HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.

ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Nhập môn
Khoa học dữ liệu
(IT4142)

PGS.TS Thân Quang Khoát & PGS.TS Phạm Văn Hải
Team lecturers

ONE LOVE. ONE FUTURE.

2

Contents

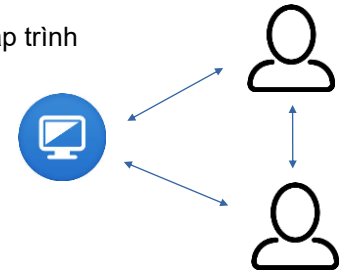
- Lecture 1: Tổng quan về Khoa học dữ liệu
- Lecture 2: Thu thập và tiền xử lý dữ liệu
- Lecture 3: Làm sạch và tích hợp dữ liệu
- Lecture 4: Phân tích và khám phá dữ liệu
- Lecture 5: Trực quan hoá dữ liệu
- Lecture 6: Trực quan hoá dữ liệu đa biến
- Lecture 7: Học máy
- Lecture 8: Phân tích dữ liệu lớn
- Lecture 9: Báo cáo tiến độ bài tập lớn và hướng dẫn
- **Lecture 10+11: Phân tích một số kiểu dữ liệu**
- Lecture 12: Đánh giá kết quả phân tích

Giới thiệu về XLNNTN

- NLP là gì?
- Các vấn đề chính trong NLP
- Các thách thức của NLP

Giới thiệu về XLNNTN: NLP là gì?

- **Ngôn ngữ tự nhiên**: công cụ để giao tiếp giữa con người với con người, giữa con người với máy tính
- **Định dạng**: tiếng nói, văn bản, hình ảnh, video
- **Thể loại**: các ngôn ngữ trên thế giới, các ngôn ngữ lập trình

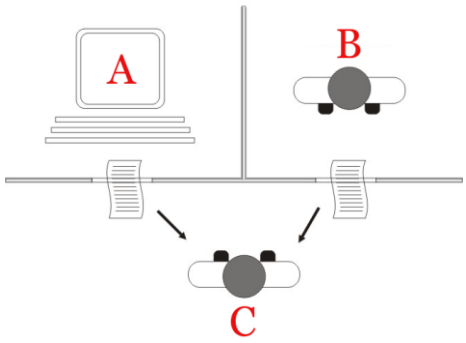


Giới thiệu về XLNNTN: NLP là gì?

- Hiểu, **mô hình hóa** được ngôn ngữ tự nhiên
- Làm cho máy tính và con người **giao tiếp** bằng ngôn ngữ tự nhiên
- Làm cho con người sử dụng các ngôn ngữ khác nhau có thể **giao tiếp** được
- **Khai thác** thông tin và tri thức được thể hiện qua ngôn ngữ để phục vụ các lĩnh vực của đời sống con người

Giới thiệu về XLNNTN: NLP là gì?

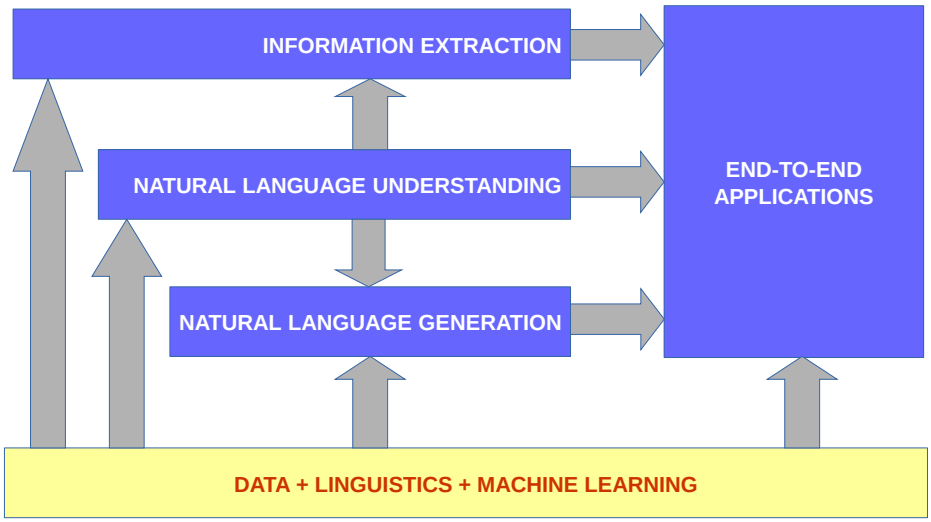
- 1950: Turing test



Giới thiệu về XLNNTN: NLP là gì?

