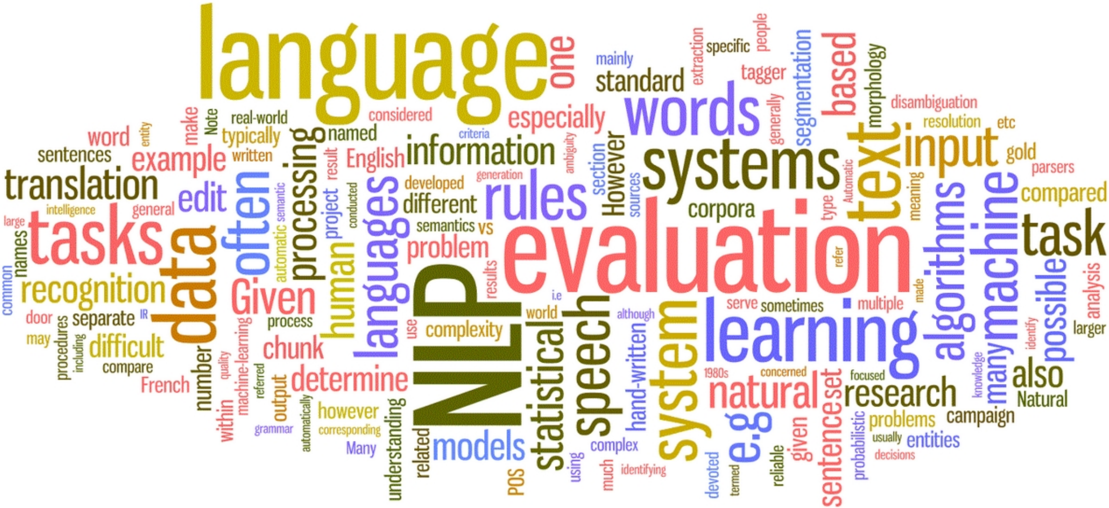
Xử lý ngôn ngữ tự nhiên

Chương 1: Giới thiệu xử lý ngôn ngữ tự nhiên

Khoa CNTT, Đại học Kỹ thuật - Công nghệ Cần Thơ
Lưu hành nội bộ

Nội dung

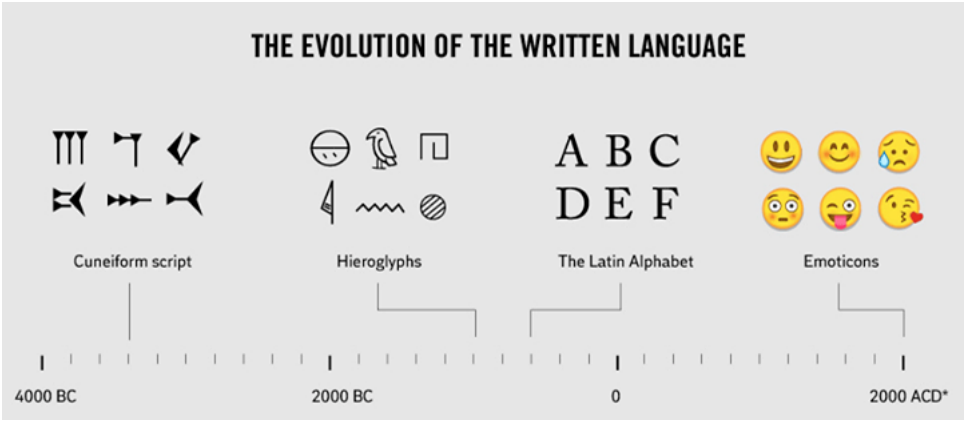
1. Ngôn ngữ tự nhiên
2. Ngôn ngữ học
3. Ngữ nghĩa học ngôn ngữ
4. Xử lý ngôn ngữ tự nhiên
5. Cài đặt thư viện scikit-learn và NLTK


