Розроблення інформаційно-обчислювального модулю категоризації неструктурованих даних

студент групи КА-41м
Дозірців В.О.
науковий керівник :
професор Панкратова Н.Д.

This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Актуальність дослідження

- Величезні обсяги неструктурованої інформації, що постійно зростають;
- Розвиток сховищ даних;
- Ефективні підходи до обробки тексту;
- Автоматизація категоризації;
- Аналіз емоційного забарвлення.

This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Сфери використання Sentiment Analysis

- CRM - системи;
- Обробка рецензій на товари і послуги;
- Рекомендаційні системи;
- Державна розвідка;
- Передбачення настроїв народу;
- Аналіз документів політиків та влади.

This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Об'єктом дослідження є великі об’єми текстових даних різної природи, а саме :

- База даних МВС;
- Національний класифікатор України : класифікатор надзвичайних ситуацій.

Предметом дослідження являється розробка моделей, методів, правил категоризації неструктурованих даних.

This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Задачі дослідження

- Приведення вхідних даних до вигляду, приємним для машинної обробки;
- Контроль і покращення якості приведення вхідних даних;
- Видобуток українського ДК 019:2010, аналіз та його первинна обробка;
- Створення правил категоризації;
- Аналіз отриманих результатів.

This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Основне означення

Думка – це п’ятірка (e_i, a_{ij}, s_{ijkl}, h_k, t_l), де e_i – це назва суб’єкту, a_{ij} - це аспект суб’єкту e_i, це емоційне забарвлення аспекту a_{ij} суб’єкту e_i, h_k — власник думки і t_l – це час, коли думка була виражена h_k. Емоційне забарвлення s_{ijkl} може бути позитивним, негативним чи нейтральним, або виражатись з різноманітним рівнем сили чи інтенсивності, наприклад по шкалі від 1 до 5 зірок найчастіше оцінюють відгуки на сайтах в Інтернеті.

Категорія аспекту суб'єкта являє собою унікальний аспект суб'єкта, в той час як вираження аспекту - це актуальною слово або фраза, яка з'являється в тексті, і вказує на категорію аспекту

This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Класифікація настроїв

- Основна мета : з огляду на документ d, виявити всі п'ятірки $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$ в d.
- Основне припущення : класифікація настроїв припускає, що думка документу d (наприклад, відкликання продукції) виражає думку з єдиним суб'єктом e і містить думки від одного власника думки h.

