

# Розроблення інформаційно- обчислювального модулю категоризації неструктурованих даних

студент групи КА-41м

Дозірців В.О.

науковий керівник :

професор Панкратова Н.Д.

# Актуальність дослідження

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

# Сфери використання Sentiment Analysis

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

**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.

# 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.

# 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

# 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  $h$ .

# Завдання в класифікація настроїв

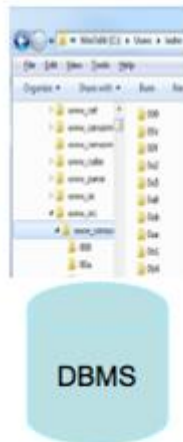
- витяг виражень і категоризація;
- витяг аспектів і категоризація;
- витяг власників думок і категоризація;
- витяг часу і стандартизація;
- класифікація аспектів настрою;
- генерація «п'ятірки» думки.



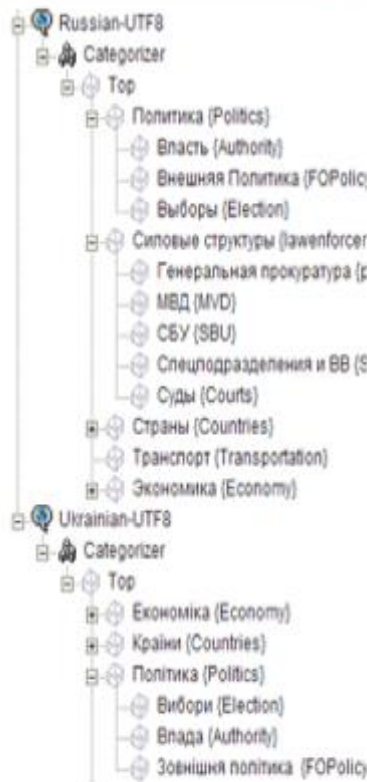
# Процес роботи з неструктурованими даними



DBMS



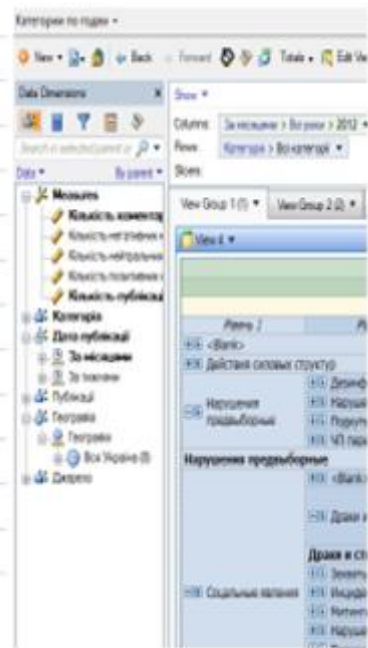
DBMS



- › Возможные причины недово
- › Дата (Date)
- › Дата публикации сообщения
- › Ключевые слова {obj}
- › Место (Place)

с{ w w - это позор:  
с{ w w - позор)  
с{ w w это позор)  
с{ w w есть позор)  
с{ w w-позор)  
с{ w w- позор)  
с{ w w -позор)  
с{obj позор)  
с{позор obj)  
с{мы против w w)  
с{я против w w)  
с{хватит :VB w w)  
с{терпеть@ :A :N)  
с{терпеть@ :N :N)  
с{терпеть@ :N)  
с{плохой@ obj)  
с{жуткий@ obj)

