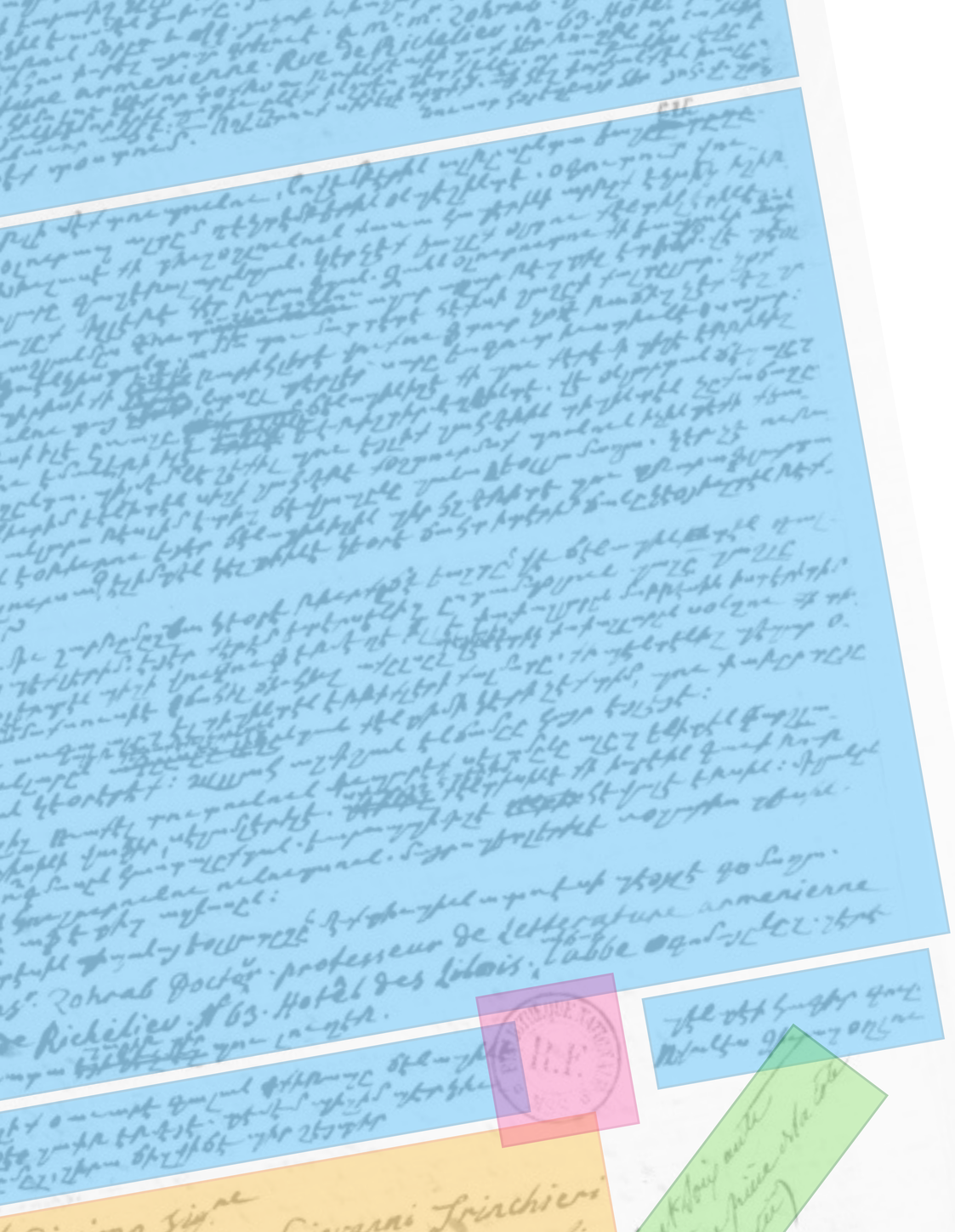


Հիմնարար գիտական գրադարան
29, Մարտ, 2024 թ.

Գրադարաններ և արհեստական բանականութիւն (AI)

Օգտագործել AI-ն հայկական հաւաքածուները զարգացնելու համար

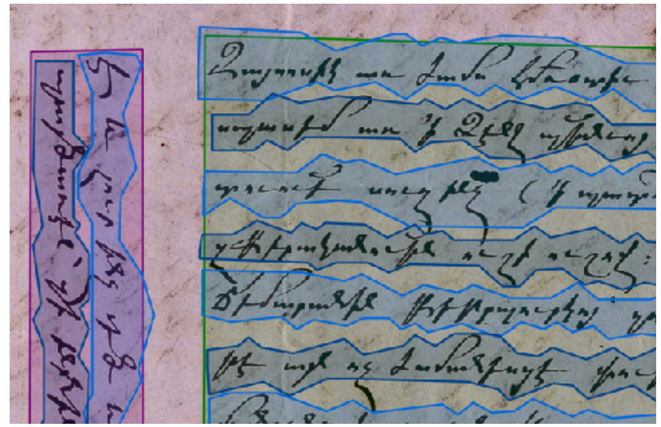
Շահան Վիդալ-Գորեն
Calfa

CALFA

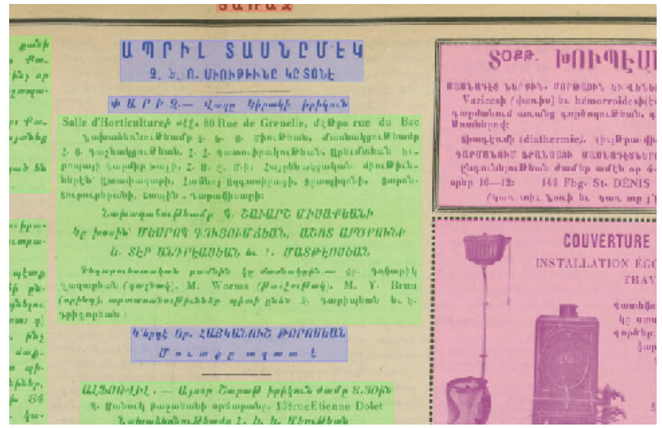
CALFA

OCR-ի (Optical Character Recognition) և արհեստական բանականութեան (AI) ճարտարագիտութիւնները արեւելեան լեզուների և հայերէնի բնագիրների ճանաչման համար

- Նուիրուած ժառանգութեան պահպահման և զարգացման:
- Գլխաւոր նպատակ. դիւրացնել հանրութեան և գիտաշխատողներու համար փաստաթղթերի և գիտելիքների հասանելիութիւնը:
- Մասնակցել թուայնացման նախագծերին և տուեալների շտեմարանի զարգացման:
- Շարունակական մշակում. արաբական և հայկական հին ձեռագիրների, հայկական ժամանակակից ձեռագիր փաստաթղթերի և թերթերի:

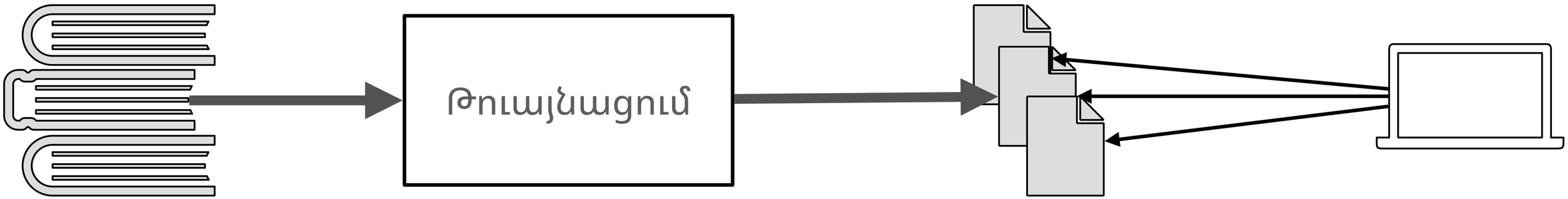


2022, Armenian, Mekhitarist Congregation



2021, Armenian, Fundamental Scientific Library

Թուայնացնել, որպէսզի հայկական ժառանգութիւնը պահպանել և զարգացնել



Փաստաթուղթ

Թուային պատճէն

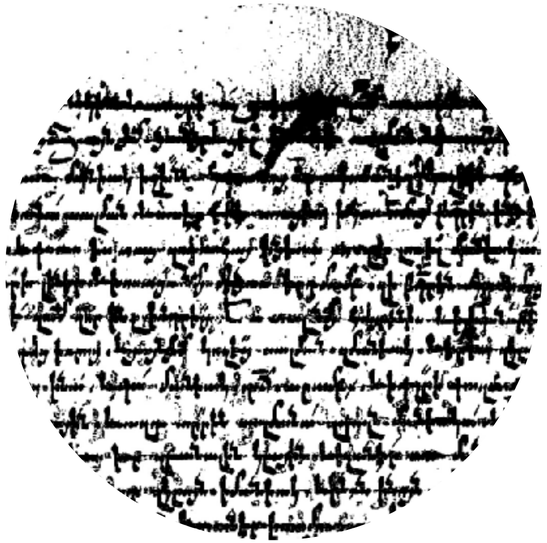
Գործածող

Գրադարան

Թուային գրադարան

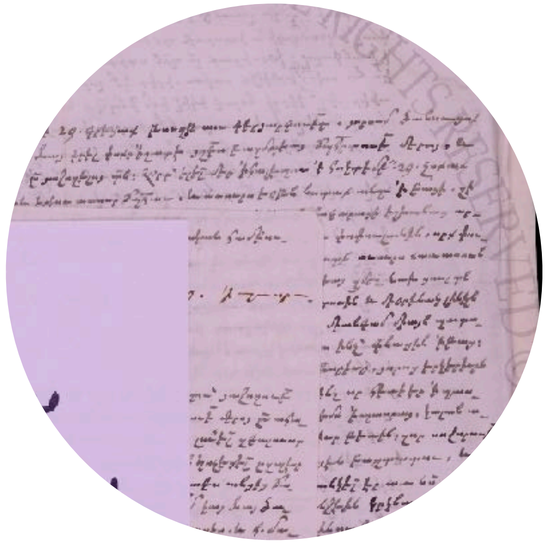
Օրինակ՝
- Pan-Armenian Digital Libary (FSL): https://arar.sci.am/dlibra
- Armenian newspapers (NLA): http://tert.nla.am/
- Armenian manuscripts, see the index of Digitized Armenian manuscripts (Calfa): https://www.armenian-manuscripts-index.com/

