Ministerul Educației, Culturii și Cercetării al Republicii Moldova
Universitatea Tehnică a Moldovei
Facultatea Calculatoare, Informatică și Microelectronică
Departamentul Informatică și Ingineria Sistemelor

Admis la susținere
Șef departament: conf. univ. dr.
V. SUDACEVSCHI

_____
„__”_____ 2021

RECUNOAȘTEREA TEXTELOR VECHI PENTRU DIGITIZAREA PATRIMONIULUI CULTURAL

Student:_____(Luca Maria)

Conducător:__ _____(Bobicev Victoria)

CHIȘINĂU – 2021

ADNOTARE

la teza de master cu tema „Recunoașterea textelor vechi pentru digitizarea patrimoniului cultural” a studentei gr. MAI-191M, Luca Maria

Digitalizarea patrimoniului cultural poate fi un instrument esențial în eforturile de astăzi în direcția conservării, renovării, studiului și promovării resurselor culturale. Recunoașterea textelor este o componentă importantă în procesul de digitizare a patrimoniului cultural. În comparație cu textele tipărite, procesul de recunoaștere a manuscriselor este foarte complicat din cauza că sunt necesare niște operații suplimentare, precum ajustarea contrastului, curățarea imaginii și segmentarea textului. Astfel, pentru manuscrise este necesar în faza incipientă de a obține un set de date care să fie utilizat ulterior pentru antrenarea algoritmilor de recunoaștere automată a textelor vechi. Teza de master este constituită din introducere și trei capitole. Scopul tezei este elaborarea unui instrument pentru suportul digitizării și recunoașterii textelor vechi din manuscrise, prin intermediul căruia va fi facilitat accesul la patrimoniul cultural al țării. Strategia prin care a fost atins scopul este realizarea unui joc cu scop.

În raport sunt prezentate rezultatele studiului și analizei modelelor și metodelor de realizare a unui joc cu scop pentru recunoașterea textelor vechi din manuscrise și instrumentele de lucru ce trebuie utilizate pentru realizarea metodologiei stabilite. Este prezentat planul de acțiuni care este urmat la elaborarea aplicației în baza metodelor studiate și cuprinde, totodată, cerințele ce necesită respectate pentru ca jocul cu scop să aibă impact și să fie atractiv utilizatorilor.

În urmare este prezentat modelul logic și conceptual al aplicației, schema ER a Bazei de Date, proiectarea aplicației, realizarea metodologiei descrise în capitolul II și rezultatele testării acesteia în forma unui studiu de caz.

Pentru a adnota cât mai multe manuscrise, în acest proiect este folosită tehnica de crowdsourcing. Aceasta ne garantează acces la experți în adnotare, dar și de simpli oameni care odată jucând, contribuie la creșterea setului de date pentru antrenare. Rezultatele acestei teze vor deveni sursă deschisă pentru a putea fi preluate de către cercetători din Republica Moldova care ulterior vor dezvolta algoritmi pentru recunoașterea automată a textelor din manuscrise.

Cuvinte cheie: recunoașterea textelor manuscrise, patrimoniul cultural, digitizarea patrimoniului cultural, joc cu scop, alfabet chirilic.

ANNOTATION

to the master thesis on topic "Recognition of old texts for the digitization of cultural heritage" of the student gr. MAI-191M, Luca Maria

The digitization of cultural heritage can be an essential tool in today's efforts towards conservation, renovation, study and promotion of cultural resources. Text recognition is an important component in the process of digitization of cultural heritage. Compared to printed texts, the process of recognizing manuscripts is very complicated because additional operations such as contrast adjustment, image cleaning and text segmentation are required. Thus, for manuscripts it is necessary to obtain a dataset, at an early stage, that can be used later for training automatic recognition algorithms for old texts. The master thesis consists of an introduction and three chapters. The aim of the thesis is to develop a tool to support the digitization and recognition of ancient texts in manuscripts, collaterally facilitating access to the country's cultural heritage. The thesis aims to achieve this goal through a purpose-crafted game.

The report presents the results of the study and analysis of the models and methods for the realization of a purposeful game for the recognition of ancient texts from manuscripts and the working tools necessary for the realization of the established methodology. This chapter describes the action plan that is followed when developing the application based on the studied methods and also includes the requirements that need to be met in order for the application to have an impact and attract users.

Finally, the logical and conceptual model of the application, the ER schema of the Database, the design of the application, the realization of the methodology and the results of its testing in the form of a case study are presented.

