УДК 656.7.085:657.71(045)

ЗАСТОСУВАННЯ ПРОГРАМИ КОНВЕРТУВАННЯ ТЕКСТОВИХ ДОКУМЕНТІВ В КОМП'ЮТЕРИЗОВАНІЙ СИСТЕМІ ПОРІВНЯЛЬНОГО АНАЛІЗУ ЕЛЕКТРОННИХ ТЕКСТІВ

Литвиненко О.Є., Лезвінський О.С.
Національний авіаційний університет
litvinen@nau.edu.ua

Розглянуто типи та принципи кодування текстової інформації. Описано систему кодування Юнікод та текстові процесори. Запропоновано простий конвертор текстових документів з DOC у TXT

In this Article were considred types and principles of code of text information. The system of Unicode is more in detail described and also was indicated about word processors, Final part is development of simple converter of text documents from DOC in TXT format.

Вступ

Інтеграційні процеси, притаманні сучасному періоду розвитку світової цивілізації, впровадження потужних засобів телекомунікацій, тотальна комп'ютеризація усіх сфер людської діяльності поставили безліч якісно нових задач у науковій галузі, яка знаходиться на стику комп'ютерних технологій та лінгвістики. Однією з таких задач є задача порівняльного аналізу електронних текстів. Вона виникає кожен раз, коли з'являється потреба у визначенні збігів або виявленні логічних протиріч у текстових документах. З проблемою визначення збігів у текстах людство зіштовхується у тих сферах своєї діяльності, де її кінцевим результатом є текстовий твір або документ. Це, в першу чергу, освіта, наука, літературна діяльність, законотворчість, діяльність видавництв, інформаційних агентств, засобів масової інформації, патентування, інноваційна та інша діяльність, пов'язана з захистом інтелектуальної власності. Друга проблема — проблема виявлення логічних протиріч у текстових документах лежить, головним чином, у площині професійних інтересів різних юридичних та інформаційно-аналітичних організацій. Особливу актуальність вона придбала останнім часом у зв'язку з перспективою вступу України до Європейського співтовариства, що зажадало, у свою чергу, гармонізації українського законодавства та його адекватності відносно загальноєвропейських нормативних актів. Очевидно, в умовах безперервно зростаючих потоків інформації вказані проблеми порівняльного аналізу текстів не можуть бути вирішені без застосування сучасних комп'ютерних технологій.

Постановка проблеми

На даний час найбільше поширенним форматом представлення текстових документів є формат .doc. Відомо, що він зручний у використанні, але потребує значних обсягів машинної пам'яті. Тому при великій кількості текстових файлів, з якими здійснюється порівняння вхідного документу у комп'ютеризованій системі порівняльного аналізу електронних текстів, отримання кінцевого результату може зажадати значних витрат машинного часу. Це обумовлює необхідність перетворення електронних текстів у більш економічний (з точки зору обсягів пам'яті) формат .txt. У свою чергу, це потребує розв'язання проблеми автоматизації процесу конвертування формату вхідних документів з формату .doc у формат .txt на основі використання сучасних методів кодування текстової інформації.

Аналіз стану проблеми

Кодування являє собою таблицю символів, де кожній літери алфавіту (а також цифрам і спеціальним знакам) присвоєно свій унікальний номер — код символу. Оскільки для кодування будь-якого символу виділено 8 двійкових розрядів (один байт інформації), таблиця містить 256 елементів — за кількістю станів, що може прийняти один байт.

На сьогодні стандартизована тільки половина таблиці, так званий ASCII-код — перші 128 символів, які містять у собі цифри, літери латинського алфавіту та інші знаки. Друга ж половина таблиці віддана під національні символи, і в кожній країні ця частина різна.

Перша російськомовна система кодування отримала назву KOI-8. Вона була розроблена у сімдесятих роках минулого століття в процесі адаптації до російської мови операційної системи UNIX. І дотепер в російськомовних країнах вона вважається основною системою кодування в операційній системі UNIX.