language
one
mainly
standard
extraction
specific
people
tagger
segmentation
based
morphology
words
especially
considered
real-world
Note
make
entity
word
sentences
typically
named
criteria
disambiguation
resolution
etc
information
systems
text
input
gold
parsers
example
written
English
ambiguity
section
However
translation
processing
result
generation
developed
different
rules
sources
corpora
compared
intelligence
general
edit
often
semantic
project
conducted
languages
semantics
vs
Automatic
type
meaning
algorithms
machine
task
large
tasks
automatic
problem
evaluation
refer
names
common
recognition
data
human
Given
results
NLP
i.e
world
serve
sometimes
multiple
made
possible
analysis
larger
door
separate
IR
process
complexity
learning
although
identify
may
procedures
including
difficult
use
many
also
machine-learning
chunk
statistical
speech
hand-written
system
compare
quality
referred
1980s
concerned
set
focused
knowledge
number
output
determine
natural
research
Natural
French
however
e.g
sentence
problems
campaign
within
grammar
automatically
corresponding
understanding
given
probabilistic
usually
entities
Many
related
models
complex
decisions
POS
using
much
identifying
devoted
termed
reliable

Ngôn ngữ tự nhiên là gì?

- Textual data (dữ liệu dạng văn bản) là dữ liệu không có cấu trúc:
  - Thuộc về một ngôn ngữ
  - Có ngữ nghĩa và cú pháp
- Natural language (ngôn ngữ tự nhiên) là ngôn ngữ được phát triển, tiến hóa, và sử dụng bởi con người.
  - Hệ thống chữ viết
  - Hệ thống tiếng nói
  - Hệ thống ký hiệu
- Ngôn ngữ nhân tạo?

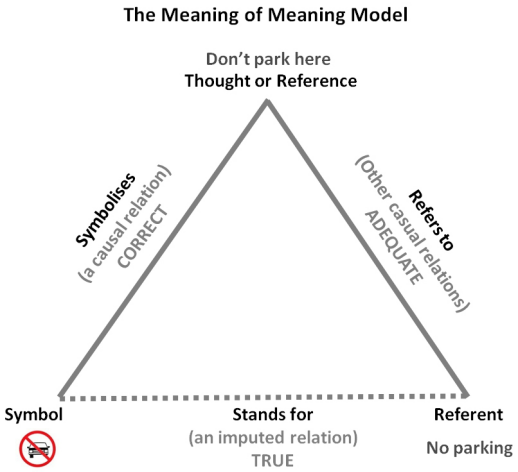
Tiến hóa của ngôn ngữ tự nhiên



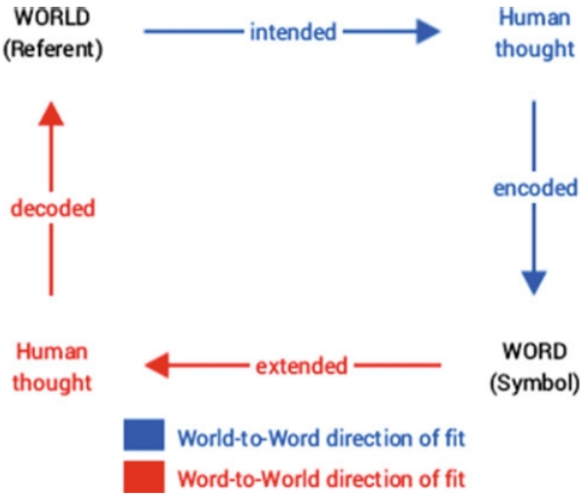
Triết học của ngôn ngữ

- Mọi ngôn ngữ tự nhiên giải quyết 4 vấn đề:
  1. Ý nghĩa tự nhiên của ngôn ngữ
  2. Sử dụng ngôn ngữ
  3. Nhận thức ngôn ngữ
  4. Mối quan hệ giữa ngôn ngữ và thực tế