Թուայնացնել, որպէսզի հայկական ժառանգութիւնը պահպանել և զարգացնել



Ձեռագիր փաստաթղթերը պահպանել պատճէներ օգտագործելով

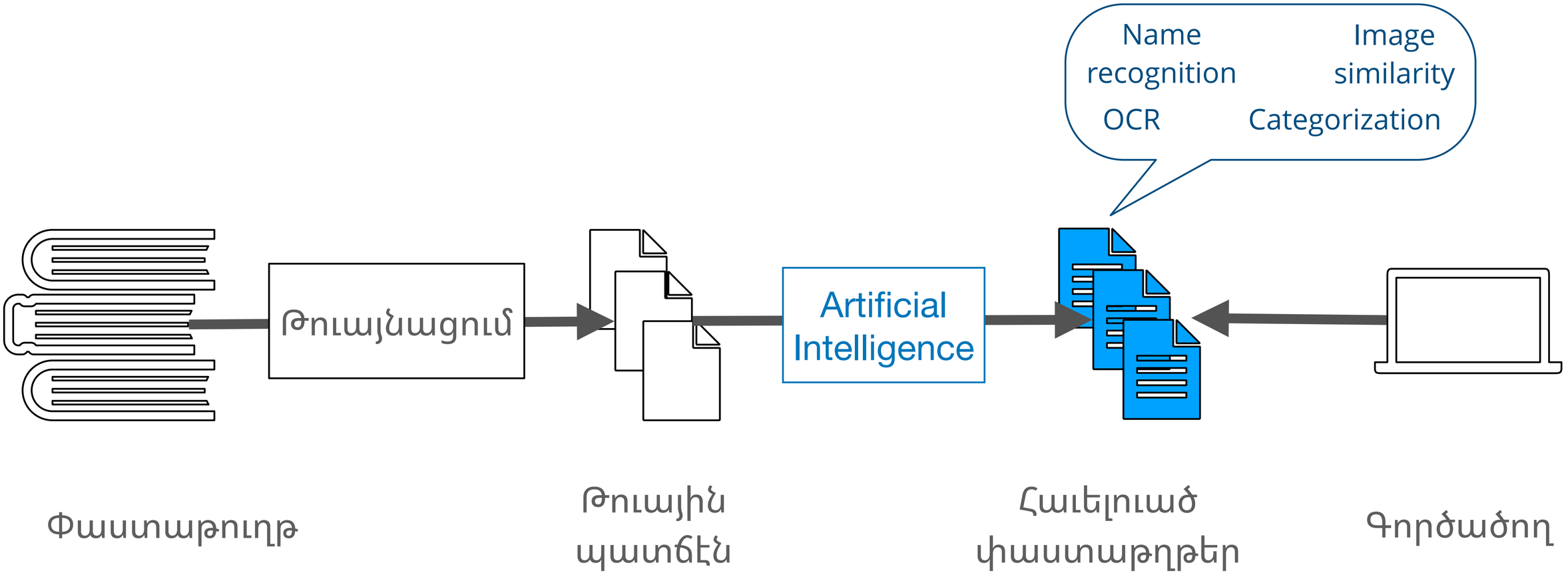
Ժառանգութիւնը մատչելի և տրամադրելի դարձնել կայքէջների միջոցով



Հայկական մշակոյթը զարգացնել

Խնդիր՝ դիւրին չէ թուային պատկերի հետ աշխատել: Իսկ եթէ առարկան աւրուած է, լինի ձեռագիր կամ հին (գրաբար), կարդալը դժուար կարող է լինել (բացի եթէ սովորել ենք այս փասթաթղթերը վերծանել):

Արհեստական բանականութիւնը օգտագործել տեղեկութիւնները գտնելու համար



Գրադարան

Թուային գրադարան

Ներկայացման ժամանակ մտքին պահել.

1

Արհեստական բանականութիւն = որոշմանը օգնելու գործիք

Արհեստական բանականութիւնը թուային հումանիտար գիտութիւններում միայն որոշում կայացնելու օգնութեան գործիք է, որն մեքենականացնում է որոշակի առաջադրանքներ, որոշակի ճշգրտութեամբ: **Երբեք կատարեալ չէ: Բայց արագ է:**

2

Հետազոտողներն ու գրադարանավարները AI-ի նոր մոդելների մշակման և սահմանման գործընթացի հիմքում են

Գրադարանավարները որոշում են, թէ ինչպէս ստեղծել և ինչպէս օգտագործել AI-ն և ինչ կարիքների համար:

3

Ընդհանուր մոդել չկայ, միայն մասնագիտացուած մոդելներ կան (արեւելեան գրութիւններ)

Բայց շատ հեշտ է մոդելը մասնագիտացնել նոր առաջադրանքի,
նոր ձեռագրի համար, և այլն:

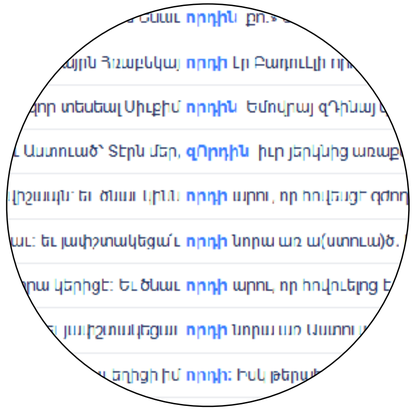
OCR (Optical Character Recognition)

Գրանշանների օպտիկական ճանաչում

Գրանշանների ճանաչում՝ ինչո՞ւ:

Մատչելի դարձնել փաստաթղթերի հասանելիությունը

Փաստաթղթերից յատուկ տեղեկություններ քաղել և գիտությունը զարգացնել



Գիտության պահպանումը և բացայայտումը ապահովել

Տպագիր փաստաթղթեր - Գրանշանների օպտիկական ճանաչում (OCR)

OCR. Տեքստի ճանաչման համակարգ տպուած թուայնացուած փաստաթղթերի, պատկերների (և այլն) համար և փաստաթղթերի ընկալման համար:

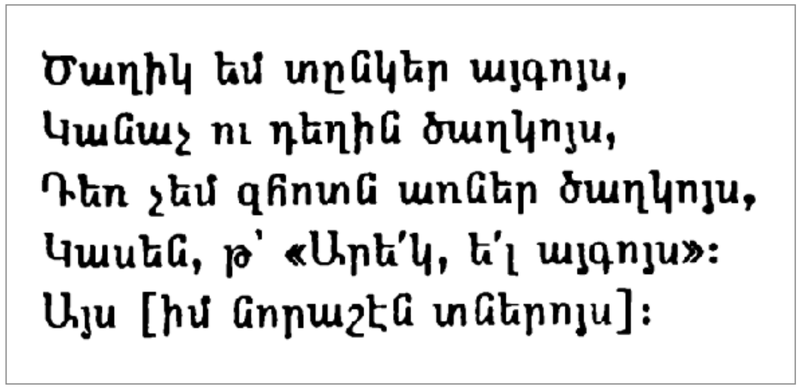
Գրանշանների սխալների տոկոս (Character Error Rare / CER). չափիչ, որն օգտագործում է սխալների տոկոսը չափելու համար (3% CER նշանակում է, որ 100 նիշի համար կայ 3 սխալ):

Կարեւոր բարելաւումներ արհեստական բանագիտութեան կիրառման շնորհիվ

1. **Խնդիրը** համարուում է **լուծուած** (յատկապէս **լատինական գրերի** համար)
2. CER < 2% հեշտութեամբ հասանելի է **ամենատարածուած տառատեսակների և պարզ էջադրումների** համար
3. **Layout analysis**-ը դեռ մարտահրաւէր է

Տպագիր փաստաթղթեր - Գրանշանների օպտիկական ճանաչում (OCR)

Գրիգորիս Աղթամարցի, Տաղեր, *p.109*
(Wikisource)


Ծաղիկ եմ տընկեր այգոյս,
Կանաչ ու դեղին ծաղկոյս,
Դեռ չեմ զհոտն առներ ծաղկոյս,
Կասեն, թ՝ «Արե՛կ, ե՛լ այգոյս»:
Այս [իմ նորաշէն տներոյս]:

Abbyy FineReader PDF (15.0.3)

Ծաղիկ եմ տընկեր այգոյս,
Կանաչ ու դեղին ծաղկոյս,
Դեռ չեմ զհոտն առներ ծաղկոյս,
Կասեն, թ՝ «Արերկ, ել այգոյս»:
Այս (իմ նորաշէն տներոյս^:

CER: 2.25%

Մի քանի կետադրական նշանների բաց թողում

Tesseract (4.0)

Ծաղիկ եմ տընկեր աղգոլս,
Կանաչ ու դեղին ծաղկոլս,
Դեռ չեմ զոոտն առներ ծաղկոլս,
Կասեն, թ՝ «Արեկ, եղ ալգոլս»:
Այս [իմ նորաշէն տներոլս]:

CER: 6.70%

Տառատեսակի նուազ ճանաչում (յ > լ)

Calfa (03/2023)

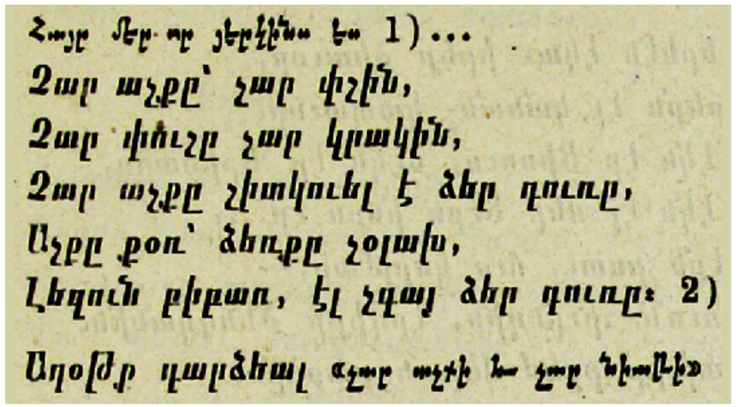
Ծաղիկ եմ տընկեր այգոյս,
Կանաչ ու դեղին ծաղկոյս,
Դեռ չեմ զհոտն առներ ծաղկոյս,
Կասեն, թ1 «Արեկ, ել այգոյս»:
Այս [իմ նորաշէն տներոյս]):

CER: 1.50%

Մի քանի կետադրական նշանների բաց թողում

Տպագիր փաստաթղթեր - Գրանշանների օպտիկական ճանաչում (OCR)

Էմինյան ազգագրական ժողովածու, 2, 1901, p. 183 (FSL of NAS RA)



Abbyy FineReader PDF (15.0.3)

Հէք «ջ էՙ I) ...
Չար աչքը չար փշին,
Չար փեւշը չար կրակին,
Չար աչը շիտկուել է ձեր դուռը>
Աչը քօւռ ձեւռը չօլահհ
Լեզուն րիբառ, էլ չգայ ձեր գսւռը: 2)
Աղօթք դարձեալ Ա

CER: 32.99%

- Անսովոր տառատեսակների և ոճերի թույլ ճանաչում

Tesseract (4.0)

Հայր մէր ոք յե կա Է» 1)...
Չար աչքը՝ չար փշին,
Զար փուչը չար կրակին,
Զար աչքը չիտկունլ է ձեր դուռը,
Աչքը քօո՝ ձեռքը չօլախ,
Լեղուն բիրառ, էլ չդայ ձեր դուռը: 2)
Աղօխը դարձհալ «օր «չի Է չոր Ֆիանի»

CER: 14.72%

- Շատ զգայուն աղմուկի նկատմամբ
- Անսովոր տառատեսակների կամ ոճերի բաց թողում

Calfa (03/2023)

Հայր մեր որ յերկինս էս 1)...
Չար աչքը՝ չար փշին,
Չար փուշը չար կրակին,
Չար աչքը չիտկուել է ձեր դուռը,
Աչքը քօռ՝ ձեռքը չօլախ,
Լեզուն բիբառ, էլ չգայ ձեր դուռը: 2)
Աղօթք դարձեալ «չոր աչքի աս չար նիաթի»

