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

Gán nhãn từ loại

- PennTreebank
- Hidden Markov model
- Đánh giá

Gán nhãn từ loại: PennTreebank

- Tạo ra bởi University of Pennsylvania
- Dự án 8 năm: 1989 – 1996
- 7 triệu từ
- Tập nhãn dựa trên bộ dữ liệu Brown

Gán nhãn từ loại: PennTreebank

CC	Coordinating conj.	TO	infinitival *to*
CD	Cardinal number	UH	Interjection
DT	Determiner	VB	Verb, base form
EX	Existential there	VBD	Verb, past tense
FW	Foreign word	VBG	Verb, gerund/present pple
IN	Preposition	VBN	Verb, past participle
JJ	Adjective	VBP	Verb, non-3rd ps. sg. present
JJR	Adjective, comparative	VBZ	Verb, 3rd ps. sg. present
JJS	Adjective, superlative	WDT	Wh-determiner
LS	List item marker	WP	Wh-pronoun
MD	Modal	WP$	Possessive *wh*-pronoun
NN	Noun, singular or mass	WRB	Wh-adverb
NNS	Noun, plural	#	Pound sign
NNP	Proper noun, singular	$	Dollar sign
NNPS	Proper noun, plural	.	Sentence-final punctuation
PDT	Predeterminer	,	Comma
POS	Possessive ending	:	Colon, semi-colon
PRP	Personal pronoun	(	Left bracket character
PP$	Possessive pronoun	)	Right bracket character
RB	Adverb	"	Straight double quote
RBR	Adverb, comparative	‘	Left open single quote
RBS	Adverb, superlative	“	Left open double quote
RP	Particle	’	Right close single quote
SYM	Symbol	”	Right close double quote

Gán nhãn từ loại: PennTreebank

- CC
  He bought a car and a house.
- CD
  Five years later, autocar will be popular.
- DT
  Pierre Vinken will join the board.
- EX
  There is no asbestos in our product now.

Gán nhãn từ loại: PennTreebank

- IN
  Mr Vinken is chairman of Elsevier N.V.
- JJ
  Rudolph Agnew was named an executive director.
- JJR
  The number of death was higher than expected

Gán nhãn từ loại: PennTreebank

- JJS
  The percentage of lung cancer appears to be highest.
- MD
  US should regulate the class of asbestos.
- NN
  It's more than three times the expected number.
- NNS
  Portfolio managers expect further declines in interest rates.

Gán nhãn từ loại: PennTreebank

- NNP
  Alexis Sanchez joined Manchester United yesterday.
- NNPS
  … the Japan Automobile Dealers' Association...
- POS
  … at Monday's auction

Gán nhãn từ loại: PennTreebank

- PRP
  It expects to obtain regulatory approval.
- PP$
  Shareholders approve its acquisition by Royal Trustco Ltd.
- RB
  … depends heavily on creativity
- RBR
  … worked for the project for more than six years

Gán nhãn từ loại: PennTreebank

- RBS
  the most mundane aspect of its workers
- TO
  He decided to stay

Gán nhãn từ loại: PennTreebank

- VB
  … to return home
- VBD
  the executives joined Mayor William
- VBG
  … before boarding the buses again
- VBN
  A buffet breakfast was held in the museum

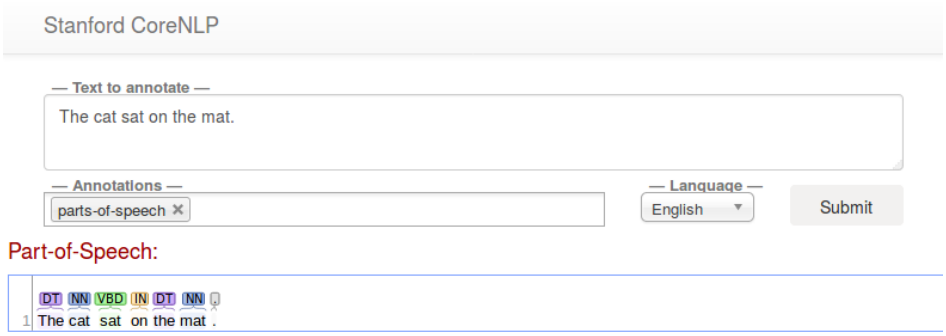
Gán nhãn từ loại: PennTreebank

- VBP
  Plans that give advertisers disscount
- VBZ
  The plan is not an attempt
- WDT
  a project that did not include Seymor
- WP
  who couldn't be reach for comment

