

Historical Text Datafication and Loss: Computational Recovery of Typographical Layout Logic on an RDF Graph Featuring ML Methods

ASIS&T SIG AI Workshop | October 29, 2022

Huapu Liu

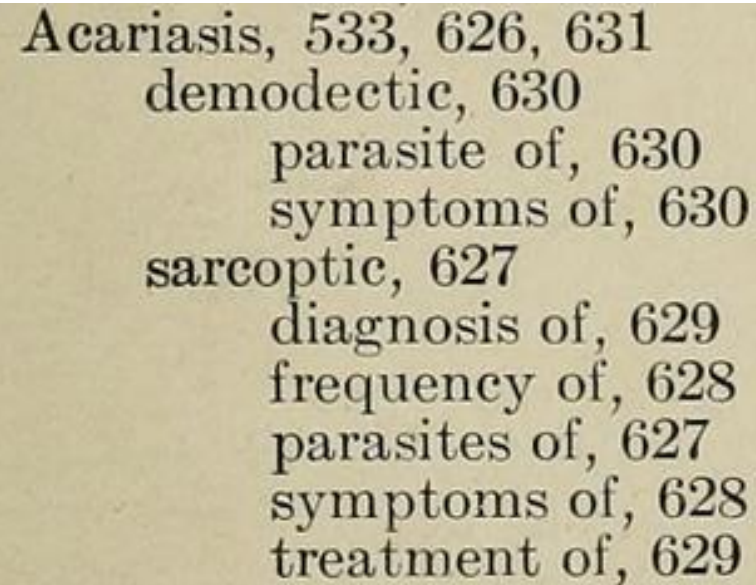
University of Alabama, USA - hliu68@crimson.ua.edu

Steven L. MacCall, PhD

University of Alabama, USA - smaccall@ua.edu

Linked Data Research Group: <https://wikibase.slis.ua.edu/>

Text Datafication and Loss of Typographical Layout Logic (TLL): Illustration of Problem



Acariasis, 533, 626, 631
demodectic, 630
parasite of, 630
symptoms of, 630
sarcoptic, 627
diagnosis of, 629
frequency of, 628
parasites of, 627
symptoms of, 628
treatment of, 629

Portion of book* index page scan
for the index main entry “Acariasis”

OCR results
for “Acariasis”

Acariasis, 533, 626, 631
demodectic, 630
parasite of, 630
symptoms of, 630
sarcoptic, 627
diagnosis of, 629
frequency of, 628
parasites of, 627
symptoms of, 628
treatment of, 629

*Osler, W., McCrae, T. (1907). Modern medicine: its theory and practice, in original contributions by American and foreign authors. Philadelphia: Lea brothers & co. Vol 1. <https://hdl.handle.net/2027/nnc2.ark:/13960/t54f2j58c>

Theoretical Framework: Navigational Paratexts

- Gerard Genette (1997) introduced concept of paratexts as functional agents that can “mediate the relations between text and reader.”
- Birke and Christ (2013) note that one important category of paratextual element is “navigational paratext:”
 - Such elements guide a reader’s reception in a more mechanical sense
 - Useful in approaching a text and for orienting a reader within the text
- Book index as navigational paratext:
 - Result of closed system indexing in contrast to open system (Klement, 2002)
 - List of entities mentioned within a single book along with locators (e.g., page #s)
 - Provide readers with granular navigation options within texts

Problem Description and Research Question

- For the Google Books Initiative, Google scanned and OCR'd every page of each book using same method.
- As a result, scanned and OCR'd book index pages have index entries without their indentations, which represents TLL that reduces intra-textual navigation compared to original print index.
- The restoration of an index's indentations would lead to testable questions in future research contexts (philology graphs).
- RQ: Can we recover TLL as logical hierarchies and represent them on a RDF knowledge graph?

High-level Workflow

1. Rescan book index pages from HathiTrust copies.
2. Connect hierarchies below index main entries with their respective subentries using partonomic logic property (“partOf”).
3. Load logical relationships on RDF knowledge graph (local Wikibase instance hosted by UA SLIS).
4. Convert page number locators to URLs for each scanned page and align with restored index entries on RDF knowledge graph.
5. Use SPARQL query to reconstruct an overlay HTML index for readers to return indexing navigation capabilities to book.