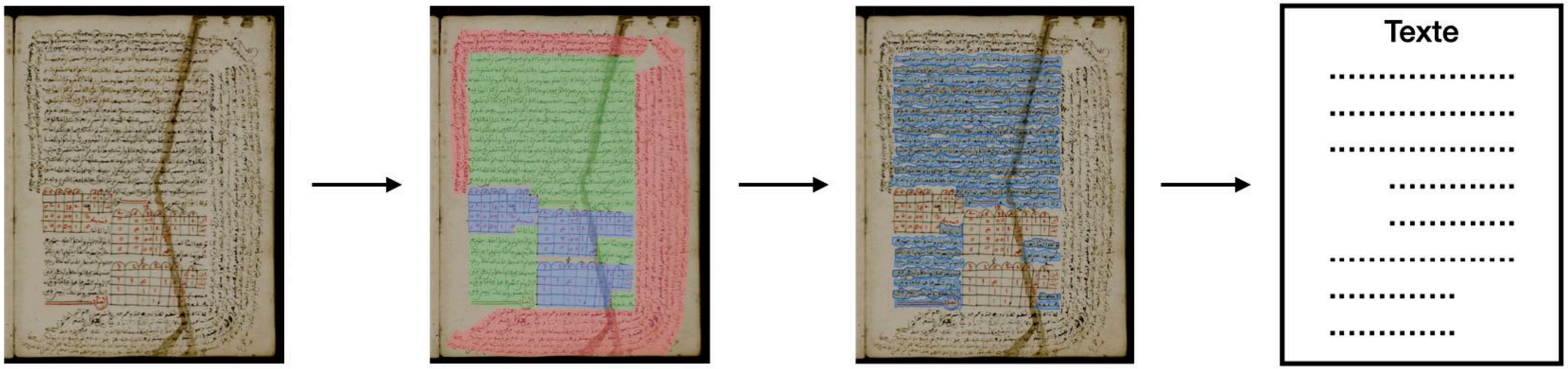
CER: 2.03%

Ձեռագիր փաստաթղթեր - Ձեռագիր տեքստի ճանաչում (HTR)

HTR. Ձեռագրերի և փաստաթղթերի ընկալման համար տեքստի ճանաչման համակարգ

Ուսումնասիրութեան բաց դաշտ. արհեստական բանագիտութեան օգնութեամբ (նեյրոնային ցանցեր) ժամանակակից համակարգեր, որոնք «հեշտութեամբ» կարող են հասնել 5% CER-ի, յատկապէս լատինատառ գրութիւնների համար

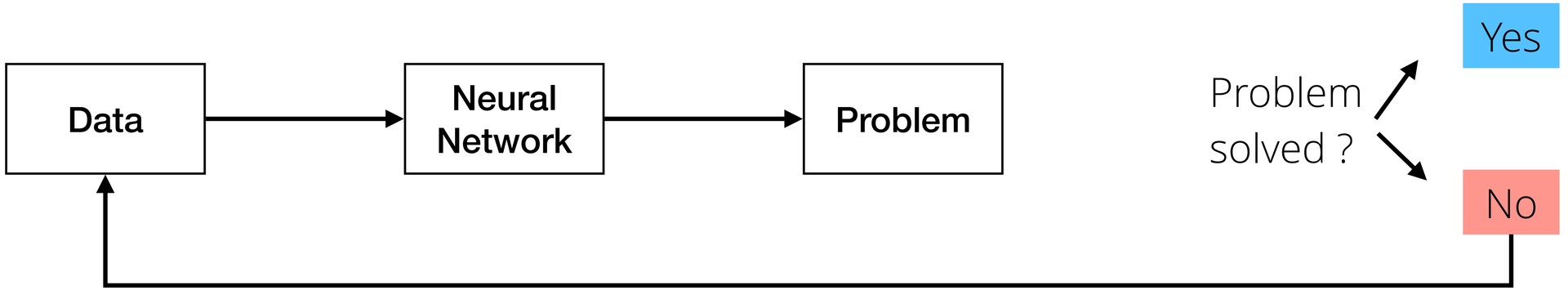
HTR-ն այժմ հասարակաց փուլ է ցանկացած Թուային Հումանիտար գիտութիւնների նախագծում լատինատառ գրութիւնների համար



BULAC MS.ARA.609, p. 124, RASAM Dataset

Ձեռագիր փաստաթղթեր - Ձեռագիր տեքստի ճանաչում (HTR)

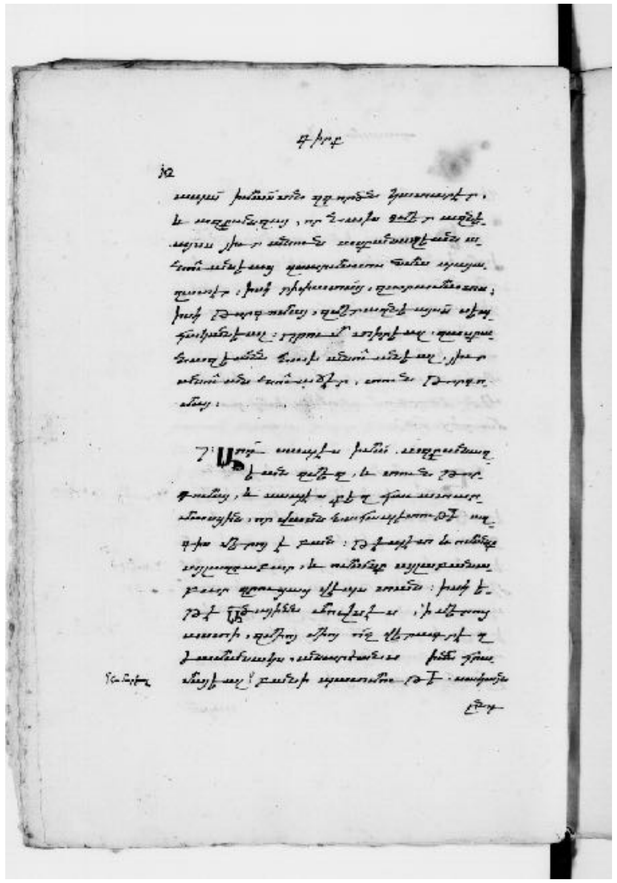
Արհեստական բանագիտութիւնը մշակել օրինակի միջոցով



Ի՞նչ է պատահում, եթէ իմ փաստաթուղթը վնասուած է / ձեռագիրը չափազանց յատուկ է / էջադրումը չափազանց բարդ է / եթէ կարելի չէ աւելի շատ տուեալներ տրամադրել:

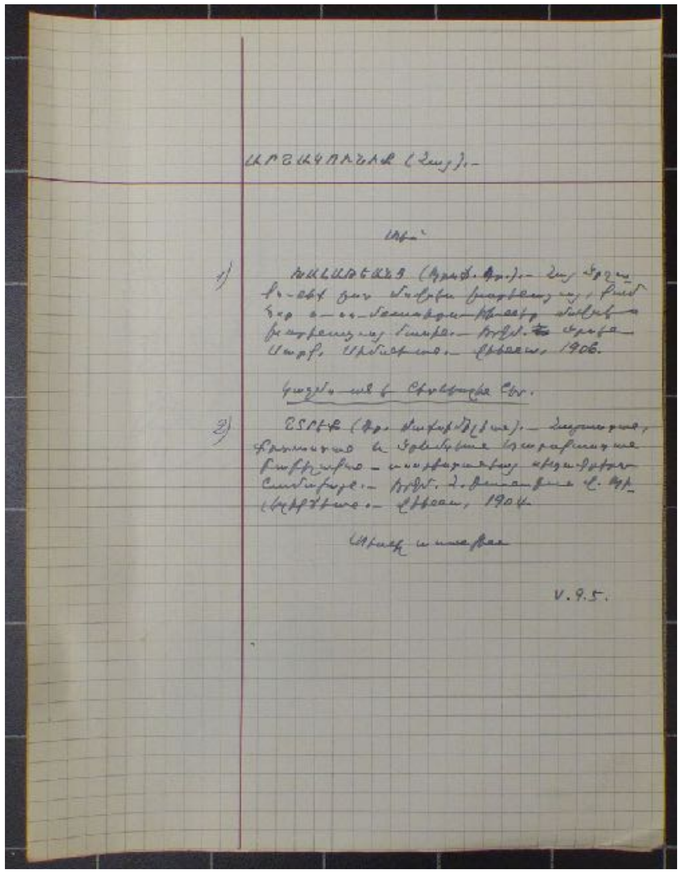
Ընդհանուր մօտեցումը պատշաճ չէ: Պահանջում է մասնագիտացուած մօտեցում:

Մեր ծրագիրները (2023-2024)



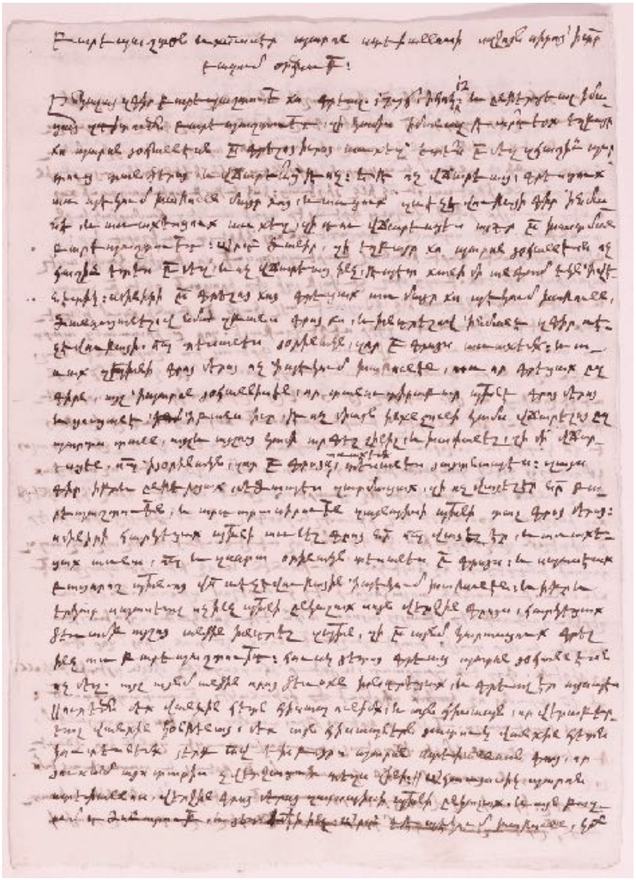
6.000 էջ
Searchable on Gallica (soon)