З появою персональних комп'ютерів широке впровадження отримала операційна система DOS. Для неї компанія Microsoft розробила спеціальну систему DOS-кодування (866-кодову сторінку) з поширеними функціональними можливостями. Наприклад, у неї були введені спеціальні символи для малювання рамок, що широко використовувалося в програмах, написаних під DOS.

Водночас паралельно з IBM-сумісними комп'ютерами розвивалося й виробництво комп'ютерів компанії Macintosh. Незважаючи на

те, що їхня частка в в російськомовних країнах відносно мала, проте потреба в русифікації систем кодування існувала. Як наслідок цього була розроблена ще одна система кодування, яка отримала назву MAC.

У 1990 році компанія Microsoft презентувала першу успішну версію Windows 3.0-3.11, а разом з нею й підтримку національних мов у вигляді нової системи Win-кодування (або кодової сторінки 1251). Де-факто вона стала найпоширенішої в російськомовних країнах.

Наступний варіант кодування зв'язаний уже не з конкретною фірмою, а зі спробами стандартизації кодувань на рівні всесвіту. Цю фукцію взяла на себе міжнародна організація зі стандартів ISO. У результаті з'явилася нова система кодування ISO-8859-5. Але вона виявилася несу-місною з існуючими широко розповсюдженими системами, тому на даний час вона практично ніде не застосовується.

У 1991 році організація Консорціум Юнікоду запропонувала свою систему кодування Unicode. Це промисловий стандарт, розроблений з метою надати текстам і символам усіх писемних систем світу узгоджене представлення і обробку комп'ютерами, тому що багато існуючих систем кодування є обмеженими в розмірі й можливостях і несумісними з багатомовними середовищами. Успіхи Юнікоду в уніфікації наборів символів призвели до його поширення і домінуючого використання в інтернаціоналізації й локалізації програмного забезпечення комп'ютерів. Стандарт був використаний у багатьох новітніх технологіях, включаючи XML, мову програмування JAVA й сучасні операційні системи.

Мета дослідження — розробка комп'ютерної програми, призначеної для автоматизованого конвертування текстових документів з формату .doc у формат .txt на базі стандарту Unicode.

Принципи конвертування текстових документів у формат .TXT

Стандарт Unicode складається з двох основних розділів — універсального набору символів і сімейства кодувань. Універсальний набір символів задає однозначну відповідність символів кодам — елементам кодового простору, що представляють негативні цілі числа. Сімейство кодувань визна-чає машинне подання послідовності кодів універсального набору символів. Стандарти наборів символів такі:

UCS-4 (англ. *Universal Character Set*) — 1 символ кодується 4 байтами, всього можна закодувати 232 символи.

UCS-2 (англ. *Universal Character Set*) — 1 символ кодується 2 байтами, всього можна закодувати 65 536 символи.

Стандарти кодувань:

1. UTF-32 (англ. *Unicode Transformation Format* — формат перетворення Юнікода) — один із способів кодування символів із Unicode у вигляді 32-бітових послідовностей. 1 символ кодується 32 бітами.

2. UTF-16 — один із способів кодування символів із Unicode у вигляді 16-бітних послідовностей. Символи з кодами менше 0x10000 (216) подаються як є (одна 16-бітова послідовність), а символи з кодами 0x10000—0x10FFFE — у вигляді двох 16-бітових послідовностей (так звана «сурогатна» пара), перша з яких лежить в діапазоні 0xD800—0xDBFF, а друга — 0xDC00—0xDFFF.

За стандартом ніякі символи не можуть мати кодів власне з діапазону 0xD800-0xDFFF, так що розшифровка кодування завжди однозначна. Утім, у переважній більшості випадків текст в UTF-16 є просто послідовністю символів з UCS-2, оскільки символи Unicode після коду 0x10000 використовуються вкрай рідко.