Gán nhãn từ loại: PennTreebank

- WRB
  where employees are assigned lunch partners

corenlp.run



http://45.117.171.213/bknlptool/

BK Parser

Please enter your text here:

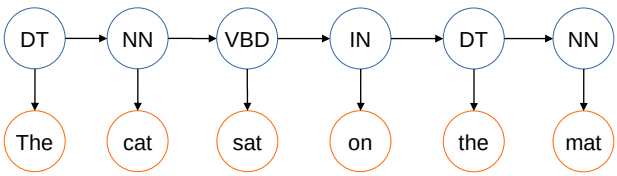
Người hâm mộ reo hò khi đội tuyển U23 đến sân Thống Nhất.

Submit Clear

Part-of-speech

	NN	VB	VB	IN	NN	NNP	VB	NN	NNP	PUNCT
1	Người	hâm_mộ	reo_hò	khi	đội_tuyển	U23	đến	sân	Thống_Nhất	.

Gán nhãn từ loại: Hidden Markov Models



Gán nhãn từ loại: Hidden Markov Models

- Xác suất chuyển đổi
  $Pr(x_t = NN \mid x_{t-1} = DT)$
- Xác suất sinh quan sát
  $Pr(o_t = cat \mid x_t = NN)$

Gán nhãn từ loại: Hidden Markov Models

- Học tham số theo tiêu chí MLE
  $argmax_{\boldsymbol{theta}} \; Pr(\boldsymbol{O}, \boldsymbol{X} \mid \boldsymbol{theta})$
  Thuật toán Baum–Welch
- Giải mã:
  $argmax_{\boldsymbol{X}} \; Pr(\boldsymbol{X} \mid \boldsymbol{theta}, \boldsymbol{O})$
  Thuật toán Viterbi

Gán nhãn từ loại: Thuật toán Baum-Welch

- Bước E
  - Pha tiến

$$\alpha_i(t) = P(o_1 o_2 .. o_{t-1}, s_t = q_i|\lambda).$$

  - Pha lùi

$$\beta_i(t) = P(o_{t+1} o_{t+2} .. o_T, s_t = q_i|\lambda).$$

Gán nhãn từ loại: Thuật toán Baum-Welch

- Bước M

$$\gamma_i(t) = P(X_t = i|Y, \theta) = \frac{P(X_t = i, Y|\theta)}{P(Y|\theta)} = \frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^{N} \alpha_j(t)\beta_j(t)},$$

$$\xi_{ij}(t) = P(X_t = i, X_{t+1} = j|Y, \theta) = \frac{P(X_t = i, X_{t+1} = j, Y|\theta)}{P(Y|\theta)} = \frac{\alpha_i(t)a_{ij}\beta_j(t+1)b_j(y_{t+1})}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i(t)a_{ij}\beta_j(t+1)b_j(y_{t+1})},$$