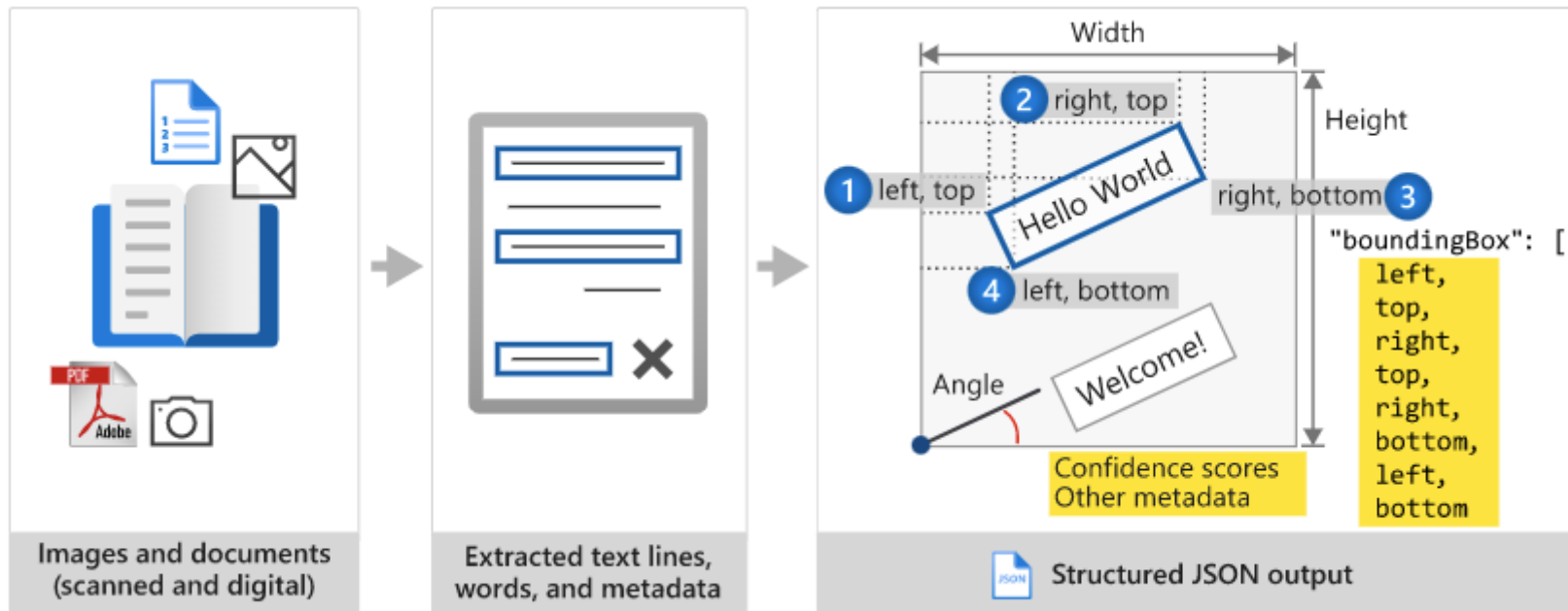
Scaling our Typographical Logic Workflow

- To scale our overlay index workflow, we developed a computational pipeline for detecting index indentations.
- Over the remaining slides, we will introduce some of the technical challenges and will describe the pipeline we developed to detect index indentations using cluster-based machine learning algorithms.

Computational Detection of Index Indentations

- Physical Structure Detection
- Technical Challenges
- Typographical Logic Recovery Pipeline
 - Noise Elimination
 - Column Segmentation
 - Skew Correction
 - Entry Classification
- Evaluation

