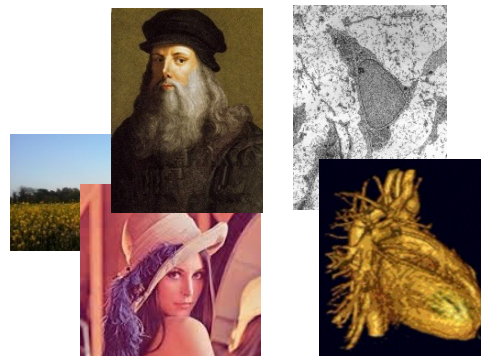
	A	B	C	D	E	F	G
1	Country	Region	Population	Under15	Over60	Fertil	LifeExp
2	Zimbabwe	Africa	13724	40.24	5.68	3.64	54
3	Zambia	Africa	14075	46.73	3.95	5.77	55
4	Yemen	Eastern M	23852	40.72	4.54	4.35	64
5	Viet Nam	Western P	90796	22.87	9.32	1.79	75
6	Venezuela (Bo	Americas	29955	28.84	9.17	2.44	75
7	Vanuatu	Western P	247	37.37	6.02	3.46	72
8	Uzbekistan	Europe	28541	28.9	6.38	2.38	68
9	Uruguay	Americas	3395	22.05	18.59	2.07	77

```
{
  "code": "1473a6fd39d1d8fa48654aac9d8cc2754232",
  "title": "[Updating] Câu chuyện xuyên mưa về :",
  "url": "http://techtalk.vn/updating-cau-chuyen",
  "labels": "techtalk/Cong nghe",
  "content": "Vào chiều tối ngày 09/12/2016 vừa",
  "image_url": "",
  "date": "2016-12-10T03:51:10Z"
}
```

texts in websites, emails, articles, tweets



2D/3D images, videos + meta



spectrograms, DNAs, ...

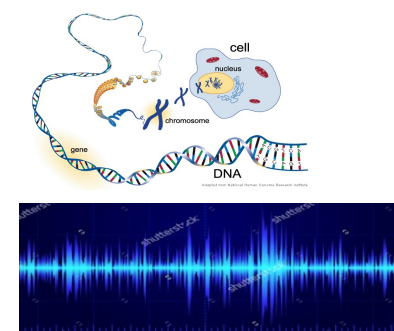


image credits: wikipedia, shutterstock, CNN

Techniques :: Transforming

Semantics?

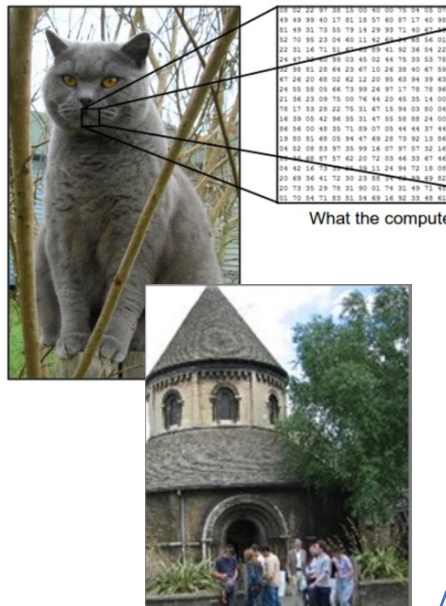
Trích xuất các **đặc trưng ngữ nghĩa**, chuẩn hóa

Semantics example: visual data

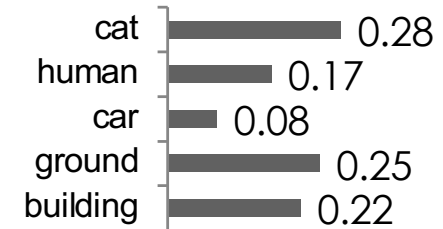
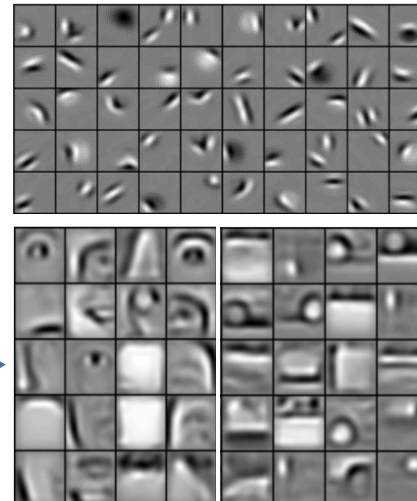
Low-level semantics
(raw pixels)