Mô hình ý nghĩa của ý nghĩa



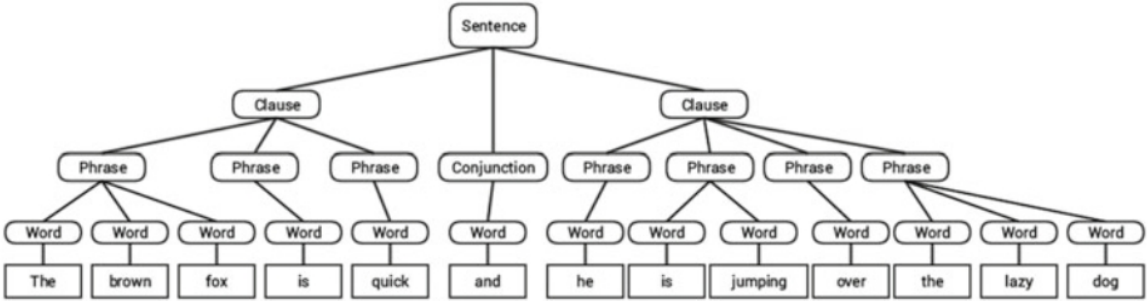
Mô hình ý nghĩa của ý nghĩa



Sự đạt được ngôn ngữ

- Language acquisition (sự đạt được ngôn ngữ) là quá trình con người:
  - sử dụng khả năng nhận thức (cognitive), kiến thức và kinh nghiệm để hiểu ngôn ngữ dựa trên nghe và nhận thức
  - sử dụng từ (word), ngữ (phrase), câu (sentence), và văn phạm (grammar) để giao tiếp
- -> thu được và sản xuất ngôn ngữ

dog the over he
lazy jumping is the fox
and is quick brown



Ngôn ngữ học

- Linguistics (ngôn ngữ học) là khoa học nghiên cứu về ngôn ngữ, bao gồm:
  - cú pháp và hình thức ngôn ngữ (form and syntax)
  - ý nghĩa (meaning)
  - ngữ nghĩa học (semantics): sử dụng (usage) và ngữ cảnh (context)

Lĩnh vực nghiên cứu của ngôn ngữ học

1. Phonetics (ngữ âm học): âm thanh của ngôn ngữ
2. Phonology (âm vị học): mẫu âm thanh (sound patterns)
3. Syntax (cú pháp): câu, ngữ, từ, cấu trúc (structure)
4. Semantics (ngữ nghĩa): ý nghĩa của ngôn ngữ
   - Lexical semantics (ngữ nghĩa từ vựng)
   - Compositional semantics (ngữ nghĩa hợp thành)

Lĩnh vực nghiên cứu của ngôn ngữ học

5. Morphology (hình thái học)
   - morpheme (hình vị): đơn vị nhỏ nhất của ngôn ngữ có thể phân biệt ý nghĩa
   - Hình thái học nghiên cứu về cấu trúc và ý nghĩa của hình vị
6. Lexicon (từ vựng): từ, ngữ, cách xây dựng từ vựng của ngôn ngữ
7. Pragmatics (ngữ dụng học): quan hệ giữa nhân tố ngôn ngữ và phi ngôn ngữ trong ngữ cảnh và tình huống; nghĩa tiềm ẩn của ngôn ngữ

Lĩnh vực nghiên cứu của ngôn ngữ học

8. Discourse analysis (sự phân tích ngôn từ): phân tích trao đổi thông tin thông qua các hình thức nói, viết, dấu hiệu

9. Stylistics (phong cách học): nghiên cứu cách sử dụng âm (tone), giọng (accent), văn phạm (grammar), loại âm (voice)

10. Semiotics (ký hiệu học): nghiên cứu về dấu (sign), biểu tượng (symbol)

Ngữ nghĩa học ngôn ngữ

- Language semantics (ngữ nghĩa học ngôn ngữ): nghiên cứu về ý nghĩa của ngôn ngữ, mối quan hệ giữa các từ, ngữ, biểu tượng và sự trình diễn kiến thức
  - Facial expression (sự biểu lộ nét mặt)
  - Body language (ngôn ngữ cơ thể)
  - Ký hiệu, biểu tượng, sự chuyển tải thông điệp

