

Wytyczne ręcznej anonimizacji

Marcin Oleksy

Stosowane metody anonimizacji tekstów (podsumowanie przeglądu literatury)	2
Budowa zanonimizowanego korpusu	3
Przygotowanie korpusu	3
Zastosowane metody (3 wersje korpusu)	4
Usuwanie (1)	4
Tagowanie (2)	4
Pseudonimizacja (3)	5
Referencje	6

Stosowane metody anonimizacji tekstów (podsumowanie przeglądu literatury)

	usuwanie	tagowanie			pseudonimizacja	
		zamiana na umowny ciąg znaków	uogólnienie informacji	tag odnoszący się do kategorii oryginalnej nazwy	zamiana na inny rzeczywisty znak	zamiana na jednostkę fikcyjną
Beckwith 2006		xxxxxxxxxxxxx xxx				
Berman 2003		***				
Morrison 2009	+					
Friedlin, McDonald 2008		<xxx>		\$YearMonthDay\$, \$patient id\$, \$patient name\$...		
Gupta 2004			zamiast usuwanych dat wstawiany przedział czasowy	NAME, SLIDE-NUMBER, PLACE, ADDRESS		
Neamatullah 2008			np. [**Age over 89**]	[**Location**], [**Name**] (gdy to możliwe uszczegółowione: (np. "last name, first name"...))	data zamieniana na inną o takim samym formacie	
Ruch 2000				każdy znak zamieniany na "x" bądź "X" w zależności od wielkości oryginalnego znaku		
Sweeney 1996					data zmieniana na inną o takim samym formacie	fikcyjne, ale sensownie brzmiące imiona
Thomas	+					

2002						
Aramaki 2006				B-HOSPITAL, I-HOSPITAL		
Taira 2002	+					
Uzuner 2008		[REMOVED]				

Pełny raport można znaleźć pod tym adresem:

https://docs.google.com/document/d/1ftbzLDWLYB9j52d6o0JENBK-EGl6_E9Gekhylz_MHjs/e/dit?disco=AAAADPbJfys&ts=5d42e672&usp=comment_email_document&usp_dm=false

Budowa zanonimizowanego korpusu

Przygotowanie korpusu

Strona techniczna:

- anonimizację wspiera system Inforex
- fragmenty tekstu nadające się do anonimizacji zostały oznaczone jedną z kategorii:
 - MAIL
 - MIEJSCE
 - OSOBA
 - USER
 - WWW
 - INNE
- proces anonimizacji polega na wpisaniu odpowiedniej formy w pole "Lemma" w atrybutach danej anotacji (wersja korpusu 2. i 3.)

W ramach preprocessingu korpus musi zostać automatycznie oznakowany w ramach nazw własnych za pomocą narzędzia Liner2.

Zastosowane metody (3 wersje korpusu)

Usuwanie (1)

Wyselekcjonowane ręcznie tokeny (oznaczone w procesie przygotowania korpusu) zostają usunięte z korpusu.

Tagowanie (2)

Wyselekcjonowane tokeny zostają zamienione na formy brzmiące jak tagi przypisane podczas przygotowania korpusu, a więc:

- MAIL
- MIEJSCE
- OSOBA
- USER
- WWW
- INNE

Na przykład:

@[kamil_rychlicki]_{USER} Widzę od razu oburzenie jakbym już powiedział
→ *@[USER] Widzę od razu oburzenie jakbym już powiedział*

Proces przeprowadzony jest w sposób automatyczny. Jeśli jedną anotacją danej kategorii oznaczone są sąsiednie tokeny (np. w przypadku wielowyrazowych nazw własnych), każdy z nich zostaje automatycznie zamieniony na tag danej kategorii, np.

[Jan Kowski]_{OSOBA}
→ *[OSOBA] [OSOBA]*.

Wskazówki:

Zakres anotacji ręcznej: jedną anotację stanowi dla nas całe wyrażenie, które możemy przypisać do jednej kategorii i odnoszące się jako całość do jednego bytu i którego wszystkie elementy składowe nazywają dany byt, np.

