



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

Khoa T. Than

Le Minh Hoa, Nguyen Van Son

School of Information and Communication Technology

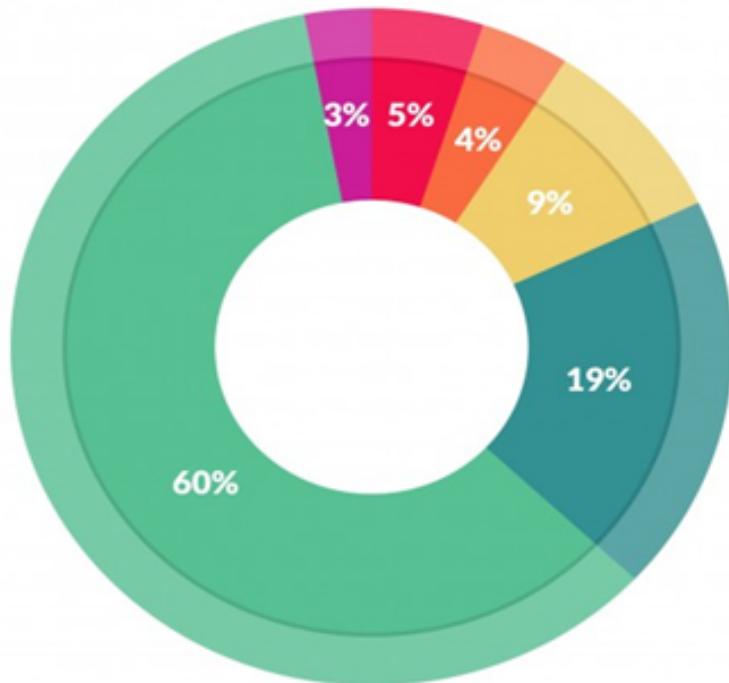
Hanoi University of Science and Technology

2020

Content

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

Quỹ thời gian



CrowdFlower Inc., 2016

- What data scientists spend the most time doing:
 - Buiding training sets: 3%
 - **Cleaning and organizing data: 60%**
 - **Collecting data sets: 19%**
 - Mining data for patterns: 9%
 - Refining algorithms: 4%
 - Others: 5%

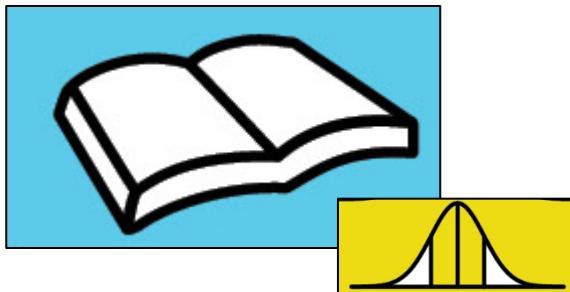
Why?

■ Tiền xử lý để làm gì

- Thuận tiện trong lưu trữ, truy vấn
- Các mô hình học máy thường chỉ làm việc với dữ liệu ma trận hoặc vectơ.
- Các mô hình học máy sẽ 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

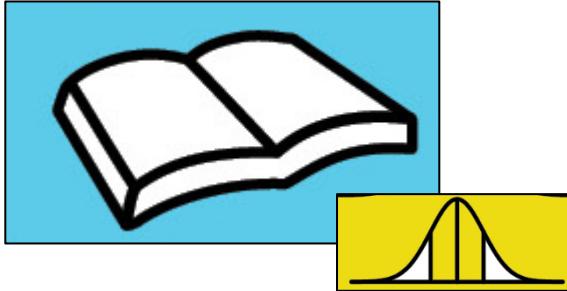
$$x^{(n)} = \begin{bmatrix} -0.0920 \\ 3.4931 \\ -1.8493 \\ \dots \\ \dots \\ -0.2010 \\ -1.3079 \end{bmatrix} \quad \mathcal{D} = \begin{bmatrix} -x^{(1)} & - \\ -x^{(2)} & - \\ \dots & \\ -x^{(n)} & - \\ \dots & \end{bmatrix}$$

How?