Mối quan hệ ngữ nghĩa từ vựng học

- Một lemma (bổ đề) trong ngôn ngữ học được hiểu là một canonical form (dạng chuẩn tắt) cho một tập các từ (word)
- Bổ đề thường là base form (dạng góc) của một tập các từ, gọi là lexeme (từ vị)
- Các từ vị được sinh ra từ word form (dạng từ) của một bổ đề
  - lemma `eat` là dạng chuẩn tắt của tập các lexeme từ nhiều word form khác nhau theo văn phạm {`eating`, `ate`, `eaten`, `eats`}

Mối quan hệ ngữ nghĩa từ vựng học

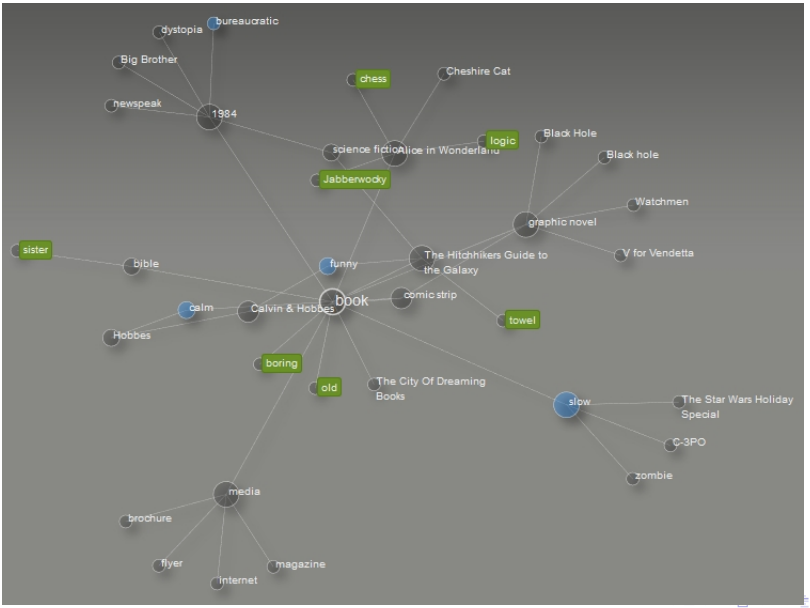
- Homonyms (từ đồng âm khác nghĩa): cùng cách phát âm nhưng khác ý nghĩa
  - *Text analytics with Python* is a really good book to learn natural language processing
  - You should book flight ticket several months before your actual journey
- Homonym bao gồm:
  - homograph (từ cùng chữ): cùng dạng viết nhưng khác cách đọc và ý nghĩa
  - homophone (từ cùng âm): cùng phát âm nhưng khác ý nghĩa

Mối quan hệ ngữ nghĩa từ vựng học

- Synonym (từ đồng nghĩa): các từ khác âm và cách viết nhưng cùng ý nghĩa
- Antonym (từ trái nghĩa): các cặp từ có mối quan hệ đối nghịch nhị phân (binary opposite relationship)
- Hyponym (từ thuộc nghĩa): các từ thuộc phân lớp dưới (subclass) của một từ lớp trên (superclass)
- Hypernym (từ nhóm nghĩa): các từ là lớp trên của các từ lớp dưới
  - `fruit` — `orange, mango`