BnF



60.000 էջ
Searchable on Intranet

ՆՈՒՊԱՐ
NUBAR
BIBLIOTHÈQUE - LIBRARY

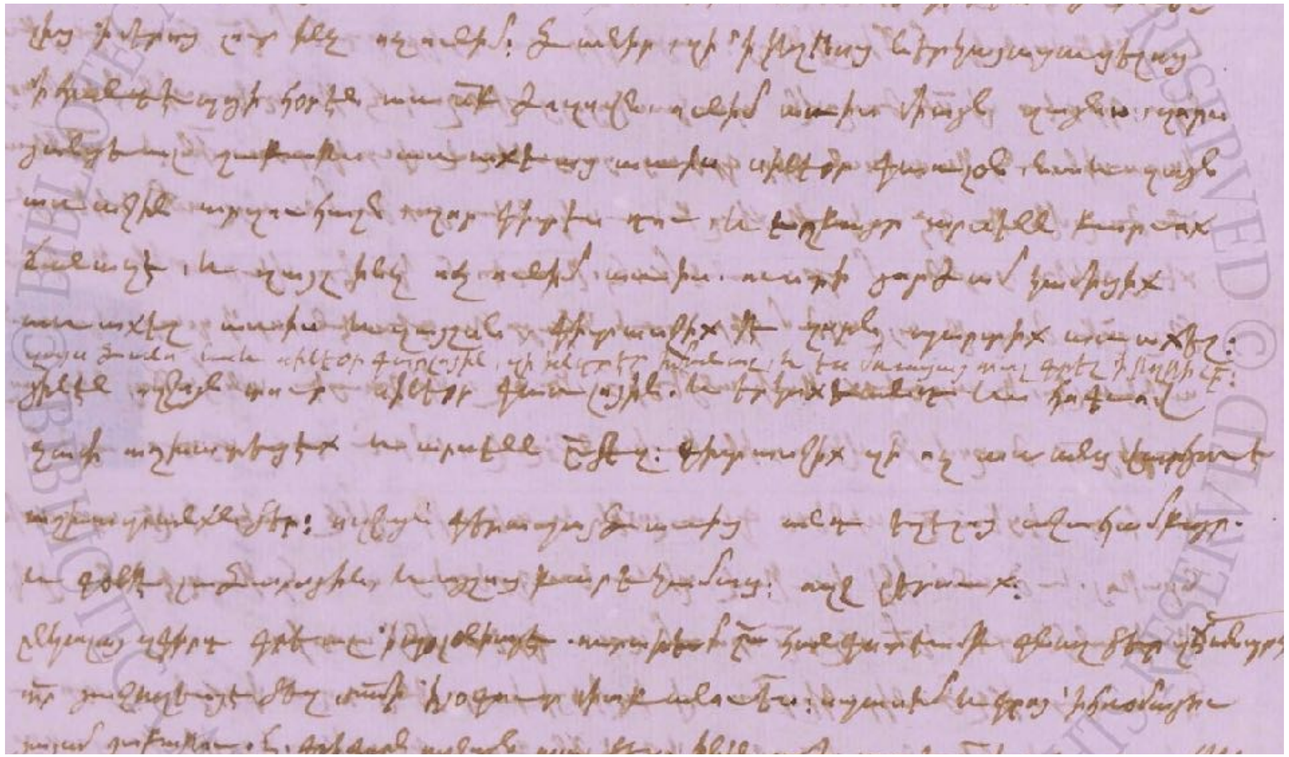


20.000 էջ
Soon available

MEKHITAR

Գրանշանների ճանաչում. օրինակներ

Մխիթարեան միաբանութեան ձեռագիր նամակների մշակումը
Հ. Վահանի և Մխիթարեան միաբանութեան գործակցութեամբ:



Melkonian letter, 1759, Mekhitarist Congregation

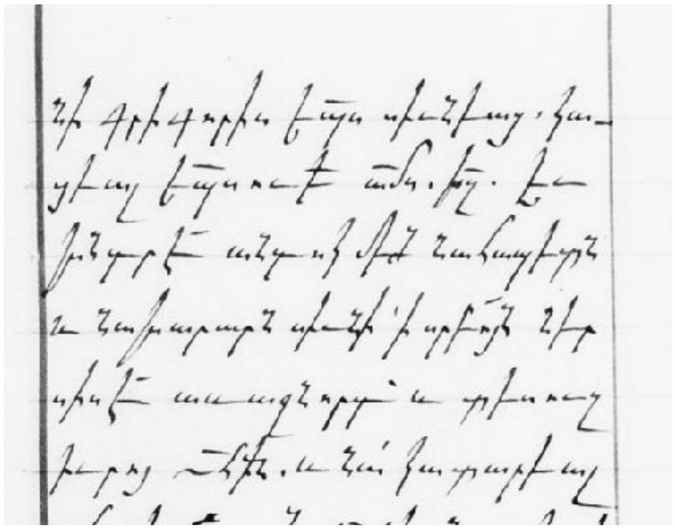
լից ի մերոց լուր ինչ ոչ ունիմ: Ծանիր, զի ի թղթոց ներկայացուցելոց հանցի հօրէն առ սբ ժողովն ունիմ առ իս միայն զայնս, զորս յանցել շաբաթս առաքեաց առ իս սինեօր Գառլօն, նաև զայն առաջին արզուհլն, զոր գիտես դու, և եղբայր Արսէնն բարւոք ճանաչէ, և զայլ ինչ ոչ ունիմ առ իս, ուստի յորժամ կամիցիք առաքել առ իս և զայլսն, գիտասջիք թէ զորն պարտիք առաքեալ: զայս ժամու նաև սինեօր Գարլոյին, զի խնդրէր խնանալ, և ես մոռացայ տալ գրել ի թղթի ի:
յինէն ողջոյն տուր սինեօր Գառլոյին, և երկոքեանդ ևս հոգւով չափ աշխատեցէք և Արսէնն ըստ ձեզ: Գիտասջիք զի ոչ առանց վարձու է աշխատանքն ձեր: Ողջոյն գերապայծառից անդ եղելոց աջահամբայր,

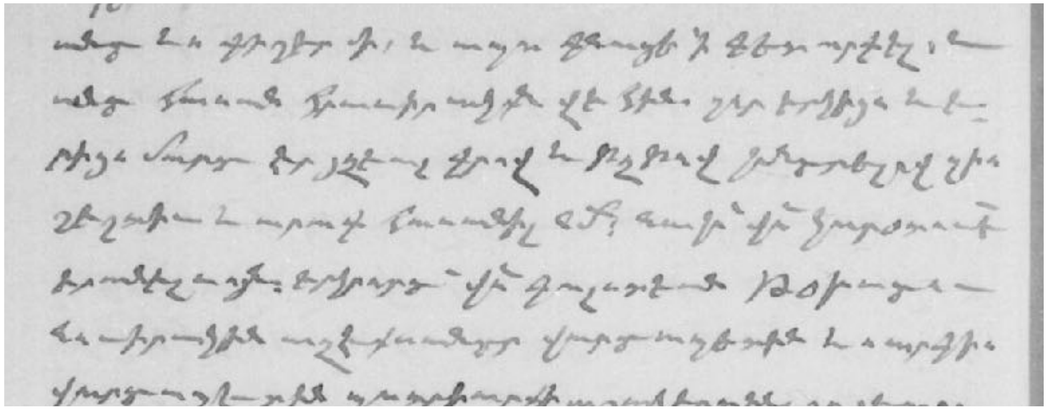
Թանաքի թափանցիկութիւն	Հապաւումները կարդացուում են	96.1%

Խնդիրը լուծուած է

Գրանշանների ճանաչում. օրինակներ

Dulaurier նախագիծը Ֆրանսիայի Ազգային գրադարանի հետ





նի գրիգորիս եպ[իսկոպո]ս սինւեաց կա
ցեալ եպ[իսկոպո]սու[թ]ե[ան] ամ[են]ս խ[ո]զ. Եւ
խնդրէ անդոկ մեծ նահապետն
և նախարարն սիւնի ի սրբոյն ներ
սիսէ առաջնորդ և տեսուչ
իւրոյ [աշխար]հին, և նա կատարեալ
անդ ևս գրշեր մի, և ապա գնացի ի գետարկել: և
անդ հասան հրաւիրակքն վեհին, զոր երկիցս և ե-
րիցս մարդ էր յղեալ գրով և թղթով խնդրելով զիս
շեշտիւ և արագ հասանիլ նմ: Նախ վ[ե]ր[այ] կարօտու[թ]ե[ան]
երանելւոյներկրորդ՝ վ[ե]ր[այ] գալստեան Թօխադու
նուիրակին աղէքսանդր վարդապետին և սարգիս

98.75%

Սկանաւորուած միկրոֆիլմեր

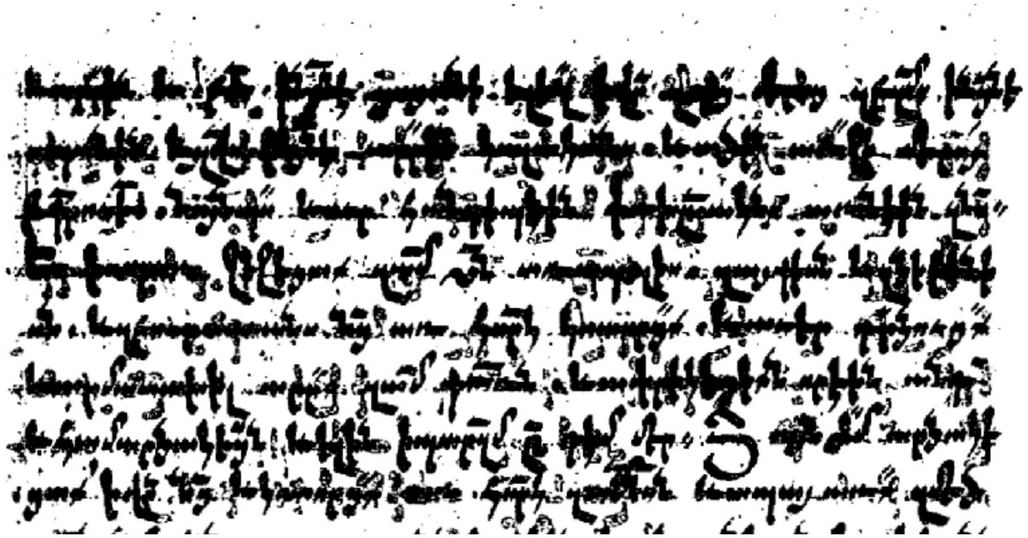
Հապաւումները կարդացուում են

96.12%

Խնդիրը լուծուած է

Գրանշանների ճանաչում. օրինակներ

Ֆրանսիայի ազգային գրադարանի աւրուած հին ձեռագիրների (միկրոֆիլմերի) մշակում:



P191, Smbat Vardapet, French National Library

և որդիX և բժ. խնդնէ յացտնի ելել ի վր զօջց մերոց զբզմս ի նցնէ սպանին: և յզվա ի նցնէ արրին կալլկանց. և ածին առջի մերոյ թգրուեX նոյնպս և ուր հանդիպէին փախըXտկն առնէին զնս: Եւ իսպառ ջնջեցաք զամ ~ն տռապօլիս զայգիսն և զձիբենի սն. և զբուրաստանս նց առ հսրկ կոտորցք. և աւեր դիձուցք և արմատախիլ արրք զամ գւռն. և ափրիկեցիքն որ էին անդ և համարձակեցն և ելին ի պտրզմ ը դէմ մեր: Յայն ժմ արձակէ ցաք ի վր նց և կոտորցք XX հսրկ զամնսն և ապայ առք զմեծ

Աղմուկը հաղթահարուած է

Հապաւումներ չկան

95.1%

Խնդիրը լուծուած է

Ընդհանուր OCR-ի մոդել ընդդեմ մասնագիտացուած OCR-ի մոդել

Իմ փաստաթուղթը **պարզ** է (կանոնաւոր փաստաթղթեր, հասարակ էջադրում - layout)

Իմ փաստաթուղթը **բարդ** է

(ձեռագիր, էջի ձեւաւորում - layout, թուայնացման որակ և այլն)