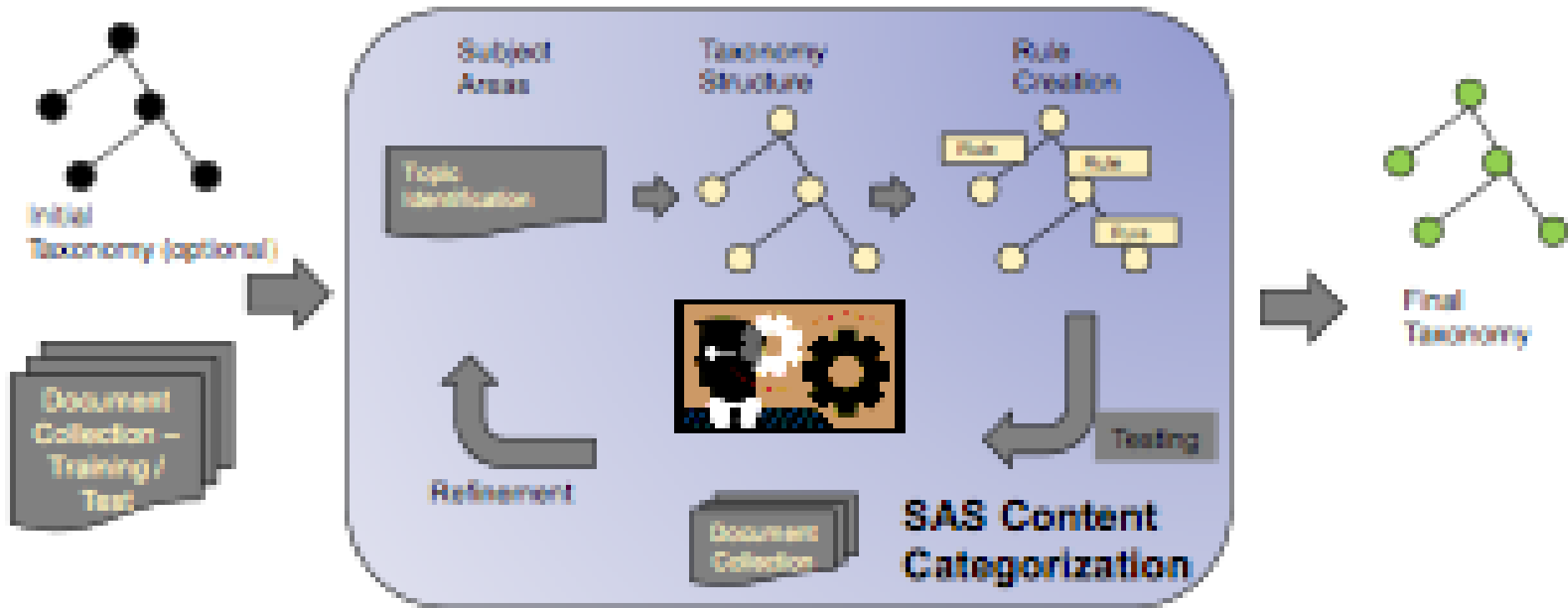
хороший@  
чистый@  
честный@  
удобный@  
родной@  
комфортно@  
удобно@  
непревзойденно@  
великий@ человек@  
настоящий@ человек@  
великий@ актер@  
круто@  
заботливый@  
счастливый@



# ТАКСОНОМІЯ

Термін «таксономія» вперше був запропонований в 1813 році Огюстеном Декандолом, 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.