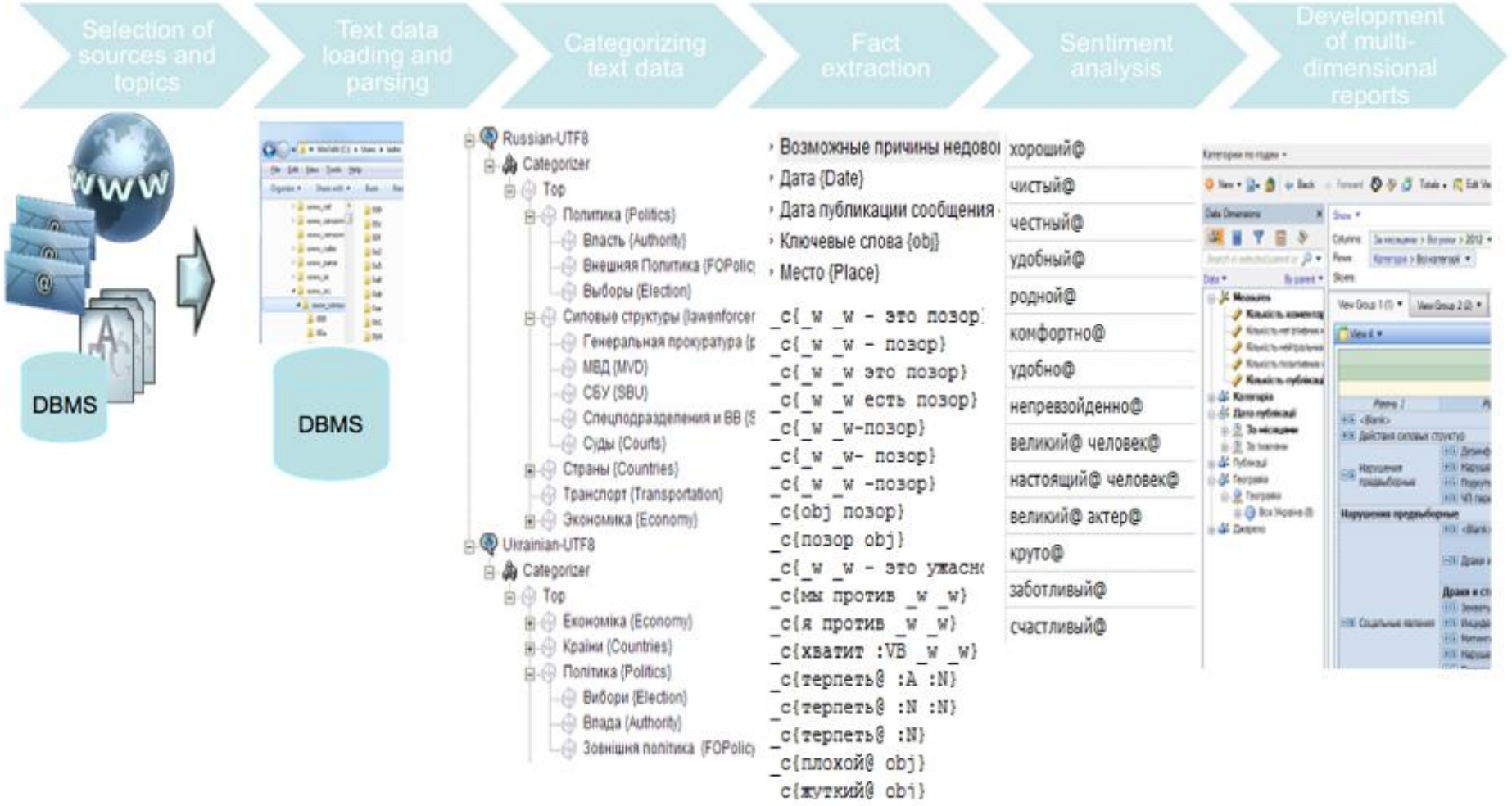
This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Завдання в класифікація настроїв

- витяг виражень і категоризація;
- витяг аспектів і категоризація;
- витяг власників думок і категоризація;
- витяг часу і стандартизація;
- класифікація аспектів настрою;
- генерація «п’ятірки» думки.

This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Процес роботи з неструктурованими даними