3. UTF-16LE та UTF-16BE. У потоці даних UTF-16 старший байт може записуватися або перед молодшим (UTF-16 Big Endian або UTF-16BE), або після молодшого (UTF-16 Little Endian або UTF-16LE). Іноді кодування Юнікода Big Endian (UTF-16BE) називають Юнікодом із зворотним порядком байтів. Аналогічно існує два варіанти 32-бітового кодування: UTF-32LE та UTF-32BE.

4. UTF-8 — поширене сьогодні кодування, що реалізовує представлення Юнікода, сумісне з 8-бітовим кодуванням тексту.

Текст, що складається тільки з символів з номером менше 128, при записі в UTF-8 перетворюється на звичайний текст ASCII. І навпаки, в тексті UTF-8 будь-який байт із значенням менше 128 зображає символ ASCII з тим же кодом. Решта символів Юнікода зображається послідовностями завдовжки від 2 до 6 байтів (реально тільки до 4 байт, оскільки використання кодів більше 221 не планується), в яких перший байт завжди має вигляд 11xxxxxx, а інші — 10xxxxxx.

Простіше кажучи, у форматі UTF-8 символи латинського алфавіту, розділові знаки і керуючі символи ASCII записуються ASCII-кодами, а решта всіх символів кодується за допомогою октетів (послідовності довжиною 8 біт) із старшим бітом 1.

У результаті, навіть якщо програма не розпізнає Юнікод, то латинські літери, арабські цифри і розділові знаки зображатимуться правильно.

Символи UTF-8 отримують з Unicode так:

Unicode UTF-8
0x00000000 — 0x0000007F: 0xxxxxxx
0x00000080 — 0x000007FF: 110xxxxx 10xxxxxx
0x00000800 — 0x0000FFFF: 1110xxxx 10xxxxxx 10xxxxxx
0x00010000 — 0x001FFFFF: 11110xxx 10xxxxxx 10xxxxxx 10xxxxxx

Також теоретично можливі, але не включені в стандарти:

Unicode UTF-8
0x00200000 — 0x03FFFFFF: 111110xx
10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx
0x04000000 — 0x7FFFFFFF: 1111110x
10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx
10xxxxxx

У форматуванні текстової інформації важливу роль відіграє текстовий процесор — комп'ютерна програма з розширеними можливостями для комп'ютерної підготовки повноцінних документів від особистих листів до офіційних паперів. Функції текстових процесорів зазвичай включають компоновку і форматування тексту, широкі можливості роботи зі змістом і сторінками, розширений набір доступних символів, перевірку орфографії, впровадження в документ гіперпосилань, графіків, формул, таблиць. У часи використання системи DOS широке розповсюдження мав текстовий процесор Lexicon, котрий міг обробити TXT-формат на довільному рівні.

Сьогодні основним інструментом для роботи з документами формату .TXT є стандартний Блокнот Windows. Основним текстовим процесором для роботи з файлами, представленими у форматах DOC, TXT, RTF є MS Word, який реалізує всі властивості цих форматів. При розробці програми конвертування електронних текстів у формат .TXT були використані такі закономірності, які були виявлені під час дослідження DOC-файлів:

1. Перші 600h байти DOC-файлів містять службові дані.
2. Два байти (0Dh, 00h) вказують на кінець абзацу.
3. Два байти, перший з яких міститься в діапазоні 20h—7Fh, а другий дорівнює нулю, означають символ першої половини кодової таблиці.
4. Два байти, перший з яких міститься в діапазоні 10h—2Fh, а другий дорівнює 4, означають символи від «А» до «Я».
5. Аналогічно, два байти, перший з яких міститься в діапазоні 30h—4Fh, а другий дорівнює 4, означають символи від «а» до «я».
6. Два байти (14h, 20h) означають символ «–«.
7. Два байти (1Ch, 20h) означають символ «лапки, які відкриваються».
8. Два байти (1Dh, 20h) означають символ «лапки, які закриваються»;.

При запуску програми потрібно вказати в командному рядку такі параметри: ім'я вихідного файлу, ім'я вхідного файлу і необов'язковий параметр «/a». Від параметру «/a» залежить, як конвертер буде визначати кінець тексту в DOC-файлі. Результати реалізації програми зображенні поетапно на рис.1—3.