- NLP liên quan đến:
  - Ngôn ngữ học, tâm lý học, triết học
  - Trí tuệ nhân tạo, học máy, dữ liệu lớn

Giới thiệu về XLNNTN: Các vấn đề chính trong NLP



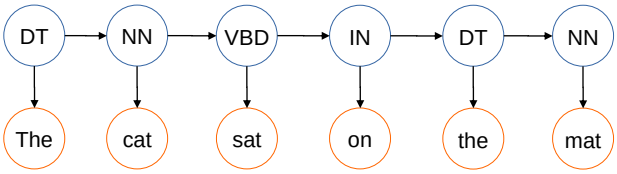
Giới thiệu về XLNNTN: Các vấn đề chính trong NLP

- Tách từ

"***xử_lý ngôn_ngữ tự_nhiên***"
(process language natural)

Giới thiệu về XLNNTN: Các vấn đề chính trong NLP

- Gán nhãn từ loại



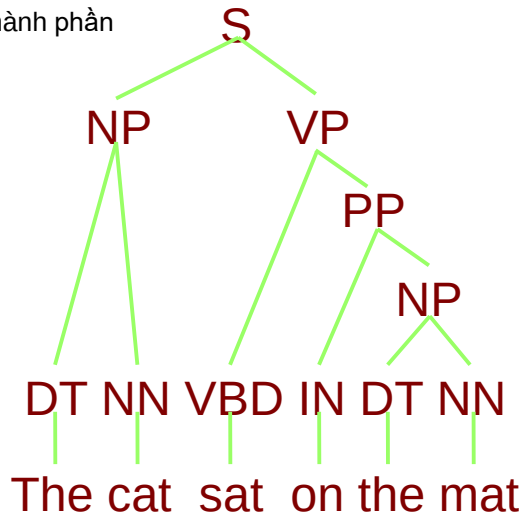
Giới thiệu về XLNNTN: Các vấn đề chính trong NLP

- Phân cụm

$$[\text{The cat}]_{NP}\ [\text{sat}]_{VP}\ [\text{on}]_{PP}\ [\text{the mat}]_{NP}$$

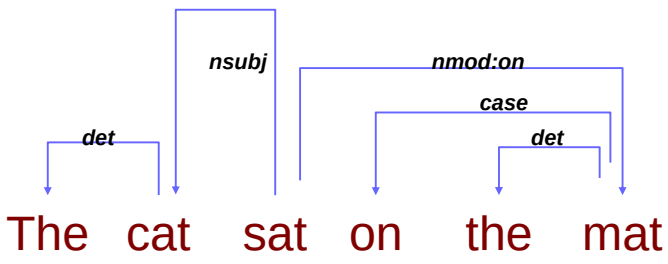
Giới thiệu về XLNNTN: Các vấn đề chính trong NLP