Physical Structure Detection



Technical Challenges

1264	INDEX
<p>Liver, hydatids of, 353 hyperemia of, 871 acute, 871 passive, 871 hypertrophy of, 875 lobar pneumonia, 97 in phosphorus-poisoning, 889 in typhoid fever, 19, 32 infiltration of, 875 albuminoid, 875 amyloid, 875 bacony, 875 fatty, 877 lardaceous, 875 waxy, 875 malformations of, 854 nutmeg, 871, 890 atrophic, 872 pseudohypertrophy of, 875 sarcoma of, 903 sclerosis of, 890 syphilis of, 384 cancer and, differentiation, 386 tuberculosis of, 275 vascular affections of, 871 Liver-fluke, 350 Lobar pneumonia, 96 abscess in, 97 acute nephritis in, 110 anomalous types, 110 antipneumococcus serum in, 118 arthritis in, 109 bacteriology, 98 blood in, 104 bronchitis in, 108 cardiac clots in, 109 catching cold and, 100 circulatory symptoms, 104 clinical history, 101 complications, 107, 120 congestion stage in, 96 course, 112 delayed resolution, 112 diagnosis, 112 differential, 112 diet in, 116 digestive system, 105 duration, 112 empyema necessitatis in, 108 endemic influence, 99 endocarditis in, 97, 108, 109 engagement stage, 96 epidemic influence, 99 etiology, 98 gangrene in, 97 gastro-intestinal complications, 109 geographic distribution, 99 gray hepatization stage, 96 heart in, 97 hydrotherapy in, 118 in influenza, 135 in typhoid fever, 34 induration in, 97 liver in, 97 mode of infection, 98 parotitis in, 109 pathology, 95 pericarditis in, 97, 108</p>	<p>Lobar pneumonia, peripheral neuritis in, 109 physical signs, 106 pleurisy in, 107 predisposing causes, 99 prognosis, 114 prophylactic inoculation in, 119 purulent infiltration in, 97 red hepatization stage, 96 respiratory stimulants in, 118 season, 99 serum treatment, 118 specific therapy, 118 spleen in, 97 sputum in, 103 stimulants in, 117 suppurative, 556 symptoms, 102 cerebral, 105 circulatory, 104 local, 102 respiratory, 102 urinary, 105 temperature in, 103 treatment, 116 local, 121 of special symptoms, 120 prophylaxis in, 116, 119 serum, 118 vaccine, 119 vaccines in, 119 varieties, 110 venesection in, 118 Lobstein's cancer, 837 Lock-jaw, 303. See also <i>Tetanus</i>. Locomotor ataxia, 1150 Ludwig's angina, 724 Luteic curve, 1146 Lumbago, 312 Lumbar plexus, diseases of, 1076 puncture in cerebrospinal meningitis, 92, 95 Lumbo-abdominal neuralgia, 1041 Lumpy-jaw, 318 Lung, abscess of, 556 brown induration of, 535 carcinoma of, 559 circulatory disturbances in, 534 diseases of, 534 collapse of, 546 compression of, 546 congestion of, 534 diseases of, 534 edema of, 536 embolism of, 543 fever, 96. See also <i>Lobar pneumonia</i>. gangrene of, 554 embolic, 554 hydatid cyst of, 561 hyperemia of, active, 534 hypostatic, 535 mechanical, 535 passive, 535 in diabetes, 405 in typhoid fever, 19</p>

INDEX	1275
<p>Rubella notha, 216. See also Rubella. Rumination, 793 Runeberg's method of examination of stomach, 753 Rupture of esophagus, 742 of heart, 671 of spleen, 907</p> <p>SACCHAROMYCES albicans, 716 Sacral plexus, diseases of, 1076 Sago spleen, 906 Saliva, hypersecretion of, 724 Salivary glands, diseases of, 724 Salpingitis, typhoid fever and, differentiation, 44 Salt test for renal function, 958 Saltatoric spasm, 1169 Salvarsan in rat-bite fever, 310 Sanatorium treatment of tuberculosis, 284 Sand, renal, 967 Sand-flea, 375 Sapremia, 165 Sarcocystis hominis, 330 mischri, 330 respiratory, 1130 of kidney, 999 of liver, 993 of lung, 561 of mediastinum, 590 of peritoneum, 932 of pleura, 589 Sarcophila carnaria, 376 Sarcoptes scabiei hominis, 374 Saturine gout, 1223 neuritis, 1045 Saturnism, 1222 Scabies, 374 Scapulohumeral type of muscular atrophy, 1208 Scarlatina, 201. See also <i>Scarlet fever</i>. anginosa, 201 sine eruptione, 205 Scarlatinal angina, 201 synovitis, 211 Scarlet fever, 201 anginose form, 205 atactic form, 205 bacteriology, 201 clinical history, 203 complications, 206 desquamation in, 203, 205 diagnosis, 218 differential, 208 disinfection in, 209 eruption, 203 etiology, 201 hemorrhagic, 205 immunity to, 203 incubation period, 203 invasion, 203 joint affections in, 206 malignant, 205 mild, 205 modes of conveyance and infection, 202 nephritis in, 206 otitis in, 206</p>	<p>Scarlet fever, pathology, 201 predisposing causes, 202 prognosis, 208 prophylaxis, 209 pyemia in, 206 serum treatment, 211 traumatic, 205 treatment, 209 types, 203 rash, 201. See also <i>Scarlet fever</i>. Schick test in diphtheria, 163 Schlammfleber, 397 Schonlein's disease, 467 Schott treatment in valvular disease, 643 Sciatic nerve, great, paralysis of, 1076 small, paralysis of, 1076 neuritis, 1042 Sciatica, 1042 Scleroactylia, 1202 Sclerolems circumscriptum, 1203 diffusum, 1202 Sclerosis, amyotrophic lateral, 1105 arterial, 693. See also <i>Arteriosclerosis</i>. combined system, 1109 diffuse, 1143 disseminated, 1140 in tuberculosis, 228 insular, 1140 multiple, 1140 of liver, 890 of pulmonary arteries, 694 of veins, 694 posterior, 1150 posterolateral, 1108 primary lateral, 1103 subacute combined, of spinal cord, 1109 Sclerotic dysentery, 441 Scorbutic dysentery, 441 Scorbutus, 440 bacteriology, 440 diagnosis, 442 etiology, 440 infantile, 443 pathology, 440 predisposing causes, 440 prognosis, 442 symptoms, 441 treatment, 442 Screw-worm fly, 376 Scrofula, 236 Seroal appendicitis, 822 Serotum, lymph-, from filaria, 372 Scurvy, 440. See also <i>Scorbutus</i>. Sea-worm, 362 Secondary anemias, 463 pneumonia, 121 Segmental anesthesia, 1179 Seminal vesicles, tuberculosis of, 277 Senile arteriosclerosis, 694 dementia, 1140 emphysema, 554 neuritis, 1045 tremor, 1171 Sensation, Bernhardt's disturbance of, 1201 Sensory aphasia, 1024 Sepsis, focal, 170</p>