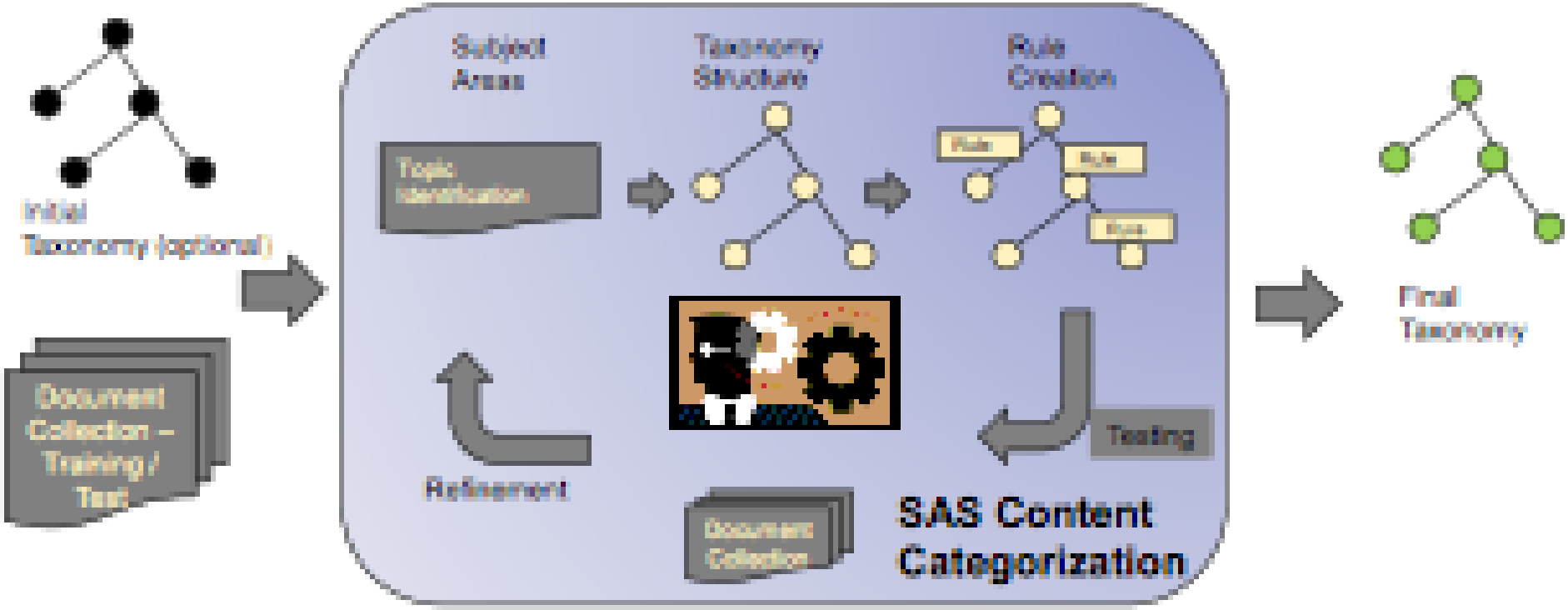
This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

ТАКСОНОМІЯ

Термін «таксономія» вперше був запропонований в 1813 році Огюстеном Декандолем, який займався класифікацією рослин, і на початку застосовувався тільки у біології. Пізніше цей термін став використовуватись для позначення загальної теорії класифікації та систематизації складних систем як в біології, так і в інших галузях науки.

This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Написання таксономії



This research was sponsored in part by the
NATO Science for Peace and Security
Programme under grant G4877