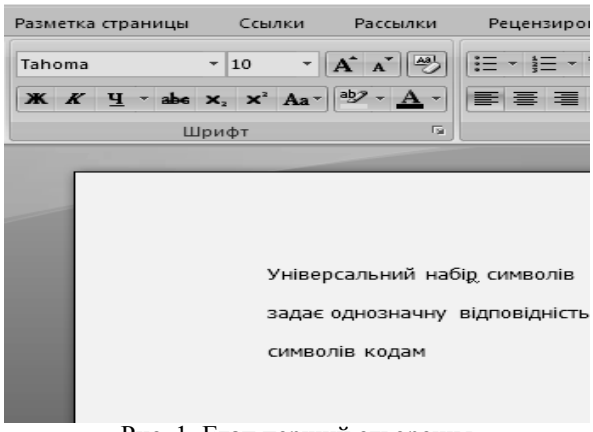

Рис. 1. Етап перший створення або використання вже існуючого DOC файла

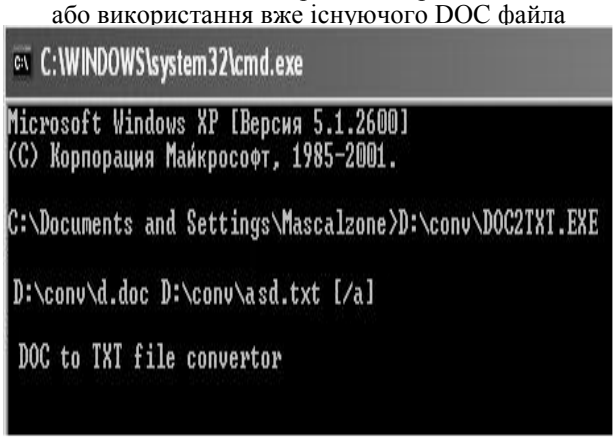

Рис. 2. Другий етап — запуск програми

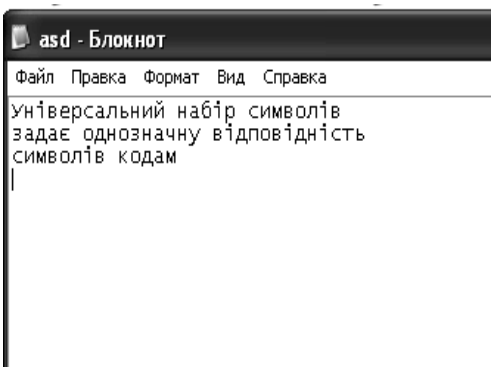

Рис. 3. Третій етап — перевірка утвореного TXT

Висновки

Експериментальні дослідження довели переваги TXT-формату перед іншими форматами представлення текстових документів. Його перевагами є економія обсягів машинної пам'яті, можливість коригування текстів будь-яким стандартним текстовим редактором, відсутність залежності від порядку байтів та довжини машинного слова на різних платформах, можливість відновлення пошкоджений файлів тощо.

Використання TXT-формату у комп'ютеризованій системі порівняльного аналізу електронних текстів призводить (в порівнянні з іншими форматами) до суттєвого зменшення витрат машинного часу для отримання кінцевого результату, що є особливо важливим при

наявності великої кількості текстів, що підлягають порівнянню.

ЛІТЕРАТУРА

1. *Розробка* сайту html, css, xml (електронний.ресур)-http://siterozrob.ru/dizajn/-nebagatoliv-pro-koduvannya/

2. *Юнікод* (матеріал з Вікіпедії вільної електронної енциклопедії)-http:/uk.wikipedia. prg/wiki/

3. *Текстовые* форматы и редакторытекстових файлов (електронний ресурс) http://www.prodtp.-ru/index.php?act=recipes&CODE=03&id=124.

Стаття надійшла до редакції 24.04.09