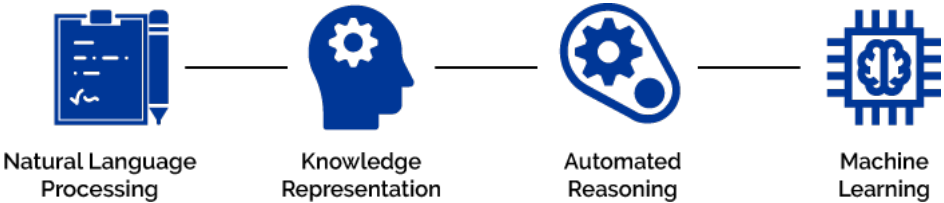
Mạng ngữ nghĩa

- Semantic networks (mạng ngữ nghĩa): hình thức hóa mối quan hệ giữa các từ và ý nghĩa; biểu diễn kiến thức và concept (khái niệm) bằng hình thức mạng và đồ thị
  - Đơn vị cơ bản của mạng ngữ nghĩa là entity (thực thể) hoặc khái niệm
  - Một khái niệm có thể cụ thể (tangible) hoặc trừ tượng (abstract)
  - Tập khác khái niệm có mối quan hệ được biểu diễn bằng cạnh có hướng hoặc vô hướng (directed or undirected edges)
  - Mỗi cạnh biểu diễn một mối quan hệ nhất định: `is-a`, `has-a`, `part-of`, `related-to`


dystopia
bureaucratic
Big Brother
chess
Cheshire Cat
newspeak
1984
science fiction
Alice in Wonderland
logic
Black Hole
Black hole
Jabberwocky
Watchmen
graphic novel
sister
bible
funny
The Hitchhikers Guide to the Galaxy
V for Vendetta
book
comic strip
calm
Calvin & Hobbes
towel
Hobbes
boring
old
The City Of Dreaming Books
slow
The Star Wars Holiday Special
C-3PO
zombie
media
brochure
flyer
magazine
internet

Tập sao lục văn bản

- Text corpora/corpus (tập sao lục văn bản): là sự sưu tầm có cấu trúc các văn bản hoặc dữ liệu văn bản bao gồm văn nói và văn viết; lưu trữ điện tử
- Được dùng để phân tích và xây dựng các công cụ xử lý ngôn ngữ
- `https://en.wikipedia.org/wiki/List_of_text_corpora`
  - Wordnet
  - Reuters corpus
  - Brown corpus


Natural Language Processing
Knowledge Representation
Automated Reasoning
Machine Learning

Xử lý ngôn ngữ tự nhiên

- Natural language processing - NLP (xử lý ngôn ngữ tự nhiên): là một nhánh của trí tuệ nhân tạo nghiên cứu, phân tích nhằm hiểu và tạo ngôn ngữ để giao tiếp với con người ở cả thể viết và nói.

- Lĩnh vực tương tác người-máy (Human-Computer Interaction)

Dịch máy

- Machine translation (dịch máy): nghiên cứu chuyển đổi giữa các ngôn ngữ của con người.
- Phát triển các kỹ thuật giúp cho việc dịch đúng ngữ nghĩa, văn phạm và cấu trúc của cặp ngôn ngữ bất kỳ
- Hệ thống dịch máy phổ biến nhất hiện nay: Google translate

Hệ thống nhận dạng lời nói

- Speech recognition systems (Hệ thống nhận dạng lời nói): hiểu được con người thông qua lời nói.
- Turing test:
  - Xác định mức độ thông minh của máy tính
  - Bài test *đạt* khi không phân biệt được người hay máy làm bài test đó

Chatbot

- Giao tiếp trao đổi với con người.
- Phân loại chatbot
  - Theo quy luật dựng sẵn
  - Thông minh dựa trên máy học + xử lý ngôn ngữ tự nhiên

Tổng hợp và phân loại văn bản

- Text summarization (tổng hợp văn bản): *nhìn* vào nhiều văn bản và cố gắng tìm kiếm từ khóa, ngữ, câu quan trọng, đại diện cho tất cả văn bản đó
  - Extraction-based summarization (tổng hợp dựa trên trích lọc)
  - Abstraction-based summarization (tổng hợp dựa trên trừu tượng hóa)
- Text categorization (phân loại văn bản): xác định loại (category) hoặc lớp (class) của một văn bản dựa trên nội dung

