Giới thiệu về
Xử lý ngôn ngữ tự nhiên

XLNNTN (NLP) là gì?

Hỏi đáp: Ứng dụng Watson của IBM

- Vô địch Jeopardy vào 16/02/2011

WILLIAM WILKINSON'S
"AN ACCOUNT OF THE PRINCIPALITIES
OF
WALLACHIA AND MOLDOVIA"
INSPIRED THIS AUTHOR'S
MOST FAMOUS NOVEL

Bram Stoker

Rút trích thông tin (Information Extraction)

Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Jura

Event: Curriculum mtg
Date: Jan-16-2012
Start: 10:00am
End: 11:30am
Where: Gates 159

Hi Dan, we've now scheduled the curriculum meeting.

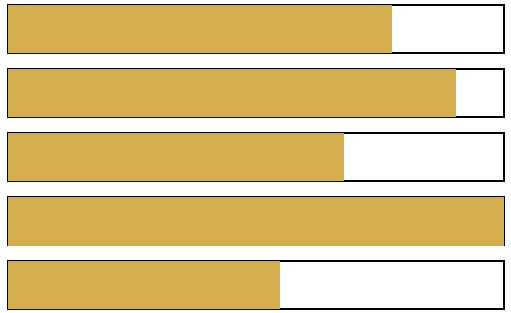
It will be in Gates 159 tomorrow from 10:00-11:30.

-Chris

Create new Calendar entry

Rút trích thông tin & Phân tích ý kiến (SA)

Đặc tính:



Size and weight

- ✓ nice and compact to carry!
- ✓ since the camera is small and light, I won't need to carry around those heavy, bulky professional cameras either!
- ✗ the camera feels flimsy, is plastic and very light in weight you have to be very delicate in the handling of this camera

Dịch máy (dịch tự động, machine translation)

- Tự động hoàn toàn

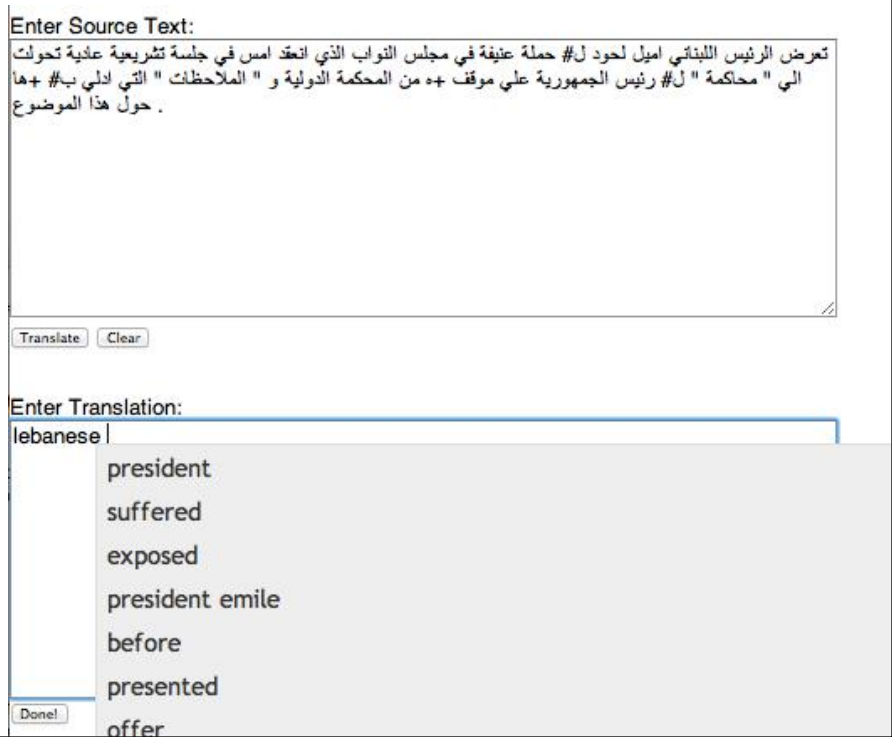
Nhập văn bản nguồn (Source Text):

这 不过 是 一 个 时间 的 问 题 .

Bản dịch:

This is only a matter of time.

- Hỗ trợ người dịch



Công nghệ xử lý ngôn ngữ (language technology)

Được giải quyết gần như hoàn toàn

Spam detection

Let's go to Agra! ✓

Buy V1AGRA … ✗

Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

Có tiến bộ nhanh

Sentiment analysis

Best roast chicken in San Francisco!

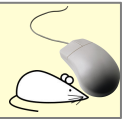
The waiter ignored us for 20 minutes.

Coreference resolution

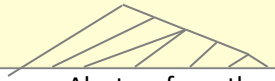
Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.



Parsing



I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕…

The 13th Shanghai International Film Festival…

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

Vẫn còn rất khó

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

⇨ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Sự nhập nhằng làm cho XLNNTN khó

Ông già đi nhanh quá.
Học sinh học sinh học.

Những thứ khác làm cho XLNNTN khó?

Không chuẩn

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

Vấn đề ranh giới từ

the New York-New Haven Railroad

the New York-New Haven Railroad

Thành ngữ

dark horse
get cold feet
lose face
throw in the towel

Từ mới - neologisms

unfriend
Retweet
bromance

Tri thức thế giới

Mary and Sue are sisters.

Mary and Sue are mothers.

Tên riêng gây rắc rối

Where is *A Bug's Life* playing …

Let It Be was recorded …

… a mutation on the *for* gene …

Nhưng đó cũng là những thứ làm cho XLNNTN thú vị!

Để đóng góp vào sự tiến bộ của lĩnh vực này

- Các tác vụ rất khó! Chúng ta cần công cụ gì?
  - Hiểu biết về ngôn ngữ
  - Hiểu biết về thế giới
  - Cách kết hợp các nguồn tri thức
- Phương pháp thường sử dụng:
  - Mô hình xác suất xây dựng từ dữ liệu ngôn ngữ
    - P(“maison” → “house”) cao
    - P(“L’avocat général” → “the general avocado”) thấp
  - May mắn thay, những thuộc tính ngôn ngữ đơn giản có thể giải quyết nửa phần việc.

Môn học

- Dạy những lý thuyết và phương pháp cơ bản về XLNNTN theo cách tiếp cận thống kê (statistical NLP):
  - Viterbi
  - Phân lớp Naïve Bayes, Maxent
  - Mô hình ngôn ngữ N-gram
  - Phân tích cú pháp
  - Chỉ mục ngược (inverted index), tf-idf, mô hình vector cho ngữ nghĩa
- Giới thiệu ứng dụng một số thực tiễn
  - Rút trích thông tin
  - Sửa lỗi chính tả
  - Truy vấn thông tin
  - Phân tích ý kiến
  - Hỏi đáp, chatbot

Kỹ năng cần có

- Đại số tuyến tính cơ bản (vector, ma trận)
- Lý thuyết xác suất cơ bản
- Lập trình Java hoặc Python