Պատրաստ է OCR-ի օգտագործման համար

Պահանջուում է **յատուկ OCR-ի մոդել**

Տուեալ

- օգտագործողը
- բաց աղբիւր
- crowdsourcing
- e.g. Calfa

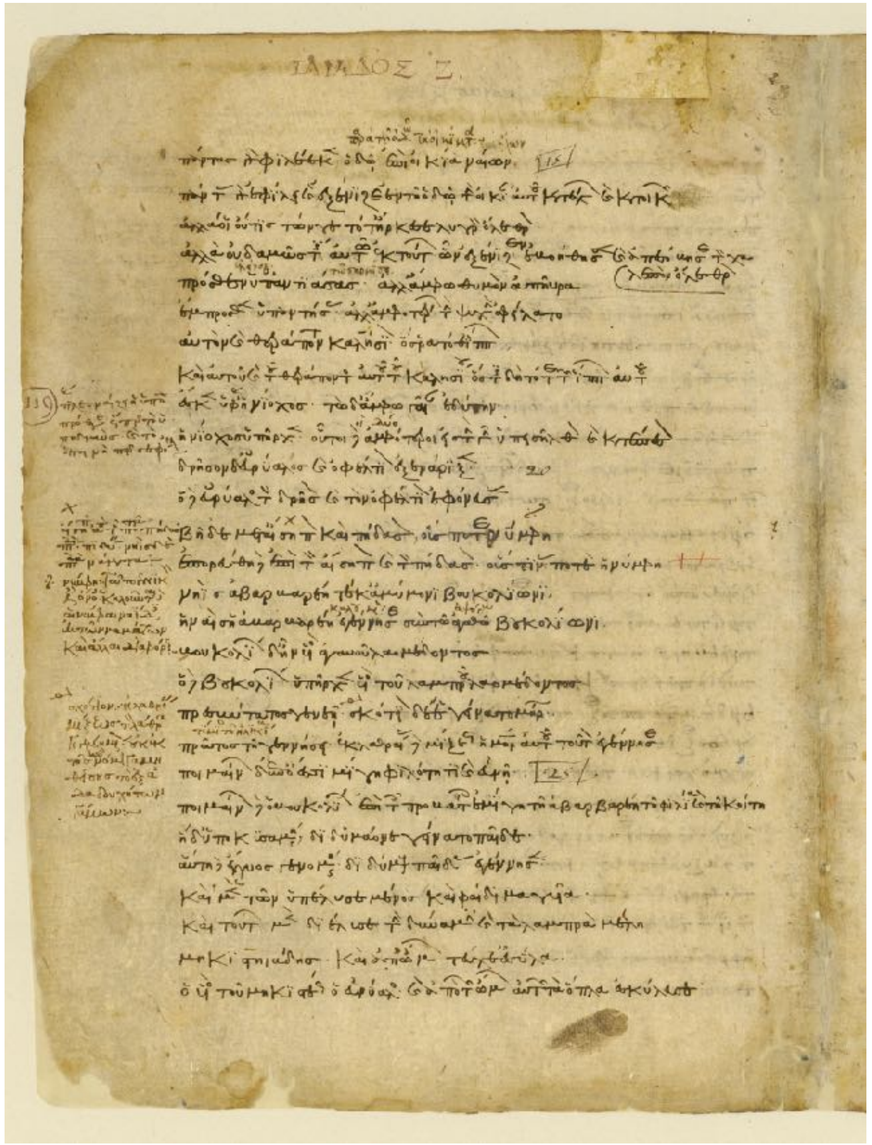
Վարժեցնել

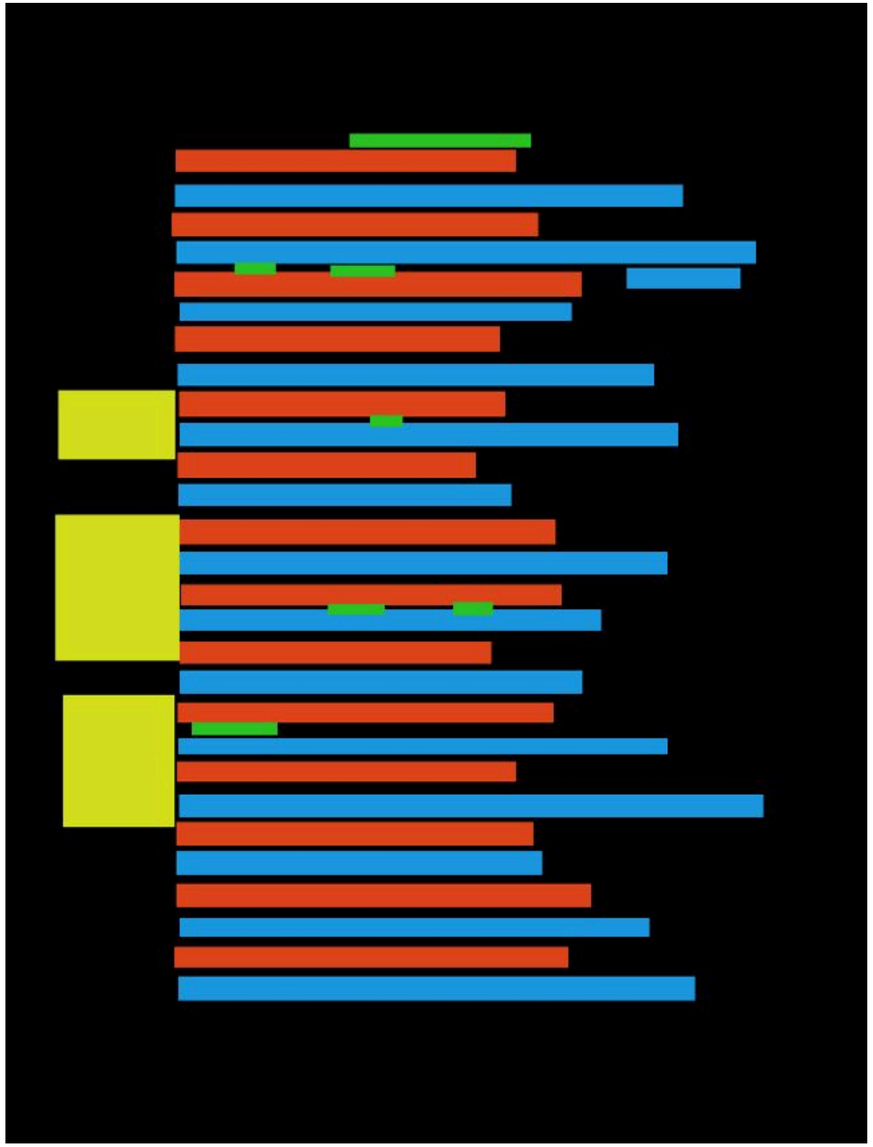
- e.g. Calfa

Արհեստական բանականութիւնը կարող է աշխատել միայն այն դեպքում, եթէ մենք գիտենք, թե ինչին ենք ուզում հասնել: Գրադարանավարը գտնուում է գործընթացի կենտրոնում, քանի որ նա է որոշում կարիքը:

Գրադարանների դերը - էջի ձեւաւորում (layout)

Ցանկացած նախագիծ, ներառեալ AI-ն և OCR-ը, սկսւում է կարիքների և խմբագրական ընտրութիւնների սահմանմամբ (մարդու որոշումն է)





gloss

main text

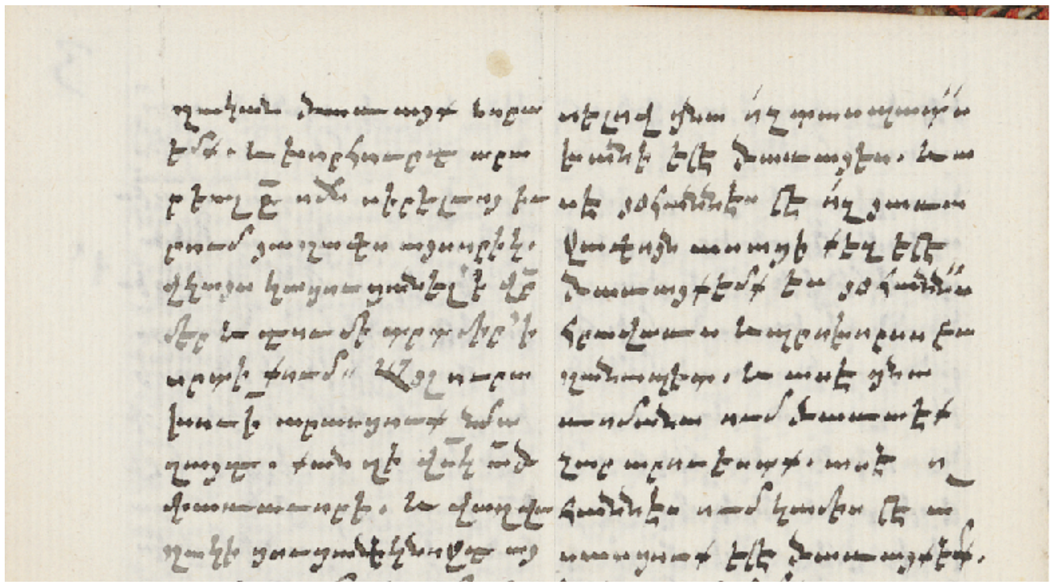
commentary

interlinear text

https://www2.unil.ch/iliade

Գրադարանների դերը - Տեքստ (text)

Ցանկացած նախագիծ, ներառեալ AI-ն և OCR-ը, սկսւում է կարիքների և խմբագրական ընտրութիւնների սահմանմամբ (մարդու որոշումն է)



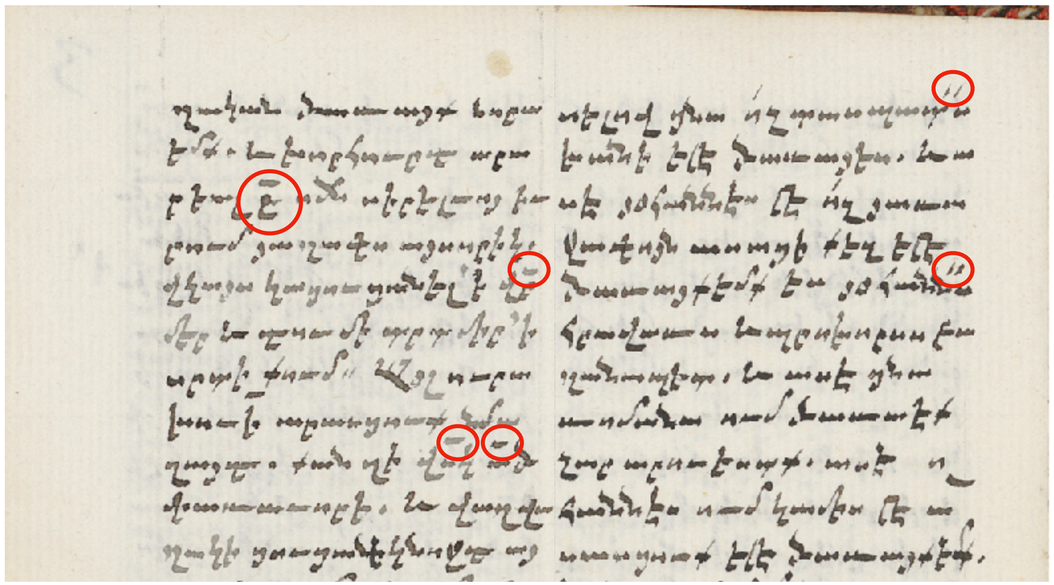
ms M5, f. 3b (Matenadaran, Erevan)

ղականծառայքնորասելովցնաոչտասպատս
եմք.ևխորհուրդարախանիեթէծառայես.ևա
րեալըոմնսիրելոյիւսէյօհաննէսթէոչյառա
րումյաղագսայսորիկ.ջագոյնասացիքեզեթէ
վկայսկացուցանել'իվրծառայքեմքեսյօհաննս
մերևդումիտրտմիր'իհրավառսևպրոխորսսբա
սրտիքում..Այլուրաղանապետ.ևասէցնա
խուիարասցուքնմառոմանաումծառաէք
զայդ.քանզիվսկածչարարուեստք.ասէո
միառաւորի.ևվաղվահաննէսումկամիսթէա
ղակիցուցանէկնոջդայսասցուքեթէծառայքեմք.