Phân tích văn bản

- Text analytics (phân tích văn bản) còn được gọi là text mining (khai mỏ văn bản): phương pháp luận và kỹ thuật trích lọc thông tin hợp lý và có chất lượng từ dữ liệu văn bản
- Kết hợp giữa NLP, sự lấy lại thông tin (information retrieval), máy học (machine learning) để phân tích (parse) dữ liệu văn bản phi cấu trúc sang dữ liệu văn bản có cấu trúc
- Dữ liệu hữu dụng cho người dùng cuối

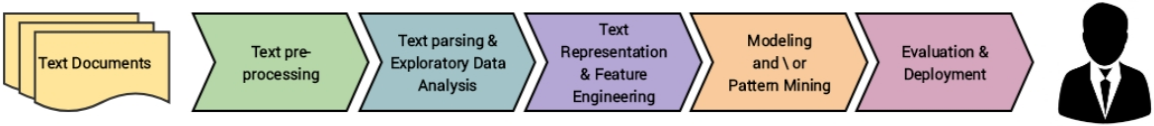
Phân tích văn bản

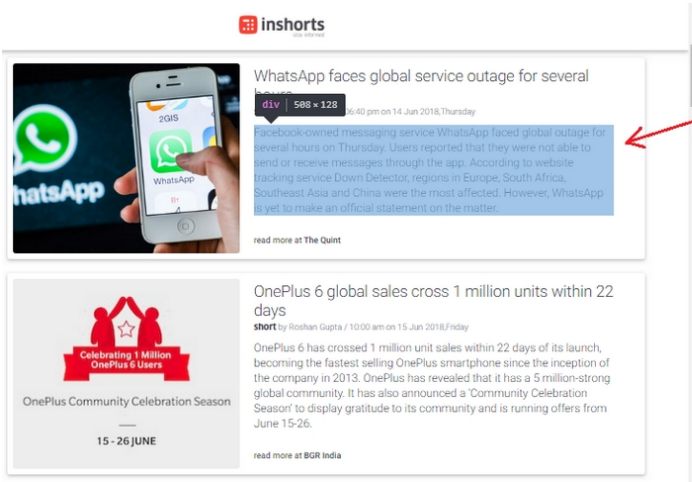
- Một số kỹ thuật và vận hành:
  - Text classification
  - Text clustering
  - Text summarization
  - Sentiment analysis (phân tích ẩn dụ, ngụ ý)
  - Entity extraction and recognition
  - Similarity analysis and relation modeling (phân tích sự tương đồng và mô hình quan hệ)

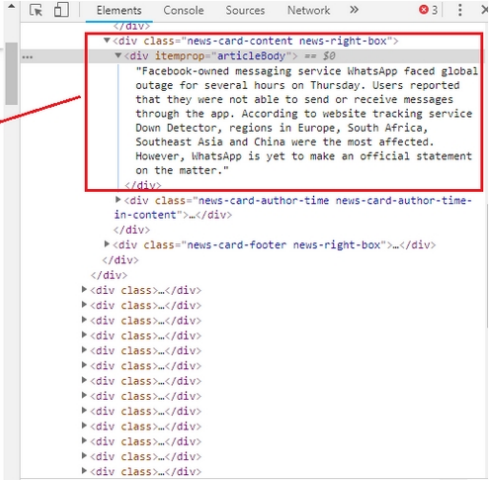
Phân tích văn bản

- Một số ứng dụng phổ biến:
  - Spam detection
  - News articles categorization
  - Social media analysis and monitoring
  - Bio-medical
  - Security intelligence
  - Marketing, ad placements
  - Chatbots
  - Virtual assistants