- Thu thập dữ liệu
 - Lấy mẫu (sampling)
 - Kỹ thuật: crawling, logging, scraping
- Xử lý dữ liệu
 - Dữ liệu cần lọc nhiễu, số hoá
 - Kỹ thuật: làm sạch, số hoá, lưu trữ

Thu thập dữ liệu

Input
Vấn đề cần giải quyết



Output
Mẫu dữ liệu

A	B	C	D	E	F	G
Country	Region	Population	Under15	Over60	Fertil	LifeExp
Zimbabwe	Africa	13724	40.24	5.68	3.64	54
Zambia	Africa	14075	46.73	3.95	5.77	55
Yemen	Eastern M	23852	40.72	4.54	4.35	64
Viet Nam	Western P	90796	22.87	9.32	1.79	75
Venezuela (Bo	Americas	29955	28.84	9.17	2.44	75
Vanuatu	Western P	247				
Uzbekistan	Europe	29541				
Uruguay	Americas					

Fundamentals :: Sampling

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

"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 ngũ cẩm của lĩnh vực.
- **Biases** – 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."

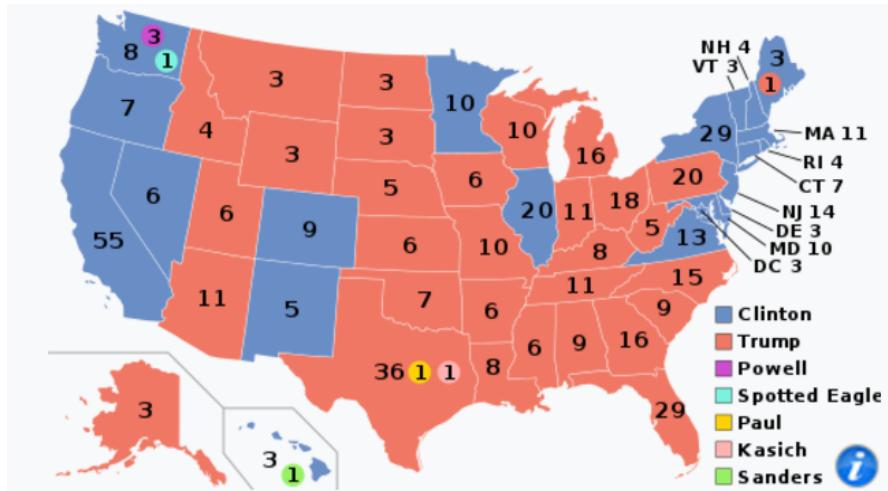
Remember to stir to avoid tasting biases.



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

Fundamentals :: Sampling :: How

- **Variety** – các mẫu đủ đa dạng để phản ánh khách quan ?