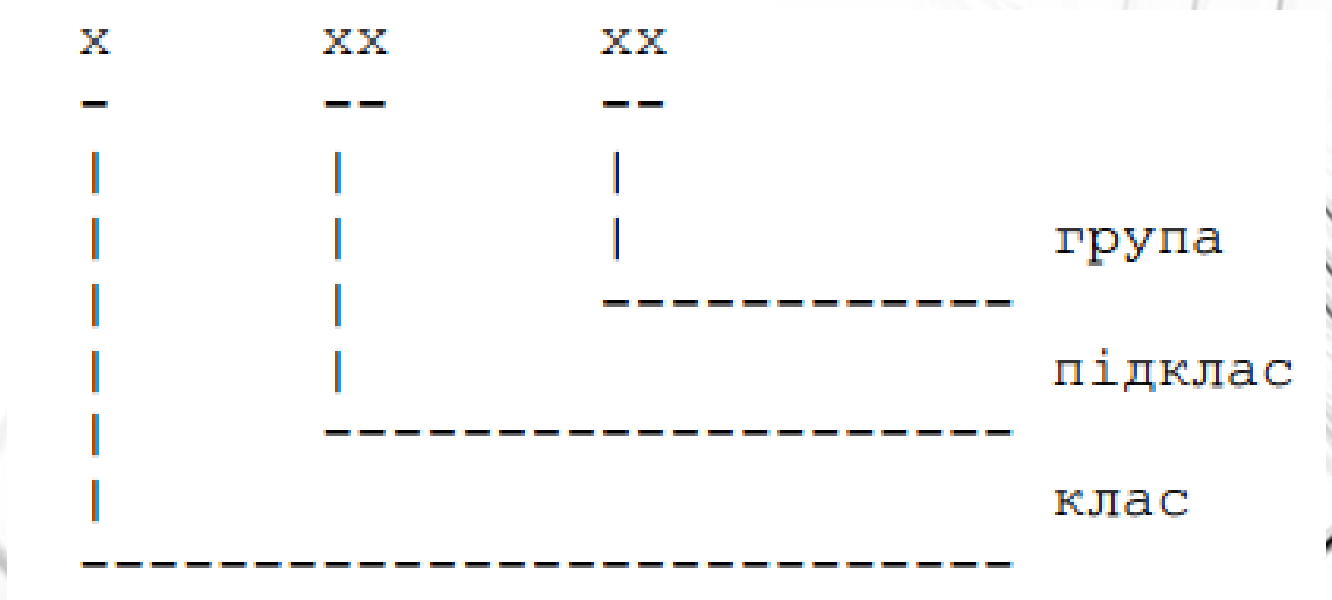
Вхідні дані - ДК 019:2010

- За структурою :
  - Клас;
  - Підклас;
  - Група.
- Методи класифікації:
  - Ієрархічний;
  - Послідовний;
  - П'ятизначний.

This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Структура класифікатора

Структура коду класифікатора має 5 розрядів і відповідає такій схемі:



This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Приклад ієрархії з ДК

- 10000 - НС ТЕХНОГЕННОГО ХАРАКТЕРУ
- 10100 НС УНАСЛІДОК АВАРІЙ ЧИ КАТАСТРОФ НА ТРАНСПОРТІ (за винятком пожеж і вибухів)
- 10110 - НС унаслідок аварії на транспорті з викиданням (загрозою викидання) небезпечних і шкідливих (забруднювальних) речовин
- 10111 - НС унаслідок аварії на транспорті з викиданням (загрозою викидання)

This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Правила побудови підкатегорій

- Підкатегорія завжди успадковує властивості категорії.
- Підкатегорії повинні містити якусь вузьку обмежену інформацію, що не виходить за рамки категорії.
- Підкатегорії завжди повинні бути паралельні між собою.
- Не слід допускати цикл категорій або протиріч.
- Підкатегорії не повинні бути синонімічні.
- Підкатегорій повинно бути значно менше, ніж понять, що наповнюють.

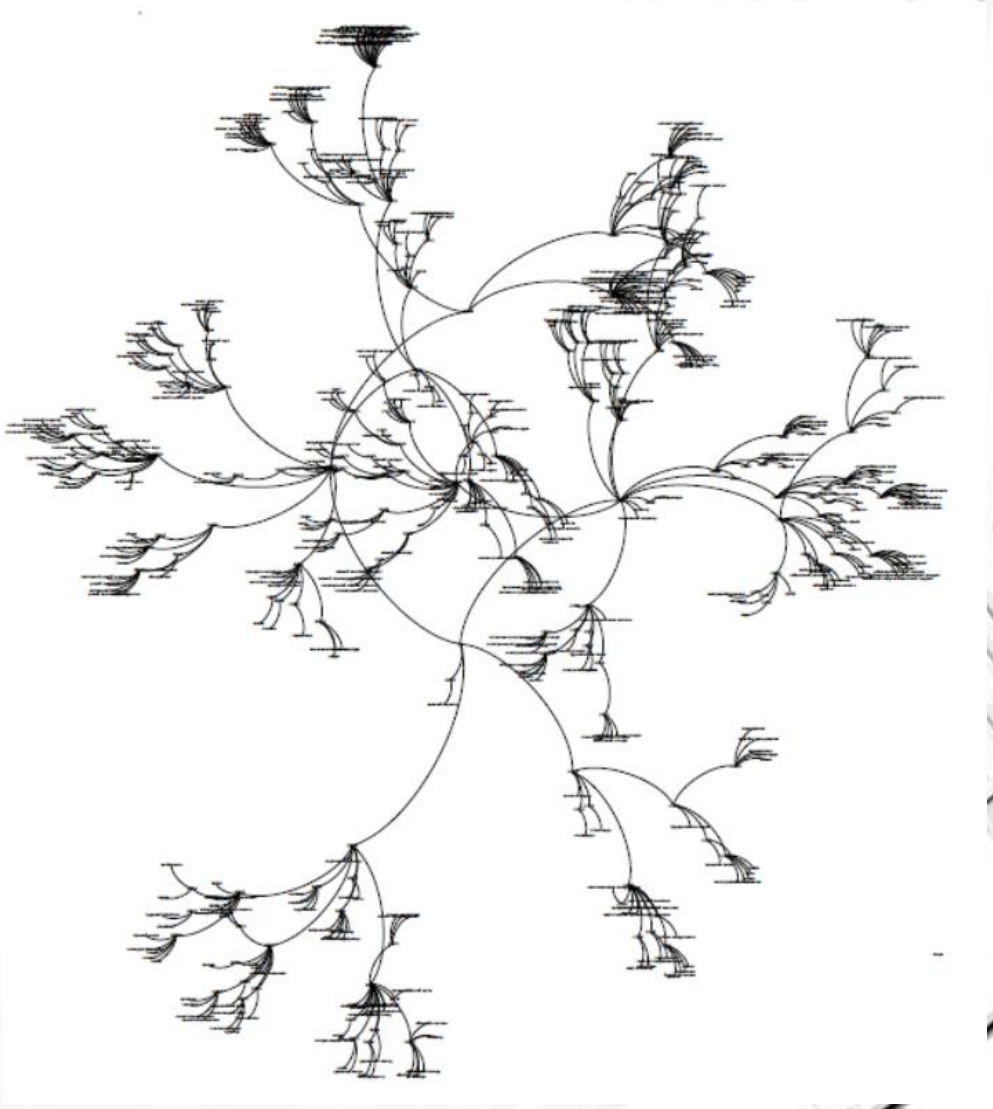
This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Розгортання категорій