# Написання таксономії

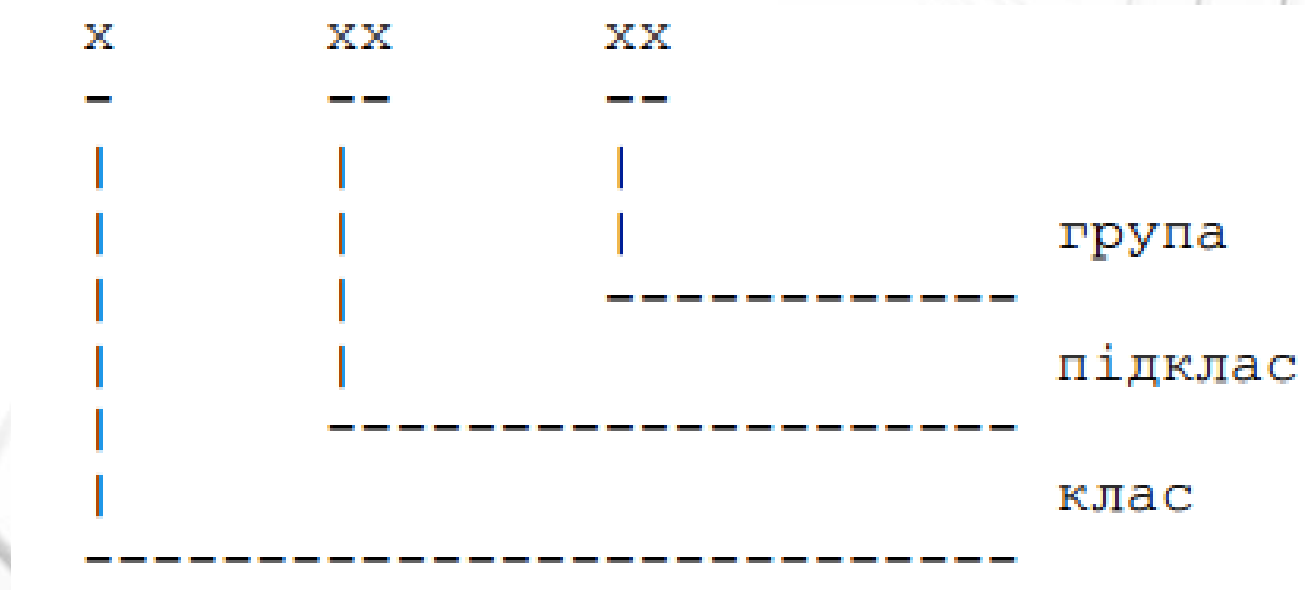


# Вхідні дані - ДК 019:2010

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

# Структура класифікатора

Структура коду класифікатора має 5 розрядів і відповідає такій схемі:



# 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)

# 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.

# Розгортання категорій

|      | A                           | B                      | C  | D         | E  | F | G      | H               | I              |
|------|-----------------------------|------------------------|----|-----------|--|---|--------|-----------------|----------------|
| 1    | Код                         | Назва                  | id | parid     | node   |   | delete | delete_word     | delete_word_id |
| 1564 | Top/20000/20700/20710/20711 | пов'язана з екзотичним | 0  | c3c059480 | Top/20000/20700/20710/20711                          |   | 0      |                 | 0              |
| 1565 | Top/20000/20700/20710/20711 | пов'язана з екзотичним | 1  | c3c059480 | c3c059480 PARO                                       |   | 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  | 0e35cc02c | 57c65d24e€ випадок                                   |   | 0      | захворювання    | 1              |
| 1568 | Top/20000/20700/20710/20711 | пов'язана з екзотичним | 5  | 2209dc095 | b7b5c72e€ окремий випадок                            |   | 0      | інфекційне захв | 1              |
| 1569 | Top/20000/20700/20710/20711 | пов'язана з екзотичним | 2  | 9138281d€ | c3c059480 небезпечним інфекційним захворюванням люде |   | 0      | захворювання л  | 1              |
| 1570 | Top/20000/20700/20710/20711 | пов'язана з екзотичним | 2  | 86639fc62 | 9138281d€ небезпечним інфекційним захворюванням      |   | 0      | інфекційне захв | 1              |
| 1571 | Top/20000/20700/20710/20711 | пов'язана з екзотичним | 5  | a4aa2e0b€ | 86639fc62 небезпечне захворювання                    |   | 0      | людей           | 1              |
| 1572 | Top/20000/20700/20710/20711 | пов'язана з екзотичним | 5  | 810c5d78€ | 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  | bc7c0b22€ | 9138281d€ небезпечне інфекційне захворювання людей   |   | 0      |                 | 0              |
| 1577 | Top/20000/20700/20710/20711 | пов'язана з екзотичним | 99 | 00000452€ | c3c059480 надзвичайна ситуація                       |   | 0      |                 | 0              |
| 1578 | Top/20000/20700/20710/20711 | пов'язана з екзотичним | 99 | be27593b€ | c3c059480 захворювання                               |   | 0      |                 | 0              |
| 1579 | Top/20000/20700/20710/20711 | пов'язана з екзотичним | 99 | 6ff83abec | c3c059480 людина                                     |   | 0      |                 | 0              |