CER = 0%, բայց ինչ ասել էջի ձեւաւորման (layout), հապաւումների, տողադարձի և բառերի մասին:

Գրադարանների դերը - Տեքստ (text)

Ցանկացած նախագիծ, ներառեալ AI-ն և OCR-ը, սկսւում է կարիքների և խմբագրական ընտրութիւնների սահմանմամբ (մարդու որոշումն է)



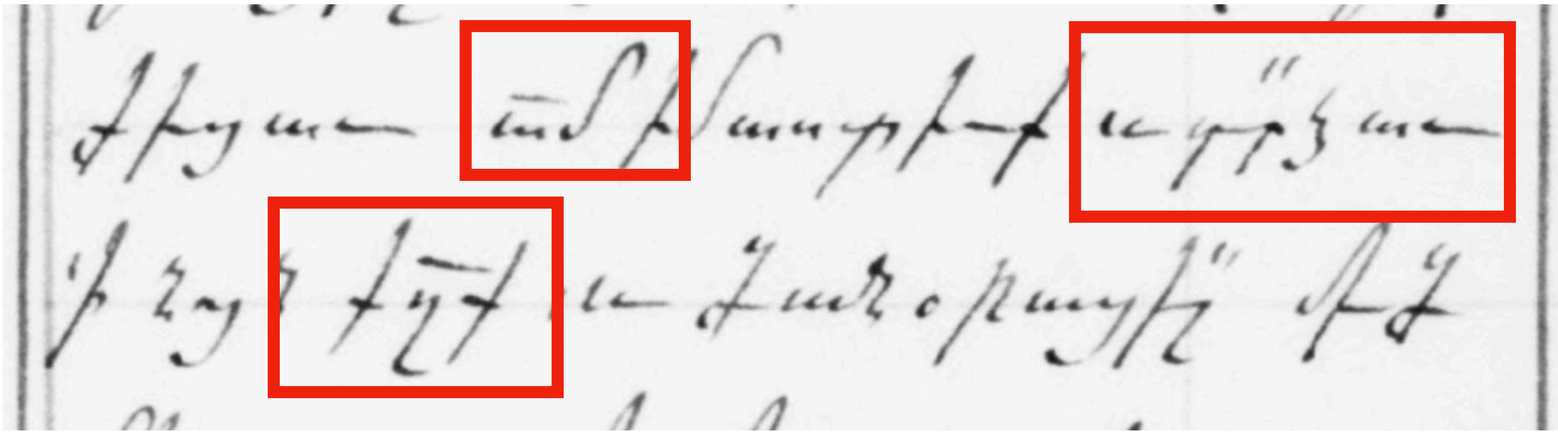
ms M5, f. 3b (Matenadaran, Erevan)

ղականծառայքնորա
եմք.ևխորհուրդարա
րեալը[ստ]ոմնսիրելւոյիւ
րումյաղագսայսորիկ.
վկայսկացուցանել'իվ[ե]ր[այ]
մերևդումիտրտմիր'ի
սրտիքում..Այլուրա
խու[թ]ի[ւն]արասցուքնմա
զայդ.քանզիվ[ա]ս[ա]կա[ստուա]ծ
միառաւորի.ևվաղվա
ղակիցուցանէկնոջդայ

սելովցնաոչտասպատ[ա]ս
խանիեթէծառայես.ևա
սէյօհաննէսթէոչյառա
ջագոյնասացիքեզեթէ
ծառայքեմքեսյօհանն[է]ս
հրավառսպրոխորսսբա
ղանապետ.ևասէցնա
ռոմանաումծառաէք
չարարուեստք.ասէո
հաննէսումկամիսթէա
սացուքեթէծառայքեմք.

Գրադարանների դերը - Տեքստ (text)

Ցանկացած նախագիծ, ներառեալ AI-ն և OCR-ը, սկսւում է կարիքների և խմբագրական ընտրութիւնների սահմանմամբ (մարդու որոշումն է)



- ամ [ենայն]
- դ [ա] րձաւ
- ք [ա] ղ [ա] ք

- 1 տառադարձում = մի քանի տարբեր գրաֆիկական ձեւեր

- 1 գրաֆիկական ձեւ = մի քանի հնարաւոր տառադարձումներ

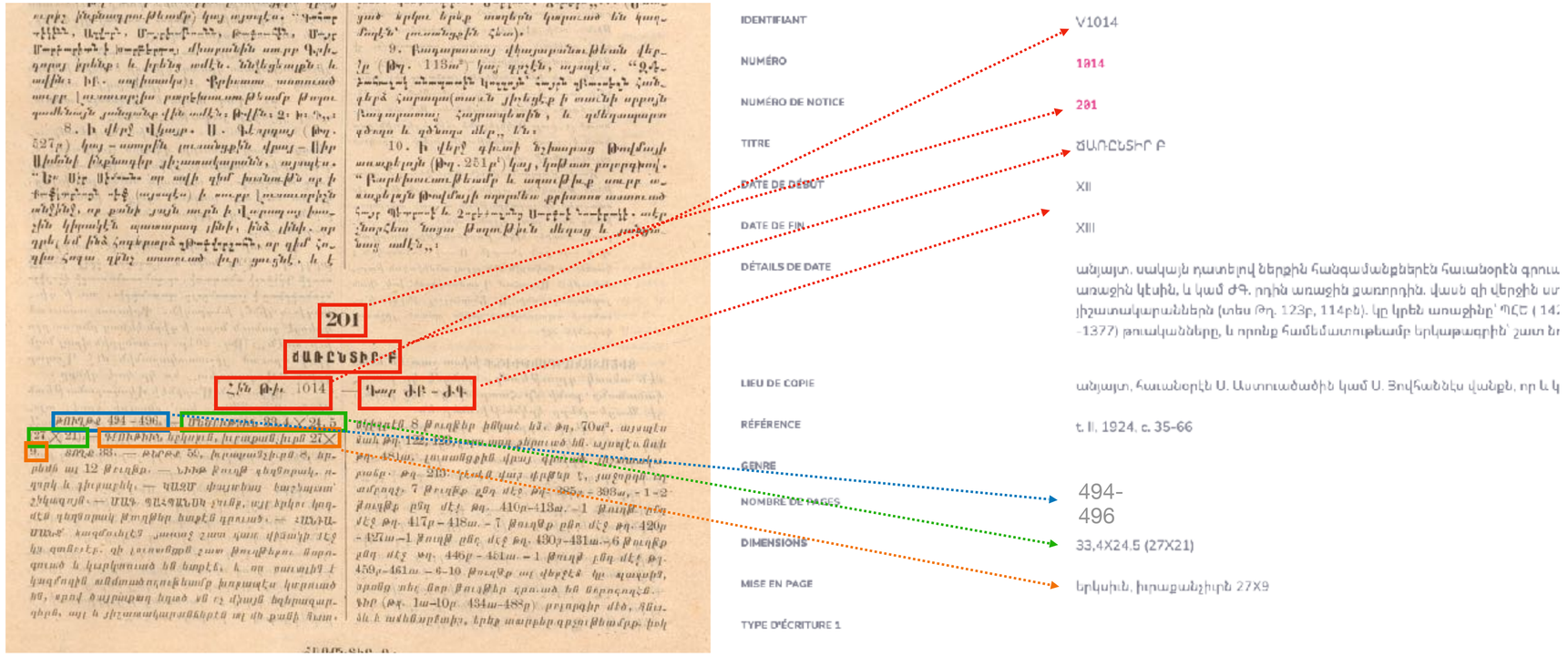
Արհեստական բանագիտություն և փասթաթղթեր

կատալոգներ • լուսանկարչական ֆոնդեր • գիտելիքների արդիունահանում

Գրանշանների ճանաչումի հնարաւորութիւնները ընդլայնել

Սուրբ Ղազարի ձեռագիրների քարտարանի թուային տարբերակի ստեղծումը
Մխիթարեան միաբանութեան համագործակցութեամբ:

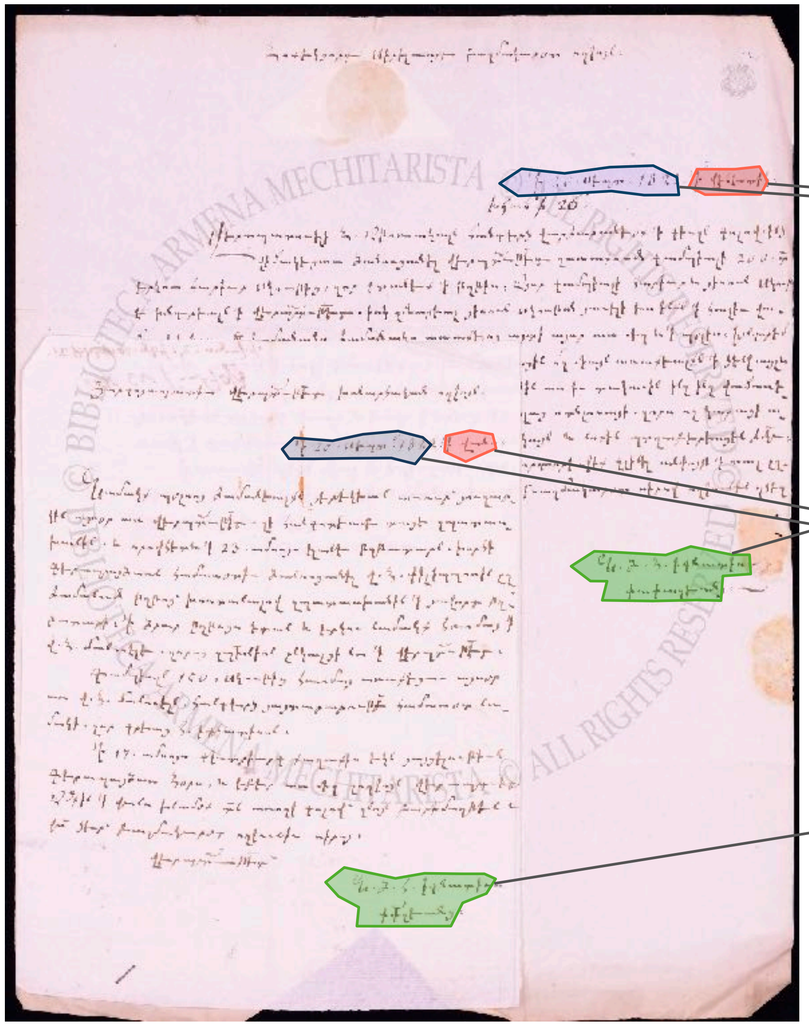
- Մասնագիտացուած Արհեստական բանականութիւն էջի ձեւաւորման, տառատեսակի և իմաստաբանութեան համար
- Միջին ճշգրտութիւնը՝ 99,5%
- Առցանց տուեալների շտեմարանի ինքնագործ լրացում