Mid-/High-level semantics
(e.g. human-interpretable features)



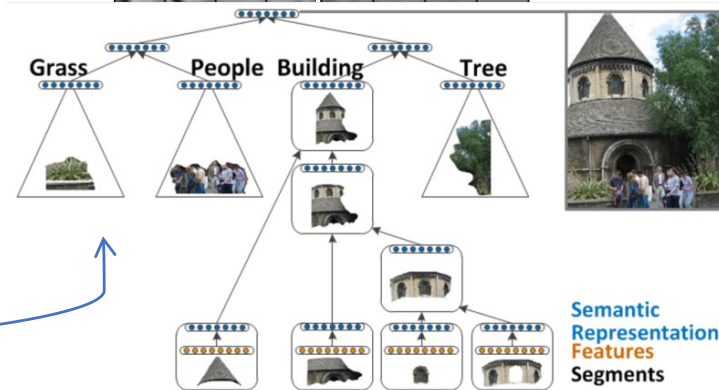
What the computer



cat → **not on** → car
 people ← **behind** ← building
 car → **is** → red

Mức ngữ nghĩa tối thiểu để có thể hiểu:

- Phân loại văn bản
- Phân tích cảm xúc
- AI Chatbot (nhiều mức ngữ nghĩa khác nhau)



Semantic Representations Features Segments

	C	D	E	F
Population	13724	40.24	5.68	3.64
Under15	14075	46.73	3.95	5.77
Over60	23852	40.72	4.54	4.35
Fertil	90796	22.87	9.32	1.79
	29955	28.84	9.17	2.44
	247	37.37	6.02	3.46
	28541	28.9	6.38	2.38
	3395	22.05	18.59	2.07

Techniques :: Transforming (cont.)

■ Mục tiêu: trích xuất các đặc trưng ngữ nghĩa.

USD điều_chỉnh trái chiều , vàng SJC quay đầu tăng

```
(0, 24506)    0.2077168092100841
(0, 23857)    0.34468369118902636
(0, 22309)    0.31713411814089415
(0, 21894)    0.3025597601047669
(0, 21265)    0.2449372095782497
(0, 20409)    0.3276089788346888
(0, 17739)    0.515839529548281
(0, 16499)    0.33820735665113805
(0, 4648)     0.3132633187744836
```

B	C	D	E	F	G
Region	Populat	Under1	Over60	Fertil	LifeExp
Africa	-0.416	0.748	-0.483	0.299	54
Africa	-0.403	1.464	-0.850	1.881	55
Eastern M	-0.060	0.801	-0.725	0.826	64
Western P	2.287	-1.169	0.289	-1.075	75
Americas	0.154	-0.511	0.257	-0.592	75
Western P	-0.888	0.431	-0.411	0.165	72
Europe	0.104	-0.504	-0.334	-0.637	68
Americas	-0.778	-1.260	2.256	-0.867	77

One-hot encoding

1 = [1 0 0 0]

3 = [0 0 1 0]

...

$$\frac{x - \bar{x}}{s}$$

- Từng lĩnh vực cụ thể, từng loại dữ liệu sử dụng các kỹ thuật xuất đặc trưng ngữ nghĩa khác nhau (dữ liệu text, hình ảnh, ...)

... and standardize

- Feature discretization* (rời rạc hoá): một số thuộc tính tỏ ra hiệu quả hơn khi được gom nhóm các giá trị.
- Feature normalization*: chuẩn hóa giá trị thuộc tính, về cùng một miền giá trị, dễ dàng trong tính toán.

Techniques :: Transforming (cont.)

- Giảm kích cỡ:
 - Giúp giảm kích thước của dữ liệu và đồng thời giữ được ngữ nghĩa cốt lõi của dữ liệu.
 - Giúp tăng tốc quá trình học hoặc khai phá tri thức.
- Vài chiến lược:
 - **Lựa chọn đặc trưng (feature selection)**: các thuộc tính không liên quan, dư thừa hoặc các chiều cũng có thể xóa hay loại bỏ
 - **Giảm chiều (dimension reduction)**: dùng một số thuật toán (ví dụ PCA, ICA, LDA,...) để biến đổi dữ liệu ban đầu về không gian có ít chiều hơn.
 - **Trừu tượng hoá**: các giá trị dữ liệu thô được thay thế bằng các khái niệm trừu tượng.