- Phân tích cú pháp thành phần

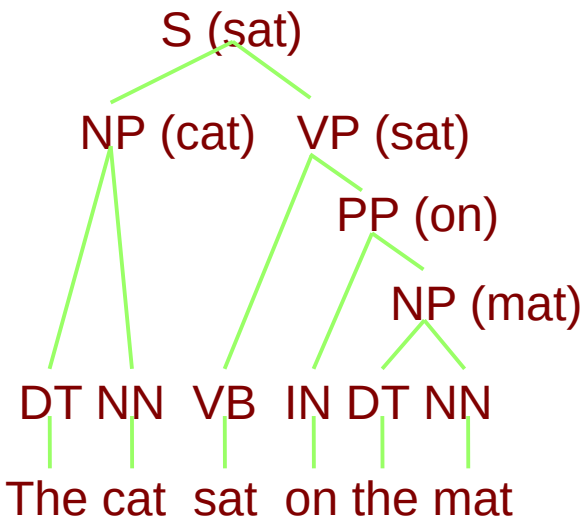


Giới thiệu về XLNNTN: Các vấn đề chính trong NLP

- Phân tích cú pháp phụ thuộc



Giới thiệu về XLNNTN: Các vấn đề chính trong NLP



Giới thiệu về XLNNTN: Các vấn đề chính trong NLP

- Phân giải đồng tham chiếu

doc#1

Alexis Sanchez stepped up **his** preparations for **Manchester United**'s FA Cup clash against **Yeovil** when **he** checked in for training on Thursday. **The Chile star** is in line to make **his** debut for the **Red Devils** in the fourth round tie at Huish Park following **his** move to Old Trafford. **Sanchez** met **his** new team-mates for the first time on Wednesday and could go straight into the matchday squad to face **the Glovers**. **Jose Mourinho** has so far given no indication on the strength of team that **he** will take to Somerset, but **Sanchez** will be keen to make **his** debut.

doc#2

Alexis Sanchez checks in for training as **he** prepares for **his** **Manchester United** debut

Giới thiệu về XLNNTN: Các vấn đề chính trong NLP

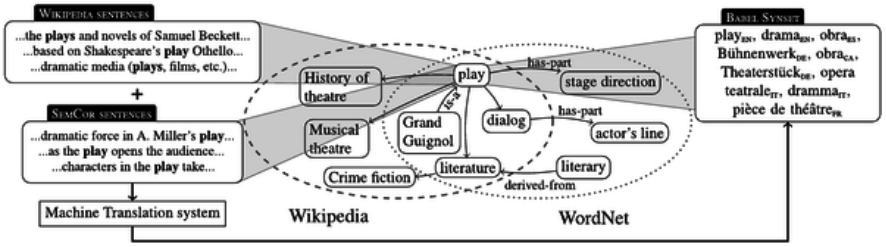
- Phân tích ngữ nghĩa

"Mouse love Rice"

"The history of computer mouse"

Giới thiệu về XLNNTN: Các vấn đề chính trong NLP

- Multilingual concept net



Giới thiệu về XLNNTN: Các vấn đề chính trong NLP

- Mô hình ngôn ngữ

"You are uniformly charming!" cried he, with a smile of associating and now and then I bowed and they perceived a`chaise and four to wish for.

Random sentence generated from a Jane Austen trigram model

From Dan Jurafsky 2018

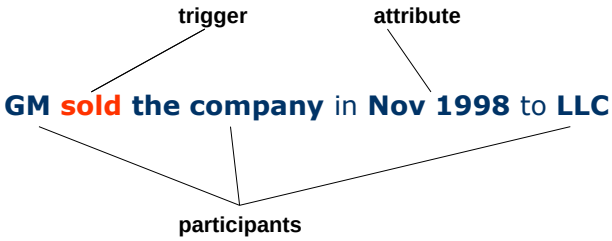
Giới thiệu về XLNNTN: Các vấn đề chính trong NLP

- Nhận diện thực thể có tên

[GM]$_{ORG}$ sold the company in [Nov 1998]$_{TIME}$ to [LLC]$_{ORG}$

Giới thiệu về XLNNTN: Các vấn đề chính trong NLP

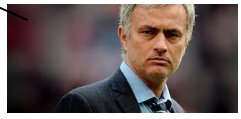
- Trích rút sự kiện



Giới thiệu về XLNNTN: Các vấn đề chính trong NLP

- Liên kết thực thể

Alexis Sanchez stepped up **his** preparations for **Manchester United**'s FA Cup clash against **Yeovil** when **he** checked in for training on Thursday. **The Chile star** is in line to make **his** debut for the **Red Devils** in the fourth round tie at Huish Park following **his** move to Old Trafford. **Sanchez** met **his** new team-mates for the first time on Wednesday and could go straight into the matchday squad to face **the Glovers**. **Jose Mourinho** has so far given no indication on the strength of team that **he** will take to Somerset, but **Sanchez** will be keen to make **his** debut.



Giới thiệu về XLNNTN: Các thách thức chính trong NLP

- Đặc tính cố hữu của ngôn ngữ là nhập nhằng
- Ngôn ngữ đa dạng và liên tục biến đổi
- Ngôn ngữ dưới tư cách đối tượng xử lý của các mô hình học máy thông kê, học sâu

THANK YOU !