Gán nhãn từ loại: Giải mã Viterbi

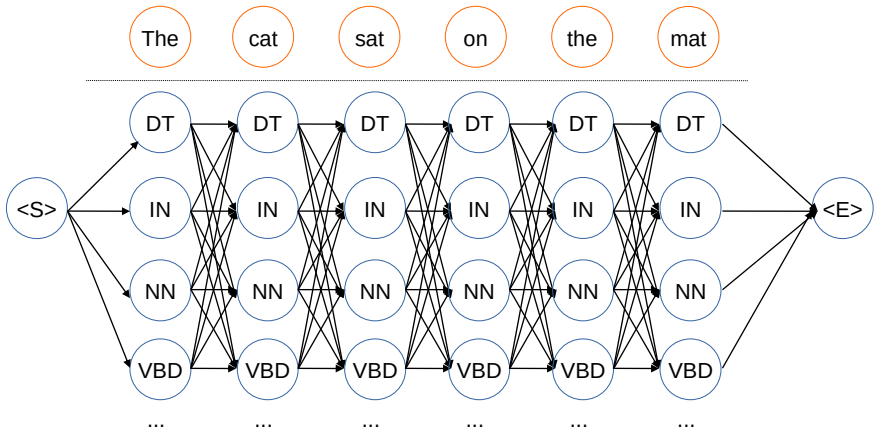
$$Best[i,t] = P(\hat{s_1}\hat{s_2}..\hat{s_{t-1}}, \hat{s_t} = q_i | o_1 o_2 .. o_t, \lambda).$$

$$Best[i,t] = max_j(Best[j,t-1] * a_{j,i} * b_{i,o_t})$$

$$Trace[i,t] = argmax_j(Best[j,t-1] * a_{j,i} * b_{i,o_t})$$

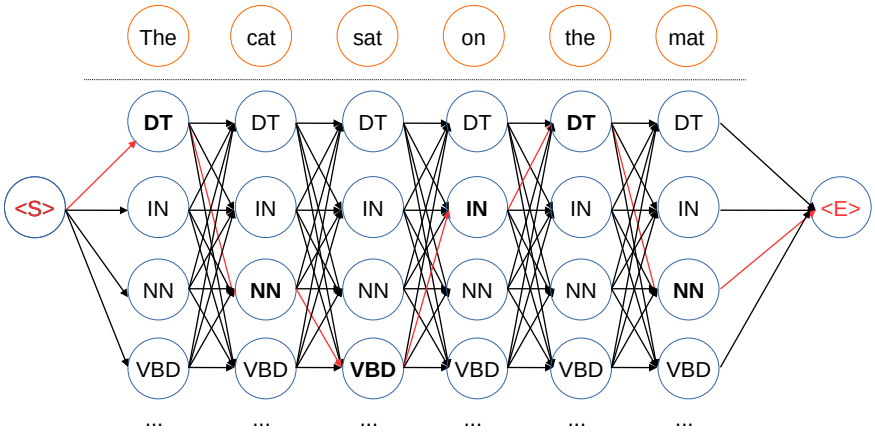
Gán nhãn từ loại: Giải mã Viterbi

*argmax$_X$ P(**X** | **O**, **theta**)*



Gán nhãn từ loại: Giải mã Viterbi

$argmax_X\ P(\boldsymbol{X} \mid \boldsymbol{O}, \boldsymbol{theta})$



Gán nhãn từ loại: Ước lượng tham số

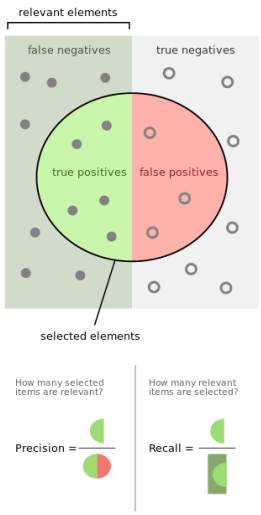
- Xác suất chuyển đổi
  $Pr(x_t=NN|x_{t-1}=DT)$
- Xác suất sinh quan sát
  $Pr(o_t=cat|x_t=NN)$
- Ước lượng tham số
  $Pr(x_t=NN|x_{t-1}=DT)=(count(DT,NN)+1)/(count(DT)+L)$
  $Pr(o_t=cat|x_t=NN)=(count(cat,NN)+1)/(count(NN)+V)$

Gán nhãn từ loại: Đánh giá

- So sánh dự đoán của mô hình với gán nhãn chuẩn
- Các tập dữ liệu:
  - Train: Huấn luyện mô hình
  - Dev: Lựa chọn siêu tham số
  - Test: Đánh giá mô hình

Gán nhãn từ loại: Đánh giá

- Precision
- Recall
- $F_1 = 2PR / (P+R)$



THANK YOU !

29