Techniques :: Transforming example & demo

Transforming text data

DEMO

Input

Mẫu dữ liệu thô: json text

```
{
  "image_url": "https://i-kinhdoanh.vnecdn.net/2018/1",
  "url": "https://kinhdoanh.vnexpress.net/tin-tuc/eb",
  "title": "Sacombank n\u00e2ng c\u00e5p nhi\u00e9u",
  "code": "db274d03b9a61aa16d70c7fd68929d799058b866",
  "domain": "kinhdoanh.vnexpress.net",
  "date": "2018-05-25, 17:00:00+07:00",
  "content": "\nHi\u00e0n th\u00e2y qu\u00e2n t\u00e2y",
  "labels": "vnexpress/Kinh doanh/Ebank\u00a0/Kinh d"
}
```

Output

Dữ liệu số theo từng ML/AI model(s)

```
(0, 24003) 0.08875917745394017
(0, 23874) 0.08543368833593054
(0, 23214) 0.06269100273800875
(0, 23085) 0.10941900286727153
(0, 22547) 0.047792971979914244
(0, 22446) 0.05082334424962779
(0, 21910) 0.08271656588481778
(0, 21905) 0.06404674731000018
(0, 21779) 0.11899134180006703
(0, 21572) 0.08401328893873479
(0, 20984) 0.0603014300399073
(0, 20928) 0.03425727291794896
(0, 20851) 0.04139691505815508
(0, 20796) 0.06515117203347312
(0, 20272) 0.09576360104259622
(0, 20254) 0.21906274633402326
(0, 19934) 0.09329205643046397
(0, 19928) 0.0815770967825164
```

DEMO :: Steps

Tokenize

Hiện thẻ quốc tế Sacombank Visa gồm các dòng thẻ tín dụng, thẻ thanh toán và thẻ trả trước. Các sản phẩm này có tiện ích chung như thanh toán, rút tiền khắp thế giới, mua sắm trực tuyến, nhận giảm giá đến 50% tại hàng trăm điểm chấp nhận thẻ liên kết. Thẻ hỗ trợ chi tiêu trước, thanh toán sau miễn lãi tối đa 55 ngày, tích lũy điểm thưởng để đổi quà, mua hàng trả góp lãi suất 0%...

Chủ thẻ có thể thanh toán nhanh chóng, thuận tiện trên phạm vi toàn cầu bằng cách chạm thẻ hoặc chạm điện thoại có cài ứng dụng Samsung Pay (đồng thời tích

['Hiện', 'thẻ', 'quốc tế', 'Sacombank', 'Visa', 'gồm', 'các', 'dòng', 'thẻ', 'tín dụng', ',', ',', 'thẻ', 'thanh toán', 'và', 'thẻ', 'trả', 'trước', ',', ',', 'sản phẩm', 'này', 'có', 'tiện ích', 'chung', 'như', 'thanh toán', ',', ',', 'rút tiền', 'khắp', 'thế giới', ',', ',', 'mua sắm', 'trực tuyến', ',', ',', 'nhận', 'giảm giá', 'đến', '50', ',', '%', 'tại', 'hàng', 'trăm', 'điểm', 'chấp nhận', 'thẻ', 'liên kết', ',', ',', 'Thẻ', 'hỗ trợ', 'chi tiêu', 'trước', ',', ',', 'thanh toán', 'sau', 'miễn', 'lãi', 'tối đa', '55', 'ngày', ',', ',', 'tích lũy', 'điểm', 'thưởng', 'để', 'đổi', 'quà', ',', ',', 'mua hàng', 'trả góp', 'lãi suất', '0', ',', '%', '...', 'Chủ', 'thẻ', 'có thể', 'thanh toán', 'nhanh chóng', ',', ',', 'thuận tiện', 'trên', 'phạm vi', 'toàn cầu', 'bằng', 'cách', 'chạm', 'thẻ', 'hoặc', 'chạm', 'điện thoại', 'có', 'cài', 'ứng dụng', 'Samsung', 'Pay', '(', '(', 'đồng thời', 'tích hợp', 'Sacombank', 'Visa', '(', '(', 'lên', 'các', 'máy', 'POS', 'NFC', 'Ngoài ra', ',', ',', 'người', 'dùng', 'còn', 'có thể', 'chi tiêu', 'thông qua', 'tính năng', 'quét', 'mã', 'QR', 'trên', 'ứng