[Bartosz Porczyk], a nie *[Bartosz] *[Porczyk]
ale

[Chestnuts Park] w [Haringey], a nie *[Chestnuts Park w Haringey] (przyimek "w" nie nazywa obiektu)

MIEJSCE

1. W przypadku rozbudowanych danych adresowych anotujemy całość, np. [Warszawa ul. Stawki 2A] (nie zaś samo miasto)
2. Anonimizacji podlegają nie tylko nazwy o charakterze geograficznym (np. nazwy miast czy wsi), ale także nazwy budynków

OSOBA

1. Anotujemy całość odniesienia do danej osoby, nawet jeśli jest to tylko imię bądź imię z inicjałem
2. Anotujemy odniesienia do osób fikcyjnych

3. Nie anotujemy nazw osobowych będących częścią innej nazwy własnej, np. Uniwersytet Medyczny im. *[Ludwika Rydygiera]

INNE

1. Nazw firm nie traktujemy jako dane wrażliwe

Pseudonimizacja (3)

Pseudonimizacja polega na zastąpieniu wskazanego fragmentu tekstu wyrażeniem, które:

1. należy do tej samej kategorii nazw własnych / jednostek identyfikacyjnych, np.
 - o imię ← imię
2. należy do tej samej klasy gramatycznej, np.
 - o subst ← subst
3. ma taką samą formę fleksyjną, np.
 - o sg:acc:m3 ← sg:acc:m3
4. ma taką samą końcówkę (?), np.
 - o Ankę ← Wiołę
5. ma taką samą liczbę segmentów, np.
 - o Bożydar Iwanow ← Jarosław Nowikow
6. zachowuje schemat zapisu dużymi/małymi literami (?), np.
 - o mBanku ← sTanku
7. zachowuje liczbę cyfr, np.
 - o 92650 ← 81769

Zasady dotyczące pseudonimizacji poszczególnych kategorii:

- **MAIL**
Człon sprzed znaku "@" zostaje zamieniony na losowy ciąg znaków przy zachowaniu ich liczby. Jeśli człon sprzed "@" składa się z kilku tokenów każdy z nich (o ile nie jest znakiem interpunkcyjnym) zostaje zamieniony według powyższej zasady. Podobnie z domeną drugiego poziomu. Domena najwyższego poziomu oraz znaki interpunkcyjne pozostają bez zmian.
- **MIEJSCE**
Token zostaje zamieniony na inną nazwę własną należącą do tej samej kategorii, do której należy nazwa źródłowa. Kategorie nazw źródłowych zostają przypisane przez narzędzie Liner2, np.
 $[Wrocław]_{MIEJSCE}$
 $\rightarrow [Wrocław]_{nam_loc_gpe_city} (Liner2)$
 $\rightarrow [Toruń]_{nam_loc_gpe_city} (wiktionary-forms)$
W celu zachowania formy gramatycznej do procesu podmiany zostanie wykorzystana baza **wiktionary-forms-with-bases-and-tags.txt** z **NELexicon2** (w miejsce podmienianego tokenu zostanie użyty inny należący do tej samej kategorii i o tej samej interpretacji gramatycznej).
- **OSOBA**
Token zostaje zamieniony na inną nazwę należącą do tej samej kategorii, do której należy nazwa źródłowa. Kategorie nazw źródłowych zostają przypisane przez narzędzie Liner2, np.

[Jan Kowalski]_{OSOBA}

→ [Jan]_{nam_liv_person_first} [Kowalski]_{nam_liv_person_last} (Liner2)

→ [Marek]_{nam_liv_person_first} [Czuchnowski]_{nam_liv_person_last} (wiktionary-forms)

W celu zachowania formy gramatycznej do procesu podmiany zostanie wykorzystana baza **wiktionary-forms-with-bases-and-tags.txt** z **NELexicon2** (w miejsce podmienianego tokenu zostanie użyty inny należący do tej samej kategorii i o tej samej interpretacji gramatycznej).

- **USER**

Pseudonimizowany według tej samej zasady, co człon (członcy) sprzed znaku "@" w adresach mailowych. Znak "@" pozostaje.

- **WWW**

Znaki w domenach zostają zamieniane według zasady: litera za literę (zachowana wielkość), cyfra za cyfrę. Nazwa hosta, domena najwyższego poziomu oraz znaki interpunkcyjne pozostają bez zmian.

- **INNE**

Token zostaje zamieniony na inną nazwę należącą do tej samej kategorii, do której należy nazwa źródłowa.

W celu zachowania formy gramatycznej do procesu podmiany zostanie wykorzystana baza wiktionary-forms-with-bases-and-tags.txt z NELexicon2 (w miejsce podmienianego tokenu zostanie użyty inny o tej samej interpretacji gramatycznej).

Referencje

1. <https://ocean.sagepub.com/blog/making-sensitive-text-data-accessible-for-computational-social-science>
2. <http://www.ep.liu.se/ecp/152/006/ecp18152006.pdf>
3. <https://onlinelibrary.wiley.com/doi/full/10.1111/jedm.12093>
4. <https://piwik.pro/blog/the-ultimate-guide-to-data-anonymization-in-analytics/>