# Схема категоризатора



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

# Результати дослідження. Побудовані категорії

The screenshot displays the SAS Content Categorization Studio interface. The main window shows a taxonomy tree on the left and a list of categories on the right. The taxonomy tree is organized into levels, with the following categories highlighted:

- 10000 (10000)
- 10100 (10100)
- 10110 (10110)
- 10111 (10111)
- 10112 (10112)
- 10113 (10113)
- 10114 (10114)
- 10120 (10120)
- 10130 (10130)
- 10131 (10131)
- 10140 (10140)
- 10141 (10141)
- 10142 (10142)
- 10143 (10143)
- 10144 (10144)
- 10145 (10145)
- 10150 (10150)
- 10151 (10151)
- 10152 (10152)
- 10160 (10160)
- 10161 (10161)
- 10162 (10162)
- 10170 (10170)
- 10171 (10171)
- 10172 (10172)
- 10180 (10180)
- 10181 (10181)
- 10182 (10182)
- 10200 (10200)
- 10210 (10210)
- 10211 (10211)
- 10212 (10212)
- 10213 (10213)

The right pane shows a list of categories, including:

- PAR
- OR
- "пожеж"
- "вибух"
- OR
- "комунікація@N"
- DIST\_8
- "вибух@"
- "у@"
- "споруді@"
- "устаткування@N"
- DIST\_7
- "технологічне@"
- "устаткування@"
- DIST\_8
- "устаткування@"
- "промислового@"
- "об'єкта@"
- DIST\_9
- "технологічне@"
- "устаткування@"
- "промислового@"
- "об'єкта@"
- DIST\_7
- "промисловий@"
- "об'єкт@"

The interface includes a menu bar (File, Edit, View, Build, Project, Category, Concept, Testing, Document, Server, Help) and a toolbar with various icons. The status bar at the bottom indicates "Ready" and "Taxonomy Dependencies".

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

# Результати дослідження. Побудова категоризатора

SAS Information Retrieval Studio

Start Stop Apply Changes Refresh

**Status**

The pipeline server is running on WIN-24QRPTCLL3S:9245.

| Pipeline Stage         | Pending | Finished | Last Busy Time   |
|------------------------|---------|----------|------------------|
| Overall                |         | 6179     |                  |
| XML parsing            | 0       | 6179     | 2016-06-21 23:53 |
| Document processing    | 0       | 0        | 2016-06-21 23:53 |
| Sending to the indexer | 0       | 0        |                  |

**Help**

The pipeline server represents a document as a set of named fields. It sends each document through a configurable list of processing stages which can add, delete, and modify these fields, split a single document into multiple documents, or invalidate a document so that it will be dropped from the pipeline completely. Once a document reaches the end of the list of processors, it will be sent to the [indexing server](#) to make it available for searching.

Different document processors perform different tasks, and each can act upon the output of previous stages, so the order of the list is significant. The `parse_html` and `document_converter` processors handle normalization of the input, so they should be placed first in the list. The `content_categorization` processor performs analysis on documents which have already been normalized, so it should come next. The `export_csv` and `export_to_files` processors take the normalized and analyzed documents and export them for consumption by other programs, so they should come last. You can also add deployment-specific custom document processors by placing them in the `bin/postprocessors` subdirectory of your installation.