	A	B	C	D	E	F	G	H	I
1	Код	Назва	l	id	parid	node	delete_	delete_word	delete_word_id
1564	Top/20000/20700/20710/20711	пов'язана з екзотичним	0	c3c059480	Top/2000(	Top/20000/20700/20710/20711	0		0
1565	Top/20000/20700/20710/20711	пов'язана з екзотичним	1	c3c059480	c3c059480	PAR0	0		0
1566	Top/20000/20700/20710/20711	пов'язана з екзотичним	2	e0f042738	c3c059480	окремі випадки	0	людей	1
1567	Top/20000/20700/20710/20711	пов'язана з екзотичним	4	0e35cc02d	57c65d24e	випадок	0	захворювання	1
1568	Top/20000/20700/20710/20711	пов'язана з екзотичним	5	2209dc095	b7b5c72e:	окремий випадок	0	інфекційне захво	1
1569	Top/20000/20700/20710/20711	пов'язана з екзотичним	2	9138281d(	c3c059480	небезпечним інфекційним захворюванням люде	0	захворювання ли	1
1570	Top/20000/20700/20710/20711	пов'язана з екзотичним	2	86639fc62	9138281d(	небезпечним інфекційним захворюванням	0	інфекційне захво	1
1571	Top/20000/20700/20710/20711	пов'язана з екзотичним	5	a4aa2e0b4	86639fc62	небезпечне захворювання	0	людей	1
1572	Top/20000/20700/20710/20711	пов'язана з екзотичним	5	810c5d788	86639fc62	інфекційне небезпечне захворювання	0	людина	1
1573	Top/20000/20700/20710/20711	пов'язана з екзотичним	5	313e86cfe	86639fc62	небезпечне інфекційне захворювання	0		0
1574	Top/20000/20700/20710/20711	пов'язана з екзотичним	3	64b01b74(	9138281d(	небезпечне захворювання людей	0		0
1575	Top/20000/20700/20710/20711	пов'язана з екзотичним	3	f21ee00f5	9138281d(	інфекційне небезпечне захворювання людей	0		0
1576	Top/20000/20700/20710/20711	пов'язана з екзотичним	3	bc7c0b22b	9138281d(	небезпечне інфекційне захворювання людей	0		0
1577	Top/20000/20700/20710/20711	пов'язана з екзотичним	99	000004526	c3c059480	надзвичайна ситуація	0		0
1578	Top/20000/20700/20710/20711	пов'язана з екзотичним	99	be27593b4	c3c059480	захворювання	0		0
1579	Top/20000/20700/20710/20711	пов'язана з екзотичним	99	6ff83abec	c3c059480	людина	0		0

