





## Методи обробки природномовної інформації

Виконав: Горб О. О.

перевірив: Миколайчук Р.А.

A decorative graphic in the top-left corner consisting of two overlapping parallelograms. The front one is blue and the back one is light green. They are set against a dark blue background with faint, larger-scale geometric patterns.


Використовуючи метод  
умовне випадкове поле ,  
розпізнати типи тегів у  
словах.



Для даної лабораторної роботи був  
вибраний набір даних з 47959  
речень, 104857 слів з  
маркуваннями.

Набір даних має такий вигляд:

Sentence #	Word	POS	Tag
Sentence: 1	Thousands	<u>NNS</u>	O
	of	<u>IN</u>	O
	demonstrators	<u>NNS</u>	O
	have	<u>VBP</u>	O
	marched	<u>VBN</u>	O
	through	<u>IN</u>	O
	London	<u>NNP</u>	B-geo
	to	<u>TO</u>	O
	protest	<u>VB</u>	O
	the	<u>DT</u>	O
	war	<u>NN</u>	O
	in	<u>IN</u>	O
	Iraq	<u>NNP</u>	B-geo
	and	<u>CC</u>	O
	demand	<u>VB</u>	O
	the	<u>DT</u>	O
	withdrawal	<u>NN</u>	O
	of	<u>IN</u>	O
	British	<u>JJ</u>	B-gpe
	troops	<u>NNS</u>	O
Sentence: 2	from	<u>IN</u>	O
	that	<u>DT</u>	O
	country	<u>NN</u>	O
	.	<u>.</u>	O
	Families	<u>NNS</u>	O
	of	<u>IN</u>	O

- 
- Кількість унікальних слів у наборі даних — 35178
  - Кількість тегів — 17

Завдання – використовуючи метод умовне випадкове поле (далі CRF), розпізнавати типи тегів у словах.

	Sentence #	Word	POS	Tag
<b>count</b>	47959	1048575	1048575	1048575
<b>unique</b>	47959	35178	42	17
<b>top</b>	Sentence: 9309	the	NN	O
<b>freq</b>	1	52573	145807	887908



## Інформація про теги:

- geo = географічна сутність
- org = організація
- per = особа
- gre = геополітична сутність
- tim = індикатор часу
- art = артефакт
- eve = подія
- nat = природний феномен



# Опис підходу

Спочатку, у "Sentence #" є багато відсутніх значень. Отже, будемо використовувати техніку pandas fillna та використовувати метод ffill, який розповсюджує останнє дійсне значення вперед.

Далі, створимо клас для отримання одного речення з набору. Кожне речення буде списком кортежів із мітками та частинами мови.

Зробимо декілька функцій word2features, sent2features, sent2labels для підготовки ознак. Ці функції за замовчуванням використовуються NER в nltk з невеликими змінами для нашого набору.

Розділимо набір даних на train (80%) та test (20%) та реалізуємо "Умовне випадкове поле" за допомогою sklearn. Параметри для CRM: тренувальний алгоритм = градієнтний спуск за методом L-BFGS; коефіцієнт регуляризації L1 = 0.1; коефіцієнт регуляризації L2 = 0.1; усі можливі переходи = False (чи генерує CRF функції переходу, які навіть не зустрічаються в навчальних даних); макс. кількість ітерацій = 100

# Реалізація

Для методу Conditional random field була навчена модель. Тестування показало результат точності середнього зваженого 97%. Результати виглядають наступним чином:

	precision	recall	f1-score	support
B-art	0.39	0.15	0.22	79
B-eve	0.54	0.42	0.47	65
B-geo	0.87	0.90	0.89	7610
B-gpe	0.97	0.94	0.95	3165
B-nat	0.65	0.37	0.47	41
B-org	0.79	0.75	0.77	3960
B-per	0.85	0.82	0.84	3387
B-tim	0.93	0.88	0.91	4080
I-art	0.31	0.16	0.22	61
I-eve	0.42	0.27	0.33	56
I-geo	0.83	0.79	0.81	1531
I-gpe	0.85	0.55	0.67	42
I-nat	0.50	0.38	0.43	8
I-org	0.82	0.80	0.81	3304
I-per	0.86	0.89	0.87	3410
I-tim	0.83	0.77	0.80	1307
0	0.99	0.99	0.99	176929
accuracy			0.97	209035
macro avg	0.73	0.64	0.67	209035
weighted avg	0.97	0.97	0.97	209035

# Як могли бачити на реалізації.

Як можна побачити результат роботи моделі доволі точний. Найменша точність моделі є при розпізнаванні артефактів (art) із-за малої кількості цих даних в наборі.

	precision	recall	f1-score	support
B-art	0.39	0.15	0.22	79
B-eve	0.54	0.42	0.47	65
B-geo	0.87	0.90	0.89	7610
B-gpe	0.97	0.94	0.95	3165
B-nat	0.65	0.37	0.47	41
B-org	0.79	0.75	0.77	3960
B-per	0.85	0.82	0.84	3387
B-tim	0.93	0.88	0.91	4080
I-art	0.31	0.16	0.22	61
I-eve	0.42	0.27	0.33	56
I-geo	0.83	0.79	0.81	1531
I-gpe	0.85	0.55	0.67	42
I-nat	0.50	0.38	0.43	8
I-org	0.82	0.80	0.81	3304
I-per	0.86	0.89	0.87	3410
I-tim	0.83	0.77	0.80	1307
0	0.99	0.99	0.99	176929
accuracy			0.97	209035
macro avg	0.73	0.64	0.67	209035
weighted avg	0.97	0.97	0.97	209035





# Висновки

Отже, під час розбору умовних випадкових полів було досліджено, що алгоритм не вимагає припущення незалежності спостережуваних змінних. Використання довільних факторів дозволяє описати різні ознаки визначених об'єктів, що знижує вимоги до повноти і обсягу навчальної вибірки. Проте, по результатам, видно, що при зменшенні кількості екземплярів, втрачається точність.

Також, реалізація алгоритму CRF має хорошу швидкість, що може допомогти при обробці великих обсягів інформації. На тренування приблизно 800,000 слів та їх міток було витрачено 4 хвилини 52 секунди.



Дякую за увагу !