Beta version : https://catalog.mechitar.org/

Գրանշանների ճանաչումի հնարաւորութիւնները ընդլայնել

Ցուցակագրել Մխիթարեան միաբանութեան նամակագրութիւնները
Հ. Վահանի և Մխիթարեան միաբանութեան գործակցութեամբ:



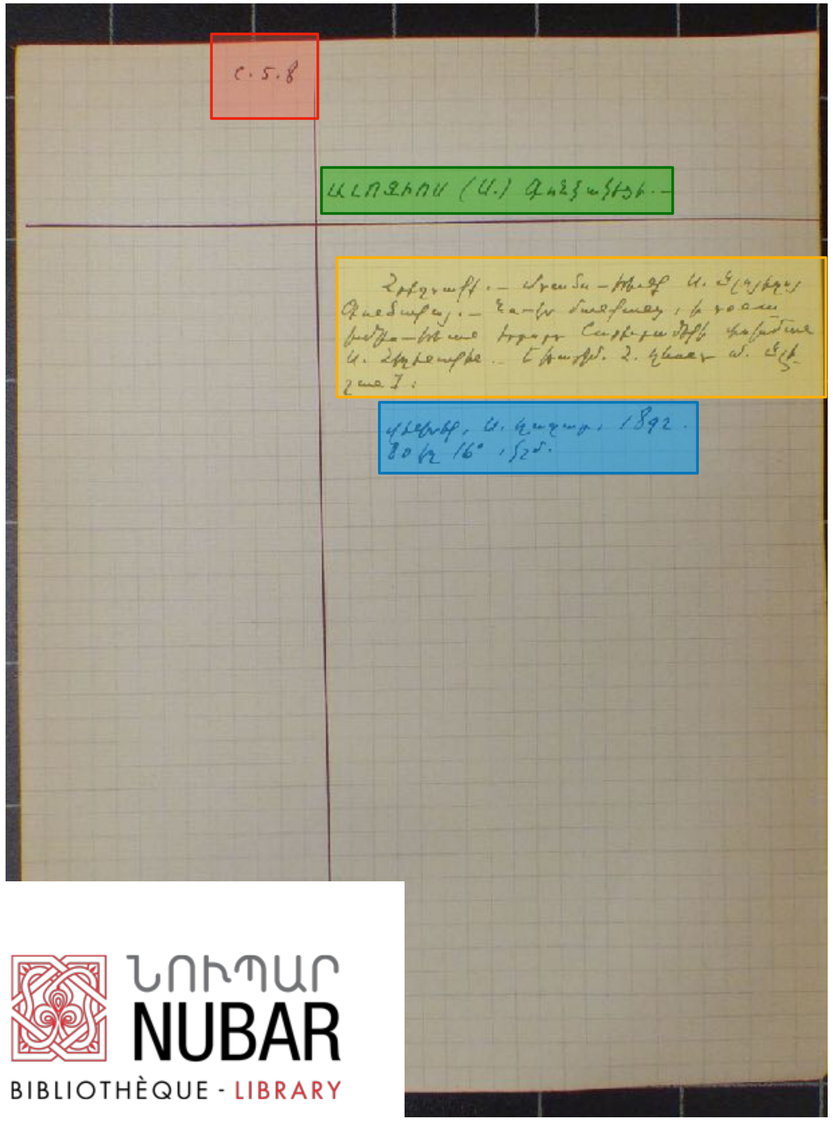
Թուական: 20 Սեպտեմբեր 1822
Վայր: Վենետիկ
Հեղինակ: Իգնատիոս Փափազեանց

Թուական: 20 Սեպտեմբեր 1822
Վայր: Վանս
Հեղինակ: Իգնատիոս Փափազեանց

Ongoing

Գրանշանների ճանաչումի հնարաւորութիւնները ընդլայնել

ՀԲԸՄ-ի Նուբար գրադարանի (Փարիզ) մատենագիտական գրառումների վերածումը
փնտրման համակարգի



OCR

Catalog number: Ը.5.8
Author Name: ԱԼՈՋԻՈՍ (Ս.) Գոնձակեցի
Content: Հրեշտակք. Մտածութիւնք
Ս. Ալոյեզոյ
Գանձակայ. Նուէր մանկանց, ի
տօնա խմբութեան երրորդ
հարիւրամեկի փոխման Ս.
Հեղինակին. [թարգմ. Հ. Ղևոնդ Մ.
Ալիշան]

Text Analysis

վենետիկ Ս. Ղազար 1892
80 էջ 16°, կզմ

ՆՈՒՊԱՐ
NUBAR
BIBLIOTHÈQUE - LIBRARY

Latin catalog done in 2023, Armenian catalog < 1986 done
Available in Nubar Library intranet

Գրանշանների ճանաչումի հնարաւորութիւնները ընդլայնել

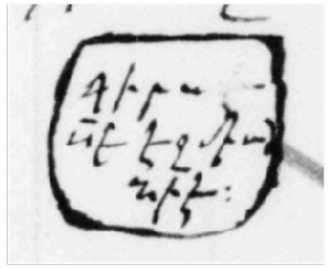
Զարգացնել լուսանկարչական ֆոնդերի հասանելիութիւնը

Ongoing

source image: Wikimedia Commons

Գրանշանների ճանաչումի հնարաւորութիւնները ընդլայնել

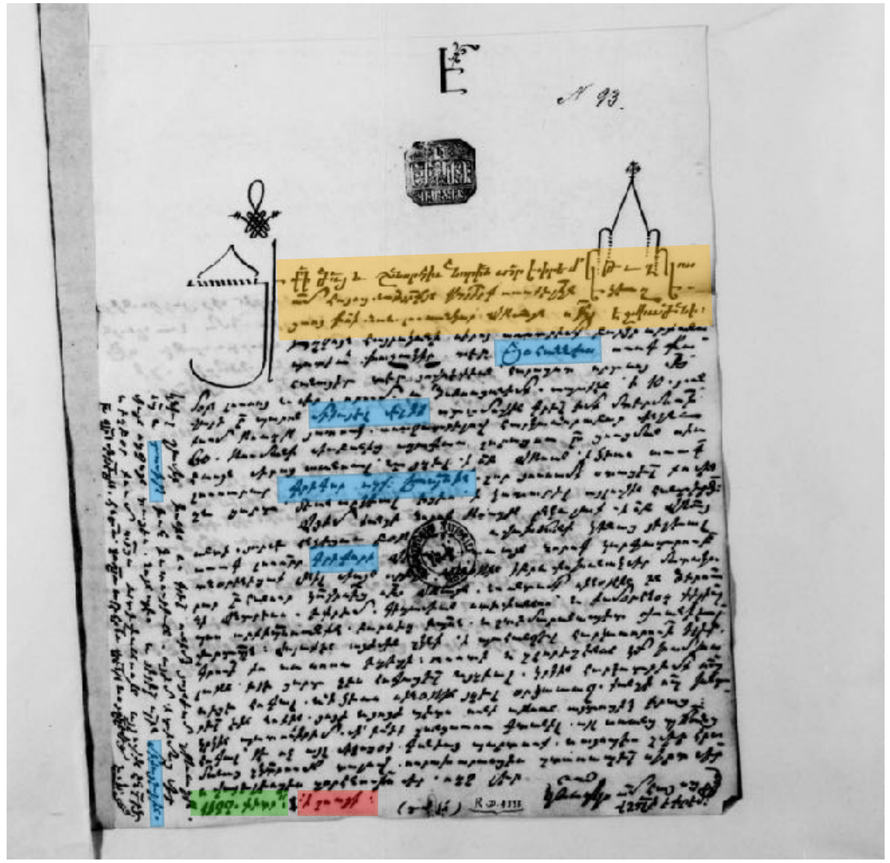
Առարկաների ճանաչում ձեռագիր արխիւներում (Dulaurier, BnF Datalab նախագիծ, 2023)