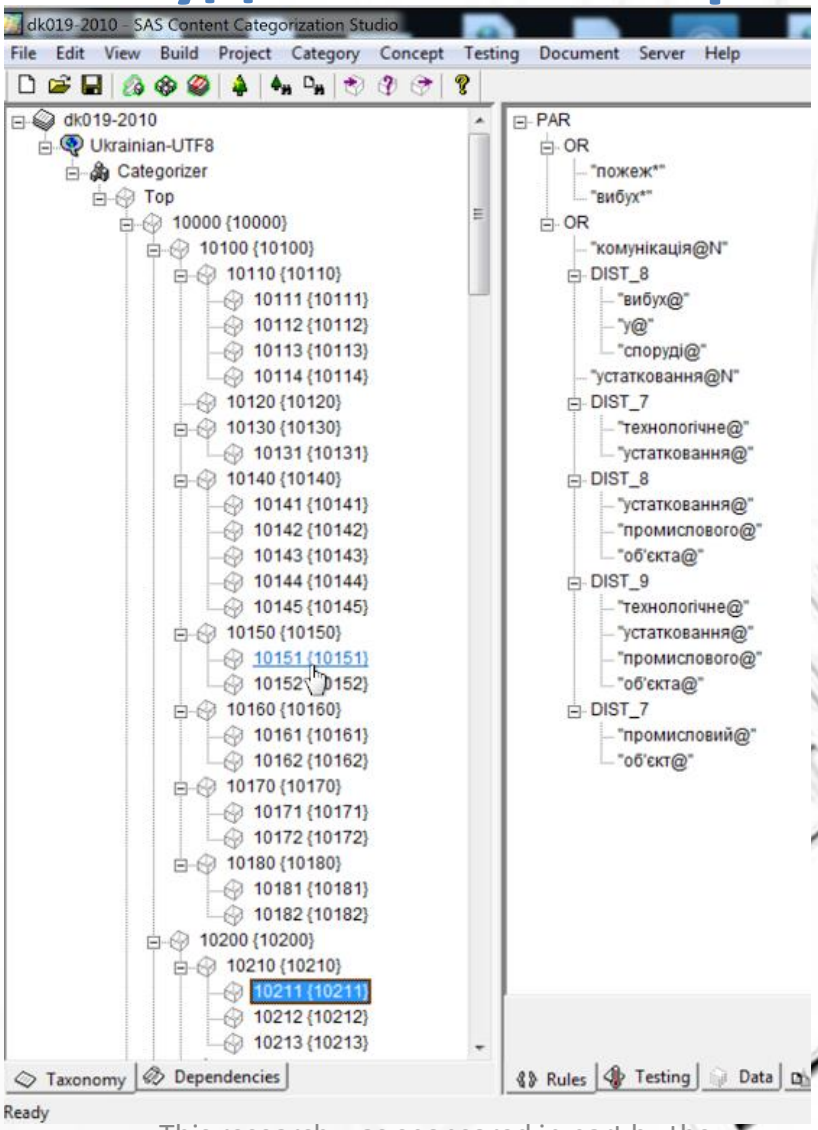
This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Схема категоризатора



This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Результати дослідження. Побудовані категорії