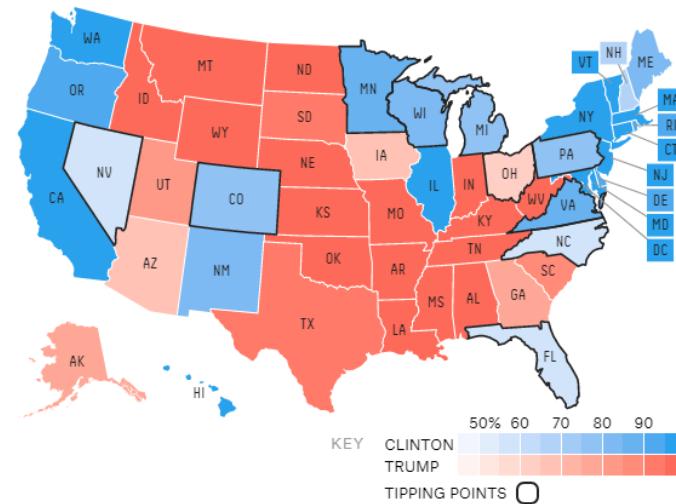
Actual results

<https://projects.fivethirtyeight.com/2016-election-forecast/>

<http://edition.cnn.com/election/results/president>

Image credit: Wikipedia, FiveThirtyEight

Chance of winning



Electoral votes

Clinton	302	2
Trump	235	0

Popular vote

Clinton	48	5%
Trump	44	9%

Techniques

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

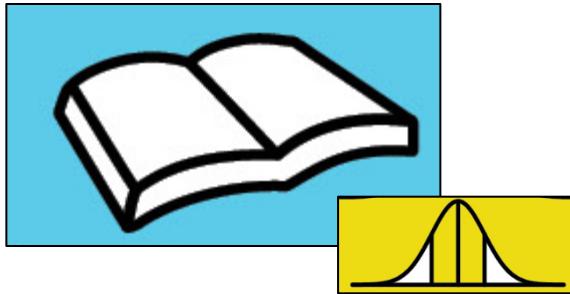
Techniques :: Scrapping :: DEMO

- **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

Danh mục	Name	Date modified
Bản đọc	2ce54553490dc5fb9a7153395793c6a648f..	5/25/2018 4:46 PM
Đời sống	7b228847f0349971fc59076def1b0eb5a9..	5/25/2018 4:46 PM
Giáo dục - Khuyến học	8af0882443701e020424acdef880cf8..	5/25/2018 4:46 PM
Khoa học - Công nghệ	949f342d8b858be7b0b6261def194d07917e..	5/25/2018 4:46 PM
Nhịp sống trẻ	146fd8057df18632a70e12bce428765504d..	5/25/2018 4:46 PM
Sức khỏe	651ab2459f0305200157b21913620f75d..	5/25/2018 4:46 PM
Sức mạnh thể thao	a1ef011572578ab4f3773a739fe5d2947..	5/25/2018 4:46 PM
Thể thao	6cb6d8552a3d7f3a733a739fe5d2947..	5/25/2018 4:46 PM
Văn hóa	e0efcbc74a582c6765077448ed7ddcc60d..	5/25/2018 4:46 PM
techtalk	e43e3696d676474946fcab0812d169e9..	5/25/2018 4:46 PM
viетbao		
vnexpress		
vtv		

DEMO :: Steps

Rss

Item

Content

Kênh do VnExpress cung cấp

Trang chủ

RSS

Thời sự

RSS

Thế giới

RSS

Kinh doanh

RSS

Startup

RSS

Giải trí

RSS

Thể thao

RSS

Pháp luật

RSS

Giáo dục

RSS

```
<rss xmlns:sisAdmin= "http://purl.org/rss/1.0/modules/sisAdmin/" version= "2.0 >
  <channel>
    <title>Kinh doanh - VnExpress RSS</title>
    <description>VnExpress RSS</description>
    <image>
      <url> https://s.vnecdn.net/vnexpress/i/v20/logos/vne_logo_rss.png
      </url>
      <title>Tin nhanh VnExpress - Đọc báo, tin tức online 24h</title>
      <link>https://vnexpress.net</link>
    </image>
    <pubDate>Thu, 07 Jun 2018 20:40:44 +0700</pubDate>
    <generator>VnExpress</generator>
    <link>https://vnexpress.net/rss/Kinh-doanh.rss</link>
  </channel>
</rss>
```

```
<article class="content_detail fck_detail width_common block_ads_connect">
  <p class="Normal">
    <span>
      Công ty TNHH MTV Xổ số điện toán Việt Nam (Vietlott) vừa trao giải cho khách hàng trúng Jackpot 1 sản phẩm Power 6/55 trị giá hơn 40 tỷ đồng (chưa trừ thuế) chiều ngày 7/6.
    </span>
  </p>
  <p class="Normal">
    <span>
      Nữ khách hàng may mắn trúng giải tên N.T, là nhân viên một ngân hàng tại TP HCM. Chia sẻ tại buổi trao thưởng, &nbsp;"</span>
    <span></span>
  </p>
  <table align="center" border="0" cellpadding="3" cellspacing="0" class="tblCaption" style="width: 100%; "></table>
  <p class="Normal">
    <span>
      Theo thông tin từ Vietlott, chi nhánh TP HCM của đơn vị này đã tiếp nhận chiếc vé trúng giải Jackpot 1 Power 6/55 từ một nữ khách hàng ngày 4/6. "
    </span>
  </p>
  <p class="Normal" style="text-align: right; font-size: small; margin-top: 10px; ">
    <span>
      "Qua kiểm tra trên hệ thống kỹ thuật và hồ sơ kèm theo, Vietlott xác định chiếc vé của chị N.T là hợp lệ và trúng giải Jackpot 1 Power 6/55 kỳ quay thứ 131. Tấm vé được phát hành tại điểm bán hàng đường số 6, phường Linh Chiểu, quận Thủ Đức, TP HCM."
    </span>
  </p>
  <p class="Normal"></p>
  <p class="Normal"></p>
</article>
```

