УДК 004(9+056.5)

Формування текстового корпусу Telegram- каналів

Ланде Д.В.

Постановка проблеми

Для проведення дієвих досліджень інформаційного простору необхідний доступ до текстових корпусів, що його репрезентують. Не всі інформаційні ресурси мережі готові надавати свій контент «роботам», які функціонують в автоматичному режимі [1]. Але ж на цей час існують чисельні ресурси, соціальні мережі, які все ж таки надають свою інформацію для подальших досліджень, наприклад, через API (Application Programming Interface), або у зручних форматах, різновидах XML (RSS різних версій, Atom, тощо) або JSON. Слід підкреслити, вільний доступ до інформації з мережі для «роботів», програмних агентів є базовою вимогою для запровадження семантичного вебу. У цій роботі розглядається технологічна можливість формування текстового корпусу із вибраних дослідником каналів месенджера Telegram.

Мета

Мета цієї роботи – представити методику формування навчального текстового корпусу на основі контенту вибраних каналів месенджера Telegram. Відомо три шляхи доступу засобів автоматизованого добування даних до цього ресурсу. Перший – доступ напряму до месенджера за адресою типу https://tlgrm.ru/channels/@kpilive. У цьому випадку kpilive – це назва каналу (конкретно у цьому випадку – каналу НТУУ «КПІ ім. Ігоря Сікорського» КПИ live). Другий, в іншому форматі, за адресою редиректу типу: https://tlg.repair/s/kpilive. Третій – доступ до інформації каналу у форматі Atom [2] через зовнішній агрегатор RSSHub: https://rsshub.app/telegram/channel/kpilive. В роботі наведено порівняльні характеристики якості сканування різними шляхами.

Методологія

1. Для створення текстового корпусу на основі контенту каналів месенджера Telegram першим кроком створюється список каналів, що можуть цікавити дослідника. Для цього можна звернутися до чисельних каталогів цих каналів, розміщених в мережі. Обираємо, наприклад, розміщений за адресою https://ru.telegram-store.com/catalog/product-category/channels/ (понад 50 тис. каналів). Вводимо в режимі пошуку слово, що нас цікавить, наприклад «Україна», отримуємо дві сторінки посилань за адресою: https://ru.telegram-store.com/?s=Україна. Формуємо список каналів у форматі:

@nowasteukraine
@onlineukraine
@sos_ua
@ua_rozvynuta
@ukraine_novosti
…

2. Скануємо ресурси вибраних каналів із допомогою вільно доступної програми. Відомо, що існує декілька програмних агентів (програм-роботів), що сканують контент із ресурсів мережі за протоколами HTTP/HTTPS. Обираємо програму wget, яка входить до складу багатьох систем. Застосовуємо вибраний у п. 1 список каналів, формуємо переліки адрес для трьох шляхів доступу до ресурсу:

а:
https://tlgrm.ru/channels/@nowasteukraine
https://tlgrm.ru/channels/@onlineukraine
https://tlgrm.ru/channels/@sos_ua
…
б:
https://tlg.repair/s/nowasteukraine
https://tlg.repair/s/onlineukraine

https://tlg.repair/s/sos_ua

…

в:

https://rsshub.app/telegram/channel/nowasteukraine
https://rsshub.app/telegram/channel/onlineukraine
https://rsshub.app/telegram/channel/sos_ua

…

Після запуску програми-робота командою [3]:

wget –i addr_list –O file,

отримуємо 3 файли, що відповідають наведеним підходам. Параметри наведеної програми задають ім'я файлу із списком адрес (–i) та ім'я вихідного файлу (–O). Найбільшу повноту дає підхід в), крім того, його формат Atom найбільш стандартний, але, враховуючи те, що він відповідає зовнішньому сервісу по відношенню до Telegram, було прийнято до реалізації підхід б), другий за повнотою.

3. На третьому, останньому етапі здійснюється перетворення зібраних найповніших даних до формату текстового корпусу, необхідного для дослідження (до формату XML 1.0), який у подальшому може завантажуватися в інформаційно-пошукову систему Sphinx [4], фрагмент якого такий:

```
<?xml version="1.0" encoding="utf-8"?>
<sphinx:docset>
<sphinx:schema>
<sphinx:field name="subject"/>
<sphinx:field name="content"/>
<sphinx:field name="source"/>
<sphinx:field name="datatime"/>
<sphinx:attr name="url"/>
</sphinx:schema>
<sphinx:document id="1_tg">
<subject>Фестивалі, концерти та літературні читання</subject>
<content>Фестивалі, концерти та літературні читання - ці події зовсім не поставлені на паузу через карантин...</content>
<source>Telegram: novoe_vremya</source>
<datatime>20200425 14:20</datatime>
<url>https://tlgrm.ru/channels/@novoe_vremya/11263</url>
</sphinx:document>
…
```

Висновки

В роботі описана методика формування текстового корпусу із вибраних дослідником каналів месенджера Telegram. Вибрані шляхи добування цієї інформації.

Наведена методика може застосовуватися для добування даних в інформаційно-пошукових системах, аналізу текстів, що публікуються користувачами месенджера Telegram. Методика після адаптації може бути використана і для формування текстових масивів із інших соціальних медіа.

1. Ланде Д.В., Субач І.Ю., Бояринова Ю.Є. Основи теорії і практики інтелектуального аналізу даних у сфері кібербезпеки: навчальний посібник. – К.: ІСЗЗІ КПІ ім. Ігоря Сікрорського, 2018. – 300 с. ISBN 978-966-2577-12-9.

2. Danny A., Andrew W. Beginning RSS and Atom Programming. – Wrox, 2005. – 769 p. ISBN 978-0-764-57916-5.

3. Hrvoje Nikshic. GNU Wget 1.20. The non-interactive download utility. –Free Software Foundation, Inc., 2018. – 76 p.

4. Aksyonoff A. Introduction to Search with Sphinx: From installation to relevance tuning. – O'Reilly Media, 2011. – 146 p. ISBN 978-0-596-80955-3.