NLP workflow




inshorts
WhatsApp faces global service outage for several
div | 508 × 128
6:40 pm on 14 Jun 2018,Thursday
Facebook-owned messaging service WhatsApp faced global outage for several hours on Thursday. Users reported that they were not able to send or receive messages through the app. According to website tracking service Down Detector, regions in Europe, South Africa, Southeast Asia and China were the most affected. However, WhatsApp is yet to make an official statement on the matter.
read more at The Quint
Celebrating 1 Million OnePlus 6 Users
OnePlus Community Celebration Season
15 - 26 JUNE
OnePlus 6 global sales cross 1 million units within 22 days
short by Roshan Gupta / 10:00 am on 15 Jun 2018,Friday
OnePlus 6 has crossed 1 million unit sales within 22 days of its launch, becoming the fastest selling OnePlus smartphone since the inception of the company in 2013. OnePlus has revealed that it has a 5 million-strong global community. It has also announced a 'Community Celebration Season' to display gratitude to its community and is running offers from June 15-26.
read more at BGR India


Elements
Console
Sources
Network

 == $0
"Facebook-owned messaging service WhatsApp faced global outage for several hours on Thursday. Users reported that they were not able to send or receive messages through the app. According to website tracking service Down Detector, regions in Europe, South Africa, Southeast Asia and China were the most affected. However, WhatsApp is yet to make an official statement on the matter."

…
…

Cài đặt máy tính cho môn học

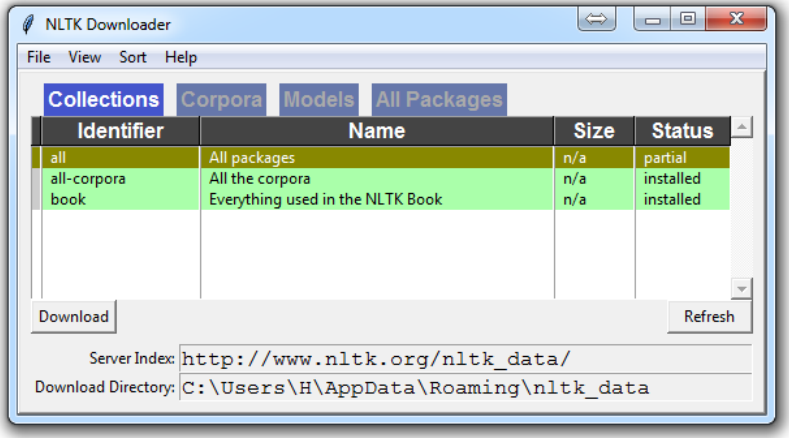
- Ngôn ngữ lập trình Python (version 3.7 trở lên)
- Bộ phân phối Anaconda
  - Đã bao gồm python interpreter
  - Đã bao gồm bộ thư viện máy học scikit-learn

NLTK

- NLTK - Natural language toolkit: thư viện xử lý ngôn ngữ tự nhiên
- `http://www.nltk.org/`
- Sách học NLTK cơ bản: `http://www.nltk.org/book/`
- Cài đặt theo các bước:
  1. Cài đặt bộ thư viện Anaconda
  2. Command prompt: `conda install -c anaconda nltk`
  3. Cài đặt data: `http://www.nltk.org/data.html`

NLTK

- Chọn **all**



NLTK - Example

```
1  from nltk.tokenize import sent_tokenize
2  from nltk.tokenize import word_tokenize
3  from nltk.corpus import wordnet
4
5  if __name__ == '__main__':
6      my_text = "Hello there. Welcome to new semester. How are you? Today
       is a good day to study NLP"
7
8      print(sent_tokenize(my_text))
9      print("-"*20)
10     print(word_tokenize(my_text))
11     print("-"*20)
12
13     syn = wordnet.synsets("NLP")
14     print(syn[0].definition())
```

NLTK - Example

['Hello there.', 'Welcome to new semester.', 'How are you?', 'Today is a good day to study NLP']

['Hello', 'there', '.', 'Welcome', 'to', 'new', 'semester', '.', 'How', 'are', 'you', '?', 'Today', 'is', 'a', 'good', 'day', 'to', 'study', 'NLP']

---

the branch of information science that deals with natural language information

NLTK

- Tài liệu tham khảo:
  - https://www.youtube.com/watch?v=FLZvOKSCkxY&list=PLQVvvaa0QuDf2JswnfiGkliBInZnIC4HL&index=1