In order to annotate as many manuscripts as possible, the project employs the crowdsourcing technique. This guarantees us access to experts in annotation, but also to ordinary people who can aid the training of the dataset by playing. The results of this thesis will become open source so that they can be taken up by researchers in the Republic of Moldova who will then develop algorithms for automatic text recognition in manuscripts.

Keywords: manuscript text recognition, cultural heritage, digitization of cultural heritage, game with a purpose, Cyrillic alphabet.

CUPRINS

LISTA ABREVIERILOR 8

INTRODUCERE 9

CAPITOLUL I. DOMENIUL DE STUDIU. NOȚIUNI GENERALE. AUDITUL DOMENIULUI DE STUDIU 11

1.1. Domeniul de studiu. Noțiuni generale. Auditul domeniului de studiu. 11

1.2. Aspecte ale recunoașterii textelor vechi pentru digitizarea patrimoniului cultural. Principii de organizare a recunoașterii textelor vechi pentru digitizarea patrimoniului cultural. 17

1.3. Probleme. Strategia, scopul și obiectivele tezei de master 18

CAPITOLUL II. MODELE ȘI METODE DE SOLUȚIONARE A PROBLEMEI MAJORE IDENTIFICATE, PRIVIND RECUNOAȘTEREA TEXTELOR VECHI PENTRU DIGITIZAREA PATRIMONIULUI CULTURAL 22

2.1 Identificarea problemei majore, privind recunoașterea textelor vechi la digitizarea patrimoniului cultural 22

2.2 Metode și instrumente, privind recunoașterea textelor vechi, utilizate în rezolvarea problemei majore în digitizarea patrimoniului cultural 27

CAPITOLUL III. REALIZAREA APLICAȚIEI DE RECUNOAȘTERE A TEXTELOR VECHI CU AJUTORUL GWAP. 46

3.1 Experiența utilizării TI&C pentru recunoașterea textelor în rezolvarea problemei digitizării patrimoniului cultural 46

3.2 Modelul conceptual și modelul logic al jocului cu scop pentru recunoașterea textelor în rezolvarea problemei digitizării patrimoniului cultural 48

3.3 Schema ER a BD a jocului cu scop pentru recunoașterea textelor în rezolvarea problemei digitizării patrimoniului cultural 57

3.4 Elaborarea aplicației de realizare a instrumentului TI&C pentru recunoașterea textelor în rezolvarea problemei digitizării patrimoniului cultural 60

3.5 Studiu de caz. 65

CONCLUZII 67

BIBLIOGRAFIE 69

INTRODUCERE

Digitizarea patrimoniului cultural reprezintă un domeniu prioritar pe agenda digitală a Uniunii Europene. Comisia Europeană a elaborat chiar și niște recomandări privind digitizarea și accesibilitatea online a materialului cultural și conservarea digitală. Digitalizarea patrimoniului cultural poate fi un instrument esențial în eforturile de astăzi în direcția conservării, renovării, studiului și promovării resurselor culturale. „Colecții uriașe de documente vechi sunt publicate de bibliotecile digitale online din întreaga lume în scopul conservării și pentru a le face disponibile online” [11]. Comunitatea internațională recunoaște importanța protejării patrimoniului cultural și reafirmă angajamentul în lupta împotriva distrugerii intenționate sub orice formă, astfel încât moștenirea culturală să poată fi transmisă generațiilor viitoare.

Digitalizarea este conversia informațiilor analogice de orice tip, fie text, fotografii, voce, în formă digitală cu ajutorul dispozitivelor electronice, astfel încât informațiilor vor fi prelucrate și transmise prin intermediul unor circuite digitale, echipamente și rețele. Prin digitalizarea colecțiilor de documente se asigură o bună promovare a valorilor naționale, o mai bună diseminare a informației și o valorificare superioară la nivel național și internațional a colecțiilor speciale, a documentelor rare.

În Republica Moldova există foarte multe documente text antice de o valoare ireproșabilă. Este cu adevărat o provocare pentru oricine să poată căuta prin aceste date pe hârtie. Documentele pe hârtie sunt scanate pentru a digitaliza datele, dar datele scanate sunt în formă picturală. Nu pot fi recunoscute de computere deoarece computerele pot înțelege caractere alfanumerice standard ca ASCII sau alte coduri. Prin urmare, informațiile alfanumerice trebuie extrase din imaginile scanate. Sistemul optic de recunoaștere a caracterelor ne permite să convertim un document în text electronic, doar dacă acesta a fost tipărit.