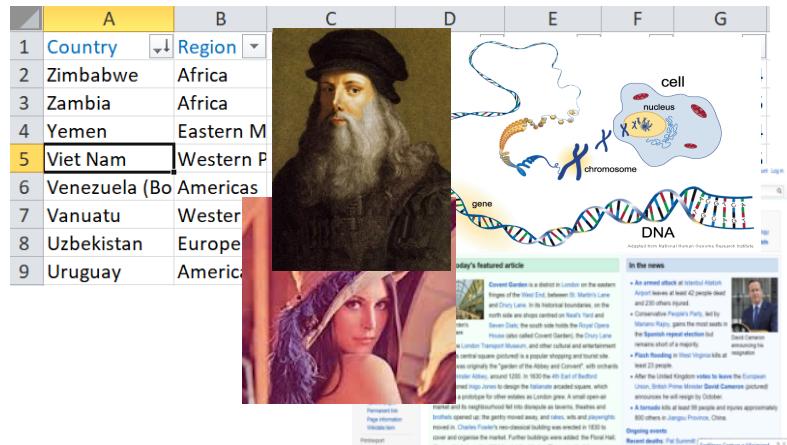
DEMO :: Sample

```
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 quâ
  image_url : "https://dantricdn.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-chayqua-mat-canh-sat-giao-thong-20180520074415778.htm"
  domain : "dantri.com.vn"
  title : "Bắc Giang: Đoàn xe cơi nới thùng 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)

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

Fundamentals :: Data “rawness”

Completeness (đầy đủ)

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

Homogeneity (đồng nhất)

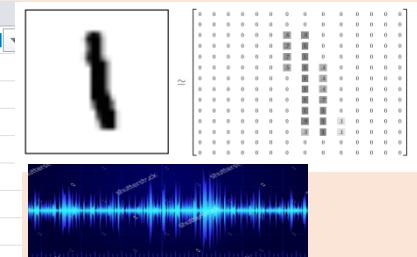
- Rating “1, 2, 3” & “A, B, C”; or Age = “42” & Birthday = “03/07/2010” (*inconsistency*)
- Heterogenous data sources / schemas

Integrity (rõ ràng)

- 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*

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

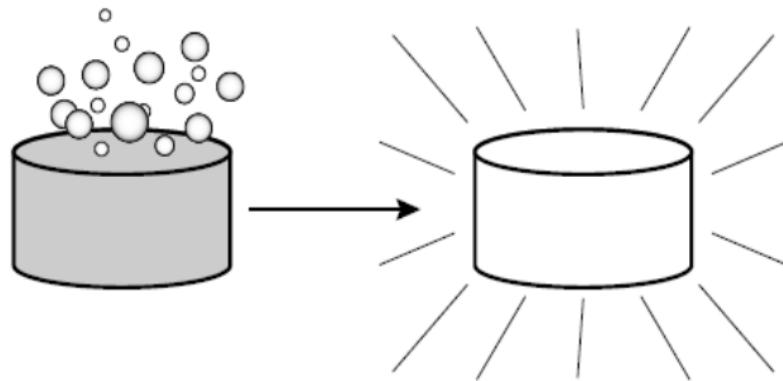
Cleaning

Integrating

Transforming

Techniques :: Cleaning

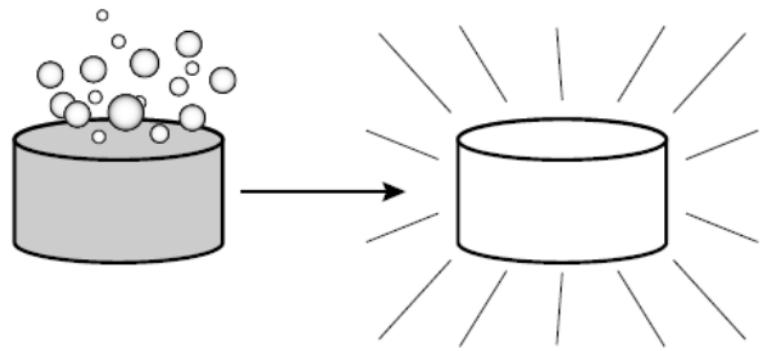
■ Tính đầy đủ + rõ ràng



- Mẫu dữ liệu cần được thu thập từ các nguồn đáng tin cậy. Phản ánh vấn đề cần giải quyết.
- Mẫu dữ liệu thu thập đôi khi không thể đầy đủ, cần có chiến lược phù hợp:
 - Bỏ qua, không đưa vào dữ liệu học.
 - Bổ sung các trường còn thiếu cho mẫu:
 - Bằng tay
 - Tự động (heuristic)

Techniques :: Cleaning (cont.)

■ Tính đồng nhất



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

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
Country	Region	Population	Under15	Over60	Fertil	LifeExp
Zimbabwe	Africa	13724	40.24	5.68 3.64		54
Zambia	Africa	14075	46.73	3.95 5.77		55
Yemen	Eastern M	23852	40.72	4.54 4.35		64
Viet Nam	Western P	90796	22.87	9.32 1.79		75
Venezuela (Bo Americas		29955	28.84	9.17 2.44		75
Vanuatu	Western P	247	37.37	6.02 3.46		72
Uzbekistan	Europe	28541	28.9	6.38 2.38		68
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

The collage illustrates the variety of unstructured data sources, including text from Wikipedia, social media profiles, news articles, and academic publications.