Dictionary

```
{'dân_trí': 6928, 'sở': 17869, 'gd': 7729, 'dt': 23214, 'tỉnh': 218, 'sgddt': 17039, 'vp': 21572, 'chấn_chính': 4971, 'tiếp_thị': 16, 'giáo_dục': 7955, 'chỉ_đạo': 5092, 'tuyệt_đối': 20254, 'phép': 0, '16194, 'mua_bán': 12653, 'dụng_cụ': 7191, 'học_tập': 9557, 'g': 63, 'tổ_chức': 20928, 'ngành': 13667, 'tham_gia': 18129, 'giới_th': 'ua': 12651, 'phát_hành': 15346, 'tham_khảo': 18130, 'phụ_huynh': 'ng': 14805, 'lành_mạnh': 11553, 'chương_trình': 4935, 'phổ_thông': 'ai_sốt': 16816, 'báo_cáo': 3493, 'hướng': 9359, 'sơ': 17704, 'đề': 'cán_bộ': 5693, 'chuyên_viên': 4681, 'đồ_dùng': 24003, 'công_khai': 'g': 15421, 'ngăn_chặn': 13743, 'báo': 3490, 'thông_tin': 18676, '5492, 'chư_păn': 4929, 'tờ': 20984, 'giấy': 8066, 'thông_báo': 18, 'thị': 18993, 'nga': 13400, 'hiệu_trưởng': 8753, 'hôm': 9267, 'xá': 004, 'chim': 4524, 'non': 14434, 'học': 9534, 'hốt': 9259, 'bảo_đ': 50, 'địa_phương': 23924, 'đặc_điểm': 23836, 'loài': 11400, 'nghiê': 12940, 'noron': 14632, 'thần_kinh': 18881, 'trách_nhiệm': 19790, 'ông_bố': 5853, 'ấn_bản': 24292, '09': 168, '12': 348, 'tạp_chí': '132, 'trúc': 19889, 'não_bộ': 14521, 'thí_nghiệm': 18628, 'tiền_s': 'học': 17142, 'đại_học': 23619, 'cornell': 5477, 'đồng_nghiệp': 24, 4520, 'vta': 21588, 'ventral': 21329, 'tegmental': 18076, 'area': 8922, 'tín_hiệu': 20537, 'giúp': 7983, 'chim_sá': 4528, 'văn': 21
```

Data Input (tfidf-Vector)

```
(0, 24003) 0.08875917745394017
(0, 23874) 0.08543368833593054
(0, 23214) 0.06269100273800875
(0, 23085) 0.10941900286727153
(0, 22547) 0.047792971979914244
(0, 22446) 0.05082334424962779
(0, 21910) 0.08271656588481778
(0, 21905) 0.06404674731000018
(0, 21779) 0.11899134180006703
(0, 21572) 0.08401328893873479
(0, 20984) 0.0603014300399073
(0, 20928) 0.03425727291794896
(0, 20851) 0.04139691505815508
(0, 20796) 0.06515117203347312
(0, 20272) 0.09576360104259622
(0, 20254) 0.21906274633402326
(0, 19934) 0.09329205643046397
(0, 19928) 0.0815770967825164
(0, 19410) 0.06593571705754445
(0, 19370) 0.03950424960970291
(0, 19345) 0.16543313176963556
(0, 18993) 0.04356540203990621
(0, 18676) 0.03464691377452836
(0, 18659) 0.050391345485324875
```

DEMO :: Exercise

- **Bài tập:** Tính vector biểu diễn của văn bản với bộ dữ liệu nhỏ.
- **Dữ liệu:** 2 bài báo từ trang dân trí.
- **Yêu cầu:**
 - Sử dụng module tách từ.
 - Build tập từ điển từ 2 văn bản
 - Sử dụng stopwords lọc từ dừng.
 - Chuyển hoá 2 văn bản thành 2 vector tfidf

Summary

(Take-home messages)

- Dữ liệu trong một lĩnh vực trước khi vào hệ thống học máy phải được thu thập và biểu diễn thành dạng cấu trúc với một số đặc tính: đầy đủ, ít nhiễu, nhất quán, có cấu trúc xác định.
- Dữ liệu thu thập cho quá trình học là tập nhỏ, tuy vậy cần phản ánh đầy đủ các mặt vấn đề cần giải quyết.
- Dữ liệu thô sau khi thu thập và tiền xử lý phải giữ được sự đầy đủ các đặc trưng ngữ nghĩa – các đặc trưng ảnh hưởng đến khả năng giải quyết vấn đề.
- Khoa học dữ liệu là một lĩnh vực rộng, ngoài việc sử dụng công cụ áp dụng, nắm vững được các kiến thức cơ bản là điều quan trọng.