Procesul de digitizare și recunoaștere este constituit din următoarele etape: digitizarea textului pentru obținerea copiilor electronice grafice, recunoașterea textului utilizând metode standardizate, precum recunoașterea optică a caracterelor sau proceduri ale inteligenței artificiale, iar la transliterarea textului se va ține cont de literele specifice utilizate în textul inițial, și verificarea textului recunoscut, care utilizează resurse lingvistice din perioada de timp respectivă.

Recunoașterea textelor este o componentă importantă în procesul de digitizare a patrimoniului cultural. Procesul de recunoaștere a manuscriselor este foarte complicat din cauza că sunt necesare niște operații suplimentare, precum ajustarea contrastului, curățarea imaginii și segmentarea textului.

În general, nu este posibilă automatizarea completă a recunoașterii textelor vechi. Astfel, obiectivul urmărit este maximizarea suportului lucrărilor manuale, cu dezvoltarea suplimentară la semi-automatizarea sub controlul uman.

În cadrul tezei de master se urmărește elaborarea unui instrument pentru suportul digitizării și recunoașterii textelor vechi din manuscrise, prin intermediul căruia va fi facilitat accesul la patrimoniul cultural al țării.

Sistemul creat este bazat pe o reducere a cheltuielilor de adnotare a manuscriselor, dar și optimizarea activității lingviștilor care, pentru adnotarea unei pagini din manuscris, pot să aibă nevoie de aproximativ o săptămână. Astfel, în cadrul tezei de master este realizată o aplicație bazată pe tehnologia jocului cu scop, prin care este redus costul adnotării datelor și crescut nivelul de participare umană.

Consultarea în permanență cu un expert în domeniu pe parcursul elaborării tezei de master a fost strategia de lucru care a servit drept suport în obținerea rezultatelor planificate ale tezei de master date. Beneficiarii tezei de master sunt potențialii utilizatori ai aplicației, pe care îi putem repartiza în grupuri specifice, cum ar fi lingviști, filologi, cercetători cointeresați, istorici, bibliotecari, preoți, etc.

BIBLIOGRAFIE

1. Cojocaru Svetlana, Alexander Colesnicov, Ludmila Malahova, Tudor Bumbu, Ștefan Ungur, 2019, Raport științific final privind executarea proiectului de cercetări științifice aplicative „Tehnologii și resurse informaționale pentru digitizarea patrimoniului românesc istorico-literar din secolele 17-20 tipărit cu alfabet chirilic”;
2. Elena Boian, Constantin Ciubotaru, Svetlana Cojocaru, Alexandru Colesnicov, Ludmila Mahlov, 2014, Digitizarea și recunoașterea și conservarea patrimoniului cultural – istoric, Institutul de Matematică și Informatică al AȘM;
3. De Luis von Ahn, Laura Dabbish, Proiectarea jocurilor cu un scop, Comunicări ale ACM, august 2008, vol. 51 nr. 8;
4. Luis vom Ahn, Game with a Purpose, Invisible Computing, iunie 2006;
5. David Styles, 15 iulie 2019, European Union’s Heritage at Risk highlights role of digital technology in restoration, European Union’s Heritage at Risk highlights role of digital technology in restoration - Museums + Heritage Advisor (museumsandheritage.com);
6. Beth Daley, 15 iulie 2019, Importance of digitising cultural heritage highlighted in „Heritage at Risk” exhibition, https://pro.europeana.eu/post/importance-of-digitising-cultural-heritage-highlighted-in-heritage-at-risk-exhibition;
7. Zaiontz C., *(W).* ianuarie 2021, *Kendall's Coefficient of Concordance,* www.real-statistics.com: http://www.real-statistics.com/reliability/kendalls-w;
8. Hotărârea nr. 478 din 04 iulie 2012 cu privire la Programul naţional de informatizare a sferei culturii pentru anii 2012-2020, Monitorul Oficial Nr. 143-148 art. 531.
9. Tezaurul Național Digital, www.digi.emoldova.org;
10. David Vickrey, Aaron Bronzan, William Choi, Aman Kumar, Jason Turner-Maier, Arthur Wang, Daphne Koller, octombrie 2008, „Online Word Games for Semantic Data Collection”, Conferința privind metodele empirice în prelucrarea limbajului natural;
11. Alfons Juan, Veronica Romero, Joan Andreu Sanchez, Nicolas Serrano, Alejandro H. Toselli, Enrique Vidal, 2010, „Handwritten Text Recognition for Ancient Documents”, Atelier de lucru privind aplicațiile de analiză a modelelor;
12. Tudor Bumbu, 2020, „On Alignment of Textual Elements in a Parallel Diachronic Corpus”, Computer Science Journal of Moldova, vol. 28, no.3(84);