2D/3D images, videos + meta



spectrograms, DNAs, ...

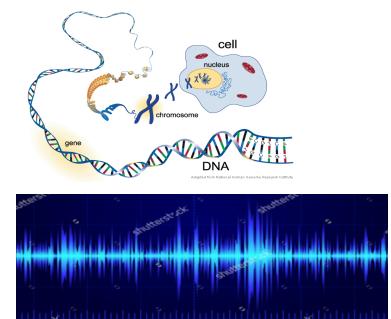


image credits: wikipedia, shutterstock, CNN

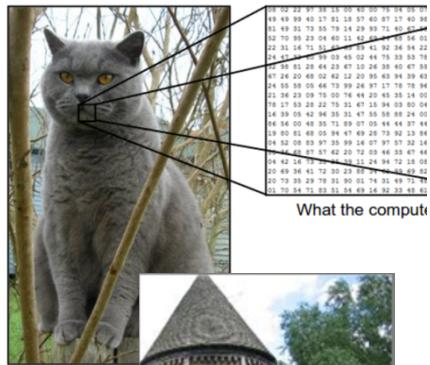
Techniques :: Transforming

Semantics?

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

Semantics example: visual data

Low-level semantics
(raw features)

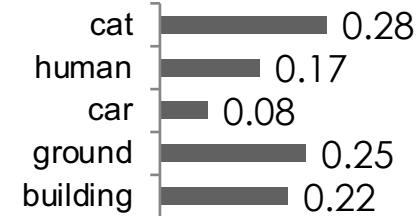
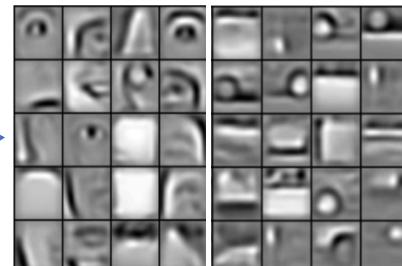
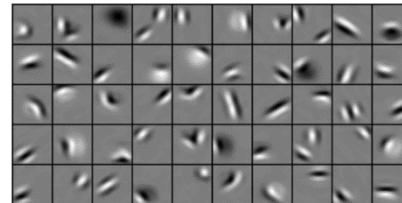


Mức ngữ nghĩa tối thiểu để có thể hiểu:

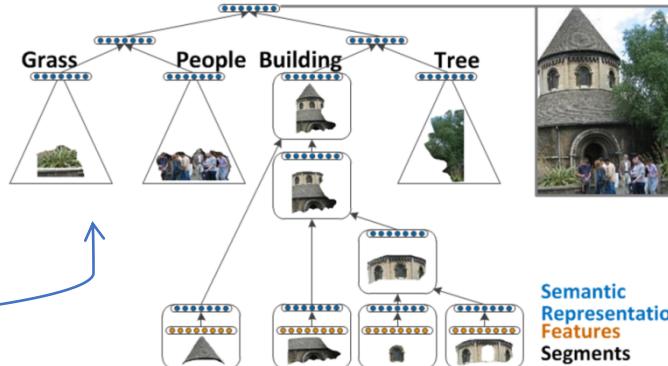
- Phân loại văn bản
- Phân tích cảm xúc
- AI Chatbot (nhiều mức ngữ nghĩa khác nhau)



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



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



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

Image credits: CS231n, Stanford University; Lee et al, 2009; Socher et al, 2011

Techniques :: Transforming example & demo

Transforming text data

Techniques :: Transforming (cont.)

- Mục tiêu: trích xuất các **đặc trưng ngữ nghĩa** của vấn đề.

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

- 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 – một số thuộc tính tỏ ra hiệu quả hơn khi được phân nhóm, sắp xếp trước.
- Feature normalization ← chuẩn hóa giá trị thuộc tính, về cùng một miền giá trị, dễ dàng trong tính toán.

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

$$\begin{aligned}1 &= [1 \ 0 \ 0 \ 0 \ 0] \\3 &= [0 \ 0 \ 1 \ 0 \ 0]\end{aligned}$$

...

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

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\u00e2n c\u00f3 m\u00e1y ATM t\u00f3n t\u00f3n v\u00e0o
  "code": "db274d03b9a61aa16d70c7fd68929d799058b866",
  "domain": "kinhdoanh.vnexpress.net",
  "date": "2018-05-25, 17:00:00+07:00",
  "content": "\nHi\u00e1n th\u00e1nh qu\u00e1n l\u00e2m t\u00f3n t\u00f3n v\u00e0o
  "labels": "vnexpress/Kinh doanh/Ebank\u00a0/Kinh do
}
```