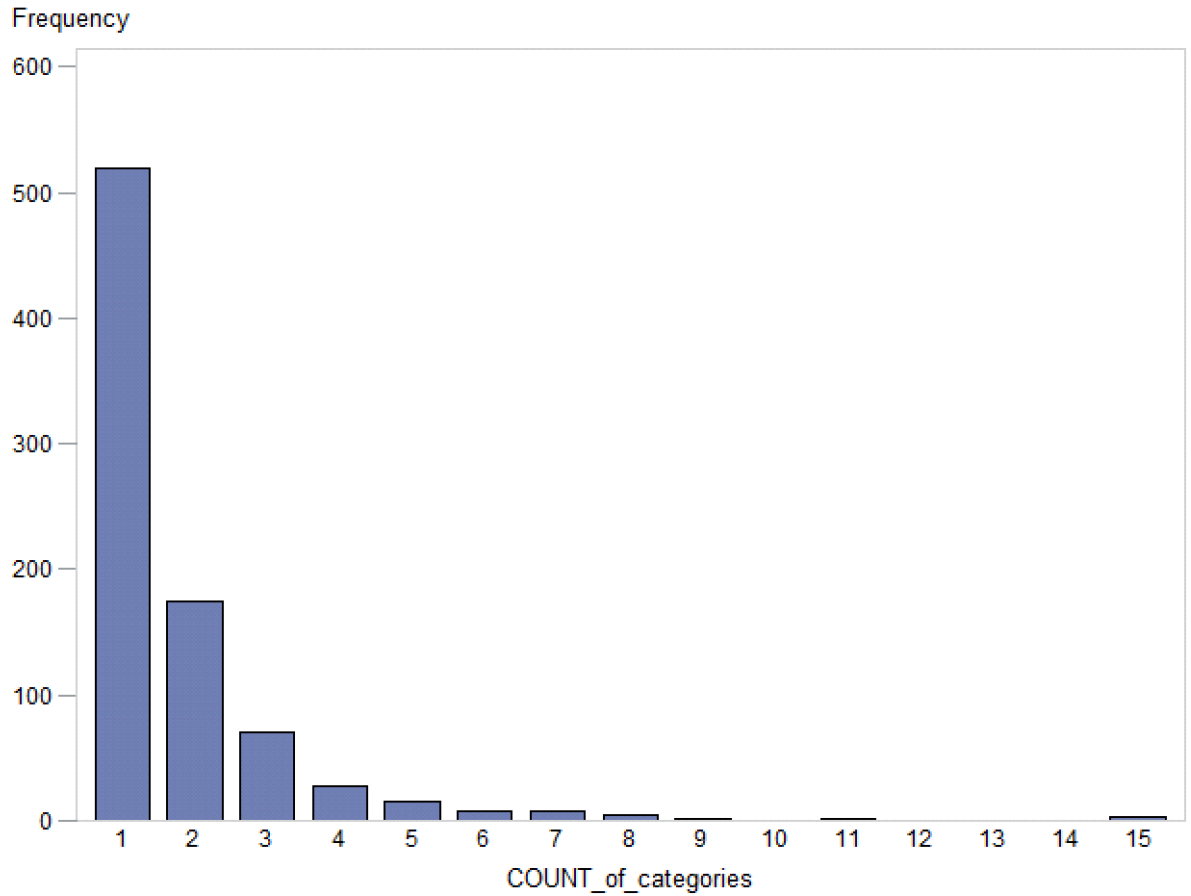


This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Результати дослідження. Побудова категоризатора

