KULTURA 2.0 - Prestakuntza-pilulak:
OCR: irudietako karaktereen ezagutza optikorako teknologia

2011ko apirila

Kultura 2.0 ekimenaren barruan, «kultura 2.0 prestakuntza-pilulak» prestatu eta hedatzen hasi da Kulturaren Euskal Behatokia, EAEn kultura-sektoreko enpresek eta profesionalek egindako eskaerei erantzun zehatza emateko asmoz.

Zer da OCRa?

OCRa (Optical Character Recognition) teknologia-mota bat da, objektuak ezagutzeko giza begiak duen gaitasuna emulatu nahi duena. Zehatzago, software horrek irudi bateko (dela eskaneatutako agiria, dela argazkia) karaktereen ezagutza-optikoa egiten du. Hala, ordenagailuak karaktere horiek uler ditzake eta, ondorioz, editatu ahal den testu-formatuko fitxategia lortzen dugu. Irteerako fitxategiaren formatua (txt, pdf, etab.), software zehatzaren aukeren araberakoa izango da.



Nola dabil OCRa?

Karaktereak ezagutzeko, softwareak pixelez pixel[1] arakatzen du irudia, karaktereen itxura daukaten formen bila. Softwarearen konplexutasun- edo garapen-mailaren arabera, programako karaktereekin eta letra-tipoekin duen antzekotasuna bilatzen du, edo ezaugarrien arabera identifikatzen ditu, bilaketa ez dadin mugatu letra-tipo zehatz batzuetara.

OCRak agiri baten elementuak azter ditzake: testu-blokeak, irudiak, taulak... Zuriuneak aztertu eta testua lerrotan, hitzetan eta karakteretan banakatzen du; horrela, programak hainbat hipotesi egin eta programaren beraren hiztegiekin alderatzen ditu (programa berrienek hainbat hizkuntzatako hiztegiak dituzte), hitzak eta testuak osatzeko.

Gaur egun, irteerako fitxategian jatorrizko egitura manten dezake OCRak, eta eskuz idatzitako agiri, diagrama, partitura eta abarretako karaktereak ezagut ditzake. Halere, ez dugu ahantzi behar eraginkortasuna ez dela erabatekoa eta, beraz, sortutako testua, gero, eskuz berrikusi eta zuzendu beharko dugula. Merkatuan OCR softwareen eskaintza zabala daukagu. Beraien konplexutasun-maila eta prezioa, berriz, karaktereak ezagutzeko zehaztasun eta eraginkortasunaren araberakoa izango da.

1 Irudi digital baten unitate edo elementu txikiena da.

Zertarako erabiltzen da OCRa?

EOCRa zeharkako teknologia da eta hainbat esparru eta sektoretan erabil daiteke, inprimakiak, administrazioko agiriak, txostenak eta abar digitalizatzeko. Izan ere, programak abantaila berberak eskaintzen dizkie horiei guztiei.

Kulturaren sektorean, ondarea zaintzeari lotuta, adibidez, OCRa batik bat agiri historikoak digitalizatzeko erabiltzen da, jatorrizko testua paperean nahiz mikroformatuan dagoenean.

OCRa era guztietako agiri-motetan erabil daiteke, beraz. Ondorenean, *«Proceso de Digitalización en la Biblioteca Nacional de España»* (Digitalizatze-prozesua Espainiako Liburutegi Nazionalean) agirian ateratako aukera batzuk zerrendatuko ditugu:

Agiri-mota	Gordetzeko fitxategia	Hedatzeko fitxategia
Testu inprimatua	X (irudi gabeko testu inprimatua)	X
Prentsa	X	X
Idazmakinaz idatzitako orrialdeak	X	X
Panfletoak	X	X
Inprimatutako partiturak	–	X

Jokabide egokiak – IMPACT PROIEKTUA

i2010 europar ekimenaren inguruan sortutako digitalizatze-proiektuetan arazoak antzeman dituztela-eta, IMPACT proiektua jarri da abian. Bertan, eskualde eta nazio mailako 11 liburutegik, 13 ikerkuntza-erakundek eta 2 enpresak hartzen dute parte. Nork bere ezagutza partekatzen du, digitalizatze masiboetako proiektuetan sortzen diren arazoak hobetu eta konpontzeko gai izango den OCR teknologia garatzeko helburuaz.

Zer abantaila eskain dezake OCR erabiltzeak?