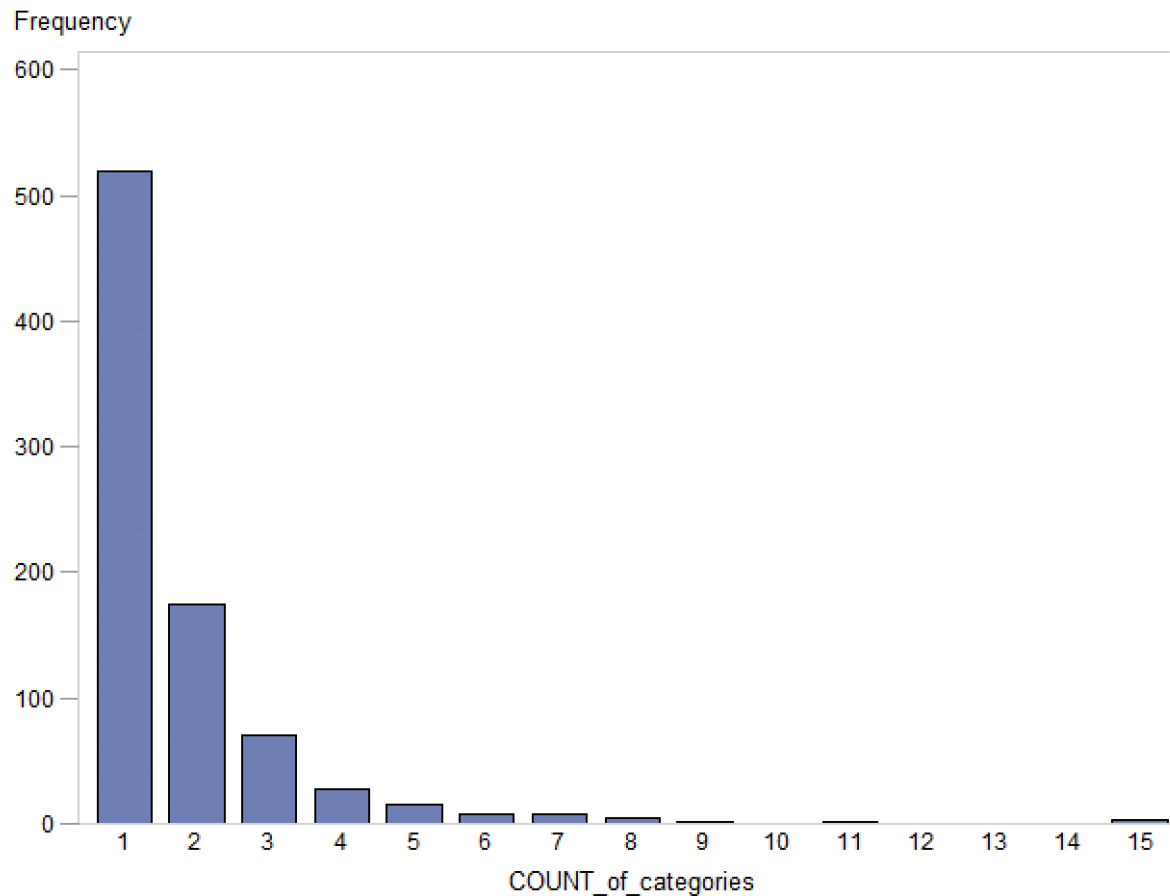
**Document Processors**

|                        |   |
|------------------------|---|
| markup_matcher         | Add<br>Remove<br>Edit<br>Move Up<br>Move Down |
| content_categorization |   |
| export_csv             |   |
| export_to_files        |   |