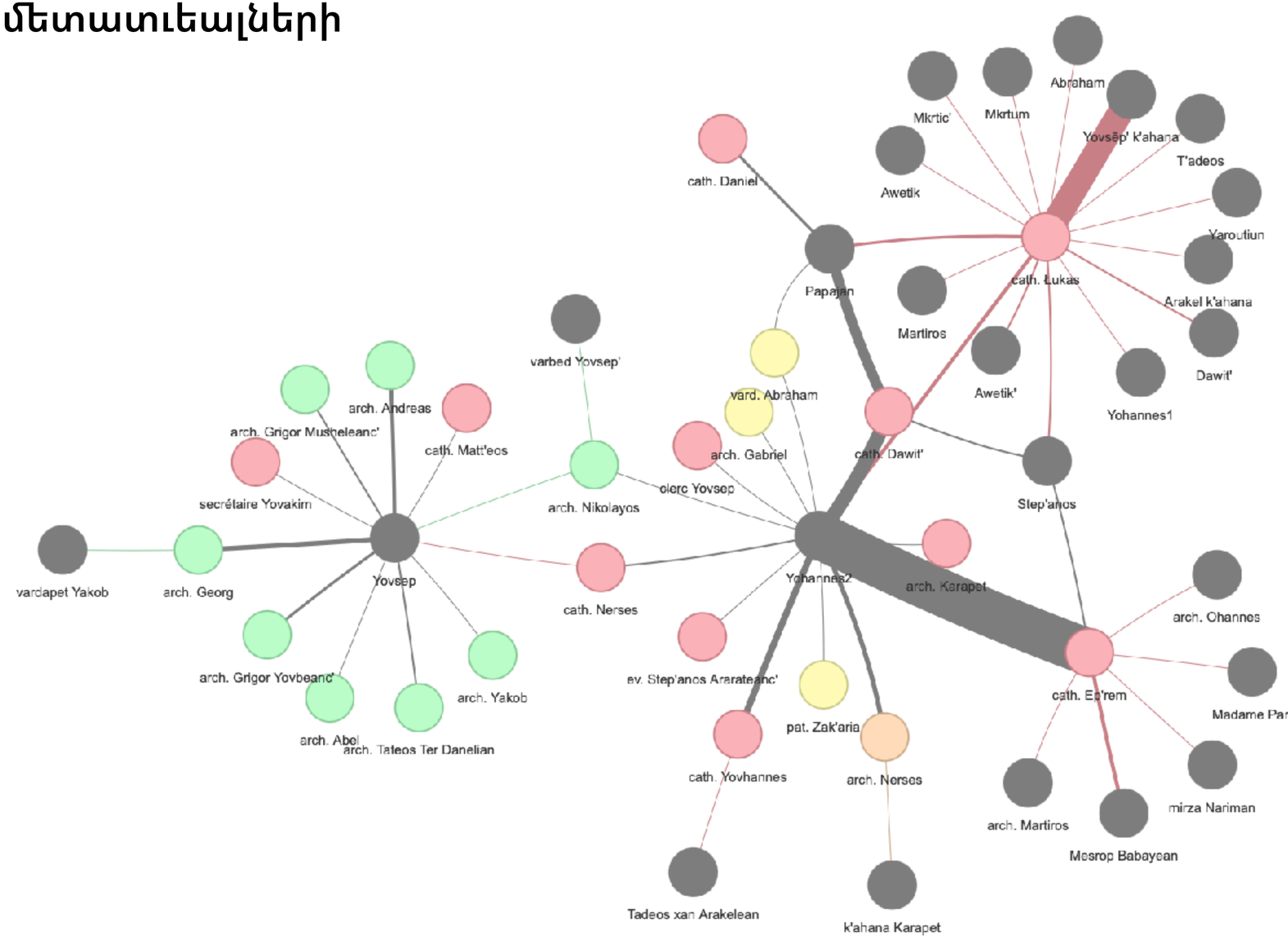


To know more: calfa.fr/blog/39

Գրանշանների ճանաչումի հնարաւորութիւնները ընդլայնել

Անուանուած յատկանիշների և մետատւեալների արդիւնահանում





BnF

CALFA

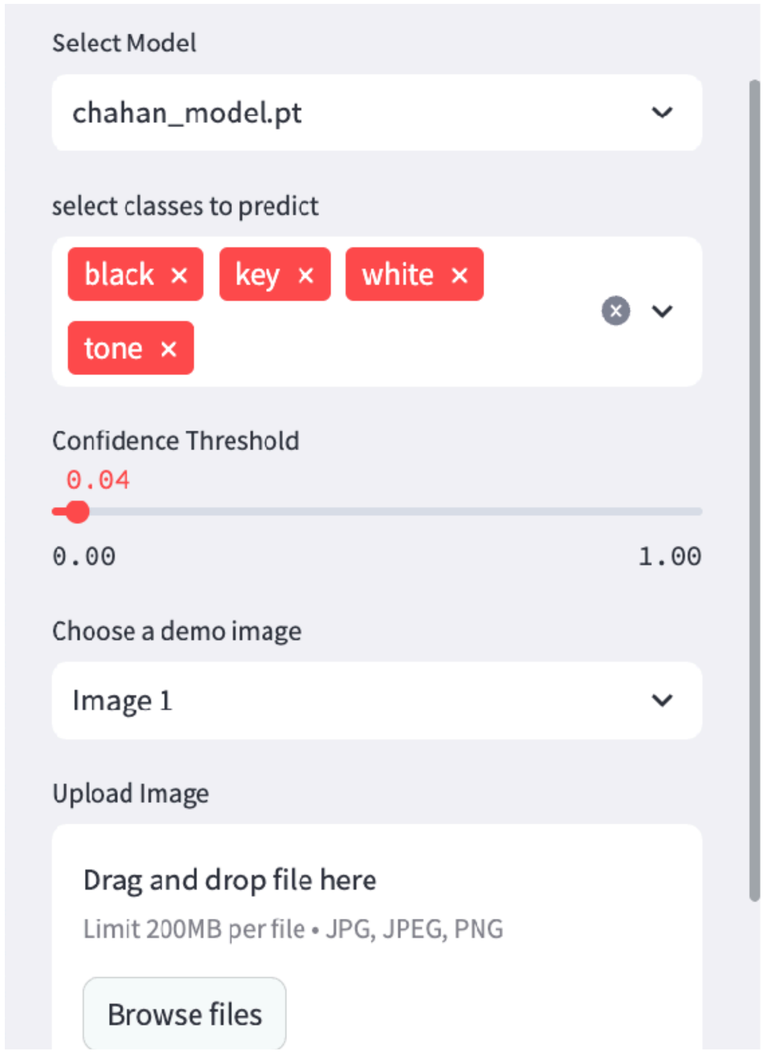
Գրանշանների ճանաչումի հնարաւորութիւնները ընդլայնել

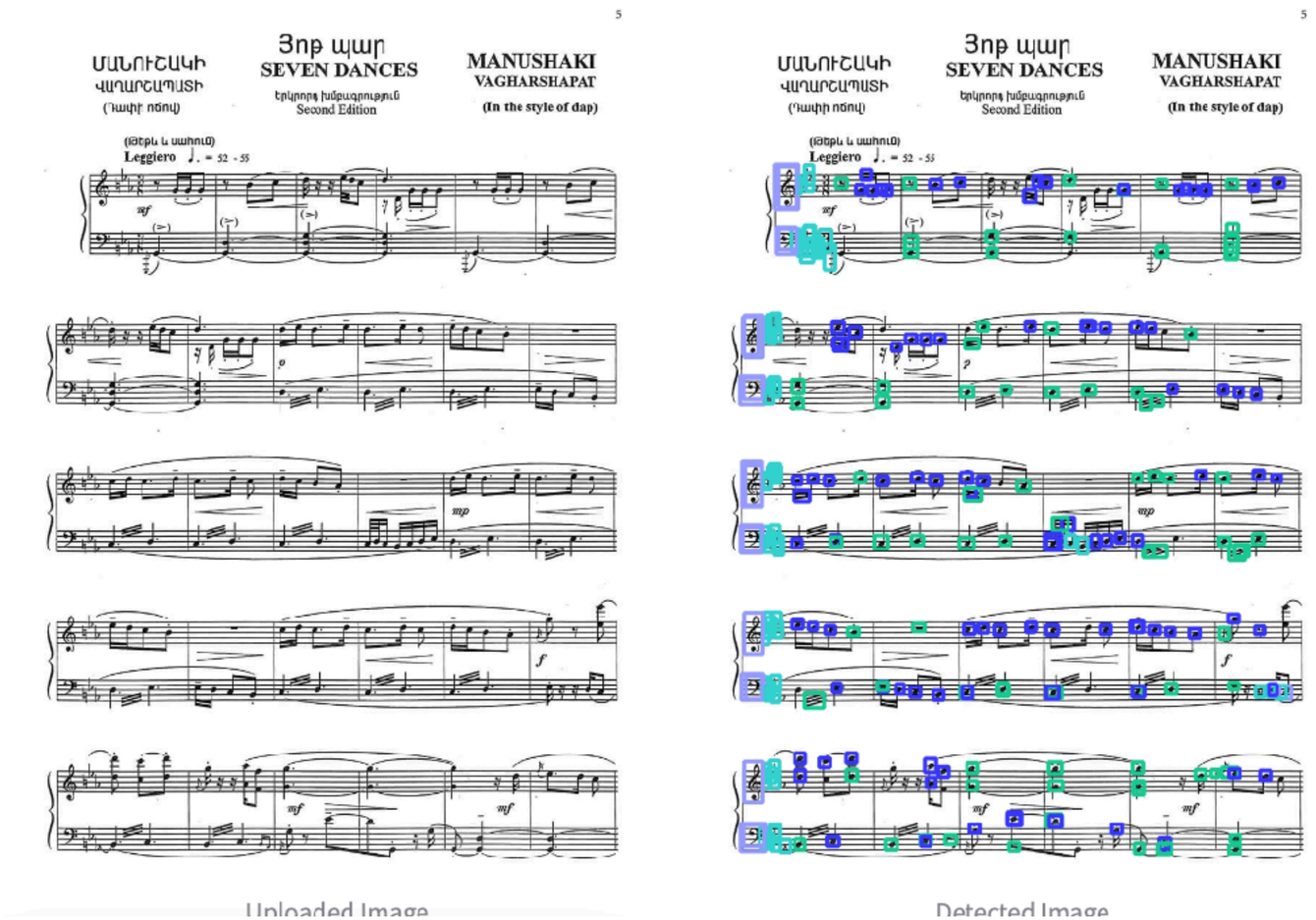
Գովազդի արդիւնահանում հայկական մամուլում

HYE-Haratch_19-1934-04_0092, ARAM, BULAC

Գրանշանների ճանաչումի հնարաւորութիւնները ընդլայնել

ԹՈՒՄՈ աշխատանոց (Մարտ 2024) - Երաժշտութիւն կարդալ և ծրագրաւորել AI-ի միջոցով



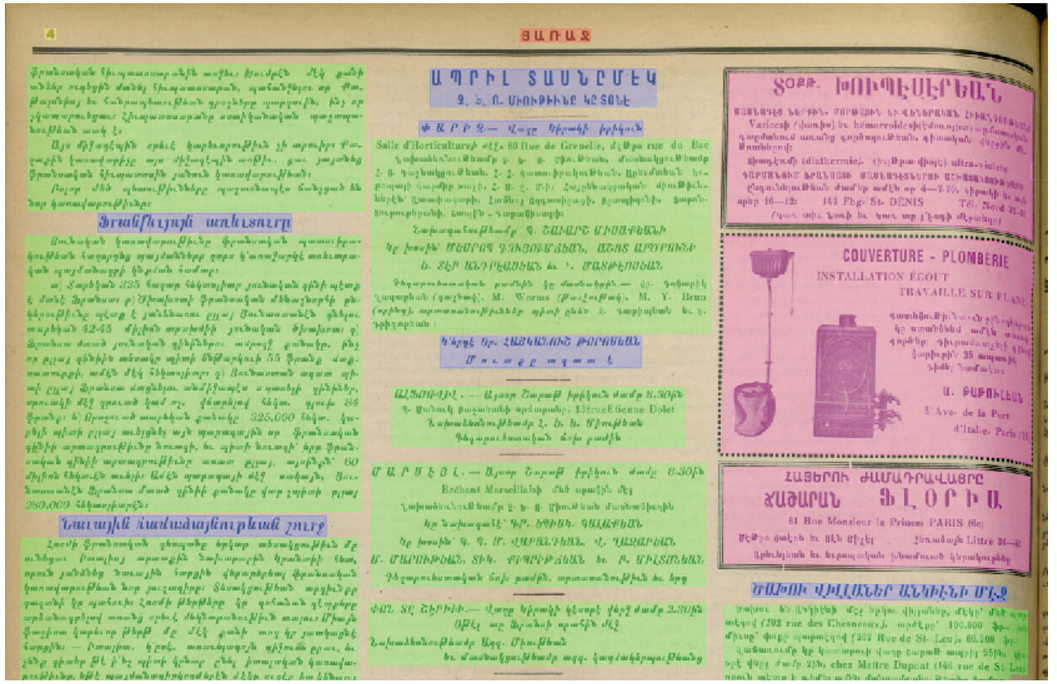


Հեռանկարներ

1. OCR / HTR-ի ճշգրտութիւն հայերէնի համար > 95% մասնագիտացուած մօտեցումներով
2. AI-ն որպէս որոշումների աջակցման/օգնութեան գործիք օգտագործողների և գրադարանավարների համար
3. Հետազոտողներն ու գրադարանավարները AI-ի նոր մոդելների մշակման և սահմանման գործընթացի հիմքում են
4. Մասնագիտացուած մոդելներով տուեալների լայն հաւաքոյթի (datasets) մշակում, ճանապարհ բացելով տուեալների հետազոտման (Data Mining) համար

Նոր նախագիծ 2024 թ. - Հ. Սամուէլեան արևելեան գրախանութ

Հայկական թերթերի թուայնացում և OCR (~ 15k էջ)



FONDATION CALOUSTE GULBENKIAN

CALFA

CALFA

Շնորհակալութիւն

To know more about Calfa and our solutions : https://calfa.fr

Contact us : contact@calfa.fr