Agiriak bilatu eta berreskuratzea

1. Agiri osoan testu librea bilatzeko aukera ematen du OCR aplikazioak.
2. Metadatuak sortzeko prozesuan, automatikoki ezagutu eta hitz-gakoen aurkibideak sortzeko erabil daiteke OCRa.

Agirien kudeaketa

3. OCRari esker, digitalizatutako agirietako testua formatu editagarria bihur daiteke.
4. OCRa ez da, berez, ikus desgaitasunak dituzten pertsonentzat agiriak egokitzeko tresna; alabaina, beste teknologia batzuekin batera erabiliz gero, ateratzen den testua braille-lerrotan edo audio-artxibotan sintetiza daiteke.

Ikuspuntu ekonomikoa

5. Datuak eskuz sartzeko aukeraren aldean, denbora asko aurrezten da (OCRak 1.200 karaktere ere irakur ditzake segundoko).
6. Fitxategiak testu-formatuan gordez gero, irudi-formatuan gordeta baino askoz espazio gehiago aurrez daiteke (irudi-fitxategiak hartzen duen lekuaren 1/3 hartzen du testu-fitxategiak, gutxi gorabehera).

Zein dira OCRaren arazoak?

Europako Batzordeak sustatutako i2010 ekimenaren testuinguruan, digitalizatze masiboko proiektu asko egin dira, eta honako arazo hauek sortu dira OCRaren inguruan:

1. Erakundeetan ez dago teknologia horren inguruko ezagutzarik eta pertsona aditurik.

2. Testu elektronikoa (ez nahastu irudi digitalarekin) funtzio guztiekin nahi izanez gero, hori sortzeko kostua handia da (prozesu hori testua tekleatuz edo OCRaren bidez –gero testua berrikusi eta zuzenduz– egin daiteke).

3. OCRak ez dauka eraginkortasun handirik agiri historikoak –XIX. mendean liburuak industrialki argitaratzen hasi aurrekoak– ezagutzeko unean.

OCRaren eraginkortasun-maila mugatuari dagokionez, softwarearen kalitateaz eta garapen-mailaz gain, jatorrizko agiriaren egoera fisikoari edo irudi digitalaren kalitateari lotutako kanpo-arazoak ere badaude, prozesua burutzeko erabakigarriak izan daitezkeenak:

Jatorrizko agiriaren egoerari lotutako faktoreak:

- Jatorrizko agiriak narriatuta egotea
- Hizkiak lausotuta egon edo irakurgaitzak izan daitezke
- Paperean orbainak edo gardentasunak ager daitezke
- Hizkiak zatikatuta edo gainjarrita egon daitezke
- Letra-tipoa arraroak izan ditzake, edo dagoeneko erabiltzen ez direnak
- Jatorrizko agiriaren tamaina (ez dio eragiten zuzenean OCRaren eraginkortasunari baina, jatorrizko agiriaren tamainaren ondorioz –adibidez egunkarien kasuan–, ezinezkoa gerta daiteke ohiko eskaner-makinak erabiltzea)

Irudi digitalaren tamainari lotutako faktoreak:

- Irudiaren bereizmen eskasa
- Eskanerra egoki konfiguratuta ez egotea

OCRa ezartzeko gomendio teknikoak

Digitalizatze-prozesuetan parte hartu duten erakundeek hainbat gomendio egin dituzte, batez ere eskaneatuko den irudiaren gutxieneko bereizmenari lotuta. Aurretik esan dugunez, faktore hori funtsezkoa da emaitza egokia lortzeko eta, zenbat eta bereizmen handiagoa izan, OCRak orduan eta zehaztasun handiagoa edukiko du.

Oro har, karaktereak zuzen ezagutzeko, gutxienez 300 dpi-ko[2] bereizmena jotzen da beharrezkoa; dena dela, agiriaren ezaugarrien arabera, bereizmen handiagoa ere aholkatzen da.

Agiri-mota	Gutxieneko bereizmena
Hizki argiak dituzten testuak	300 dpi
Hizki txikiko letra-tipoak, edo kalitate eskaseko originalak (prentsa)	600 dpi

¿Kultura 2.0ren inguruko zer beste gai aipa genitzake hurrengo prestakuntza-piluletan? (ahalik eta hobekien zehaztu)

Bidal iezaguzu zure iradokizuna honako e-mail honetara: kulturabehatokia@ej-gv.es

2 Hazbete bakoitzeko puntuak.