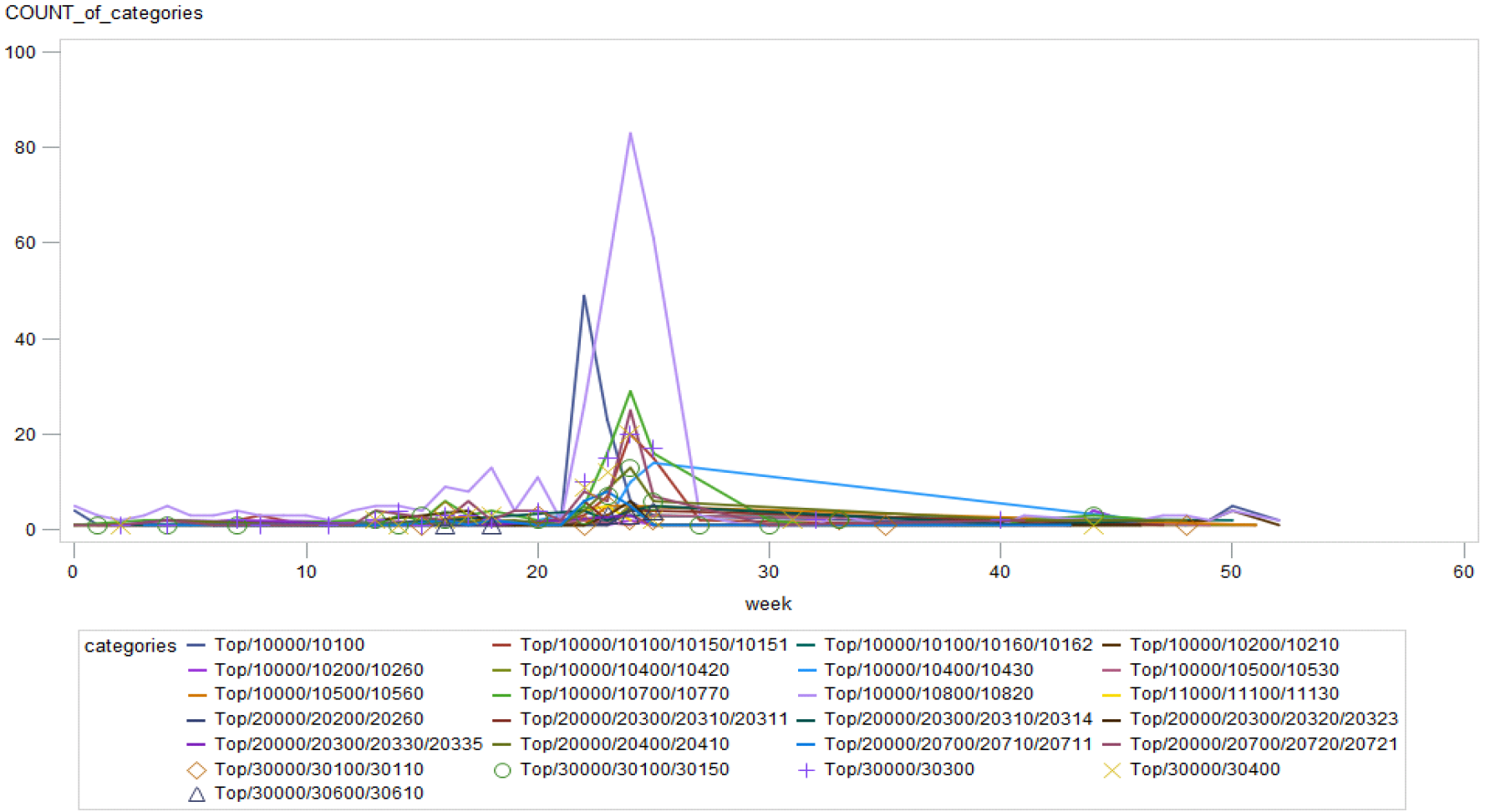
This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Результати дослідження. Статистичні дані



This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Результати дослідження. Аналіз даних по часу : ТОП25 категорій



This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Висновки

- Проаналізовано ДК 019:2010 та формалізовано правила побудови категорій для даного класифікатора;
- Запропонований підхід до синтезу правил побудови категорій через автомат;
- Побудовано категоризатор та прокатегоризовано 83 відсотки документів.

This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Подальші дослідження

- Аналіз інших категоризаторів;
- Узагальнення правил створення категорій для української мови;
- Пришвидшення методів визначення емоційного відтінку тексту;
- Автоматизація приведення вхідних даних до приємного виду для аналізу.

This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Перелік публікацій

1. Pankratova N.D., Dozirtsiv V.O. Application of methods for text analysis of the emotional tone to identify social disasters./ Pankratova Natalia, Dozirtsiv Volodymyr // System Analysis and Information Technology — 2016. — vol. 18. — P. 38

This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877

Дякую за увагу

This research was sponsored in part by the NATO Science for Peace and Security Programme under grant G4877