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

Dictionary

Data Input
(tfidf-Vector)

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 trực, thanh toán sau miễn lãi tối đa 55 ngày, tích lũy điểm thường 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', ',', '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%', '%', 'tai', 'hang', 'trảm', 'diễn', 'chấp nhận', 'thẻ', 'liên kết', ',', 'Thẻ', 'hỗ trợ', 'chi tiêu', 'trả trước', ',', 'thanh toán', 'sau', 'miễn', 'lãi', 'tối đa', '55', 'ngày', ',', 'tích lũy', 'diễn', 'thường', 'để', 'đổi', 'qua', ',', 'mua hàng', 'trả góp', 'lãi suất', '0', '%', '...', 'Chủ', 'thẻ', 'có thẻ', 'thanh toán', 'nhanh chóng', ',', 'thuận tiện', 'trên', 'phạm vi', 'toàn cầu', 'bằng', 'cách', 'chạm', 'thẻ', 'hoặc', 'chạm', 'diện thoại', 'cố', 'cài', 'ứng dụng', 'Samsung', 'Pay', '(', 'đồng thời', 'tích hợp', 'Sacombank', 'Visa', ')', 'lên', 'các', 'máy', 'POS', 'NFC.', 'Ngoài ra', ',', 'người', 'dùng', 'còn', 'có thẻ', 'chi tiêu', 'thông qua', 'tinh năng', 'quét', 'mã', 'QR', 'trên', 'ứng

```
{"dân_trí": 6928, "sở": 17869, "gd": 7729, "dt": 23214, "tỉnh": 28, "sgđt": 17039, "vp": 21572, "chẩn_chính": 4971, "tiếp_thị": 16, "giáo_dục": 7955, "chỉ_dạo": 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_khoa": 18130, "phụ_huynh": 14805, "lành_mạnh": 11553, "chương_trình": 4935, "phô_thông": ai_sót": 16816, "báo_cáo": 3493, "hướng": 9359, "số": 17704, "đè_cán_bộ": 5693, "chuyên_viên": 4681, "đồ_dùng": 24003, "công_khai g": 15421, "ngăn_chặn": 13743, "báo": 3490, "thông_tin": 18676, "5492, "chú_päh": 4929, "tờ": 20984, "giấy": 8066, "thông_báo": 18 'thị': 18993, "nga": 13400, "hiệu_trưởng": 8753, "hôm": 9267, "xâ 004, "chim": 4524, "non": 14434, "học": 9534, "hốt": 9259, "bảo_d 50, "địa_phương": 23924, "đặc_diểm": 23836, "loài": 11400, "nghie 12940, "noron": 14632, "thần_kinh": 18881, "trách_nhiệm": 19790, ông_bô": 5853, "ấn_bản": 24292, "09": 168, "12": 348, "tập_chí": 132, "trúc": 19889, "não_bô": 14521, "thí_nghiệm": 18628, "tiến_s học": 17142, "đại_học": 23619, "cornell": 5477, "đồng_nghiệp": 24 4520, "vta": 21588, "ventral": 21329, "tegmental": 18076, "area": 8922, "tín_hiệu": 20537, "nhiều": 7983, "chim_sả": 4528, "vẫn": 21 /0 186561 0 050391345485224875
```

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.