**Document Inspector**

Log

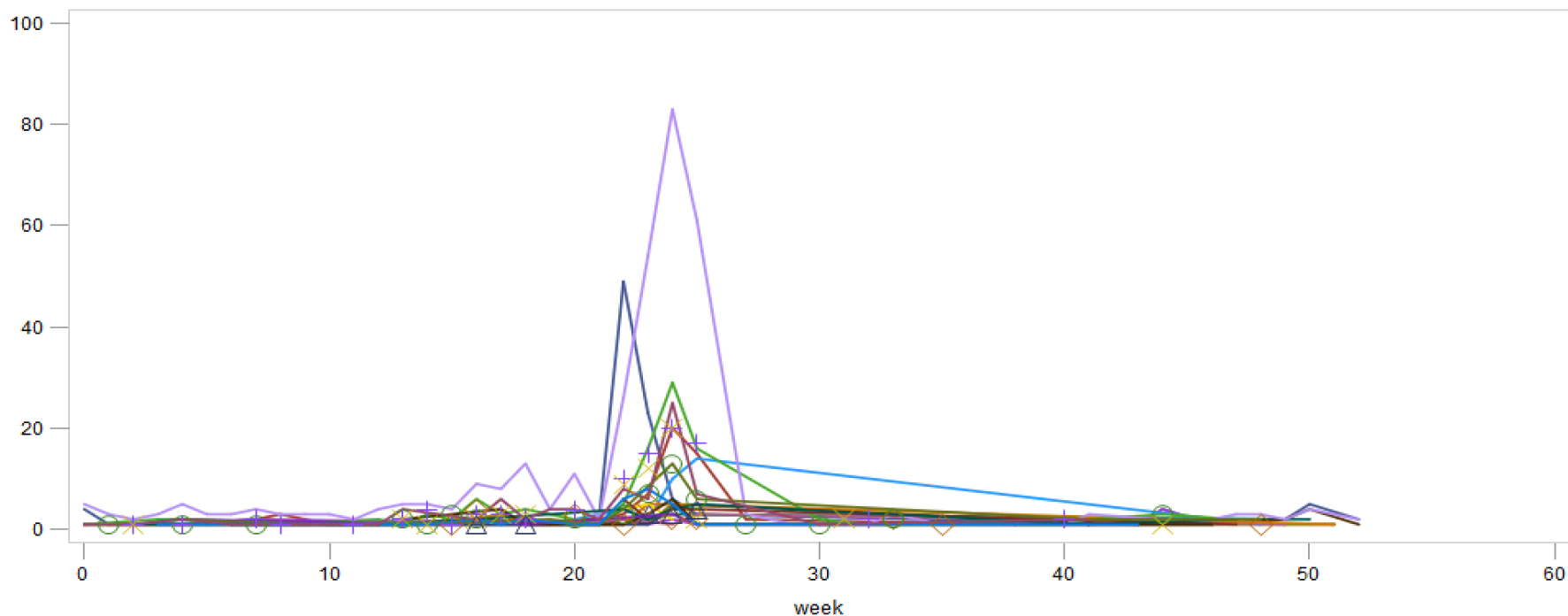
# Результати дослідження. Статистичні дані



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

COUNT\_of\_categories



| categories | Top/10000/10100             | Top/10000/10100/10150/10151 | Top/10000/10100/10160/10162 | Top/10000/10200/10210       |
|------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
|            | Top/10000/10200/10260       | Top/10000/10400/10420       | Top/10000/10400/10430       | Top/10000/10500/10530       |
|            | Top/10000/10500/10560       | Top/10000/10700/10770       | Top/10000/10800/10820       | Top/11000/11100/11130       |
|            | Top/20000/20200/20260       | Top/20000/20300/20310/20311 | Top/20000/20300/20310/20314 | Top/20000/20300/20320/20323 |
|            | Top/20000/20300/20330/20335 | Top/20000/20400/20410       | Top/20000/20700/20710/20711 | Top/20000/20700/20720/20721 |
|            | Top/30000/30100/30110       | Top/30000/30100/30150       | Top/30000/30300             | Top/30000/30400             |
|            | Top/30000/30600/30610       |                             |                             |                             |

## 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.

# 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.

# 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





**Дякую за увагу**

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