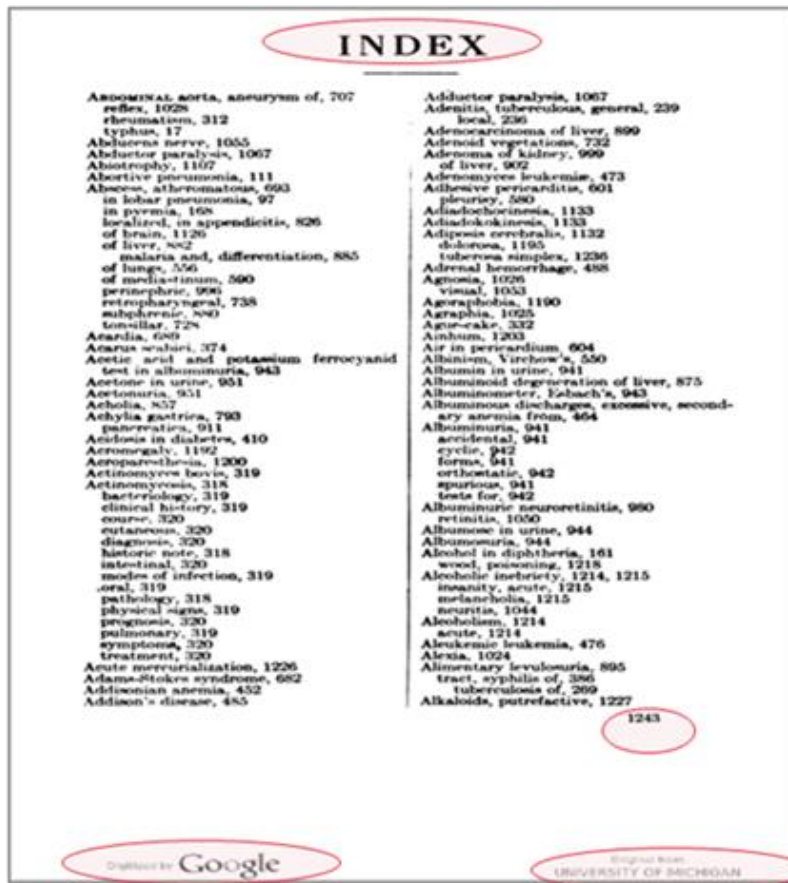
Technical Challenges

	X1	Y1	X2	Y2	X3	Y3	X4	Y4	text
2	178	720	740	720	740	757	178	757	ABDOMINAL aorta, aneurysm of, 707
3	211	755	391	754	391	787	211	788	reflex, 1028
4	210	787	477	787	477	821	210	821	rheumatism, 312
5	205	824	379	823	379	855	206	856	typhus, 17
6	178	853	522	853	522	888	178	888	Abducens nerve, 1055
7	179	888	566	887	566	921	179	923	Abductor paralysis, 1067
8	179	922	450	921	450	956	179	957	Abiotrophy, 1107
9	177	954	569	954	569	989	177	989	Abortive pneumonia, 111
10	176	989	606	987	606	1020	176	1022	Abscess, atheromatous, 693
11	205	1020	573	1020	573	1056	205	1056	in lobar pneumonia, 97
12	210	1055	445	1053	445	1087	210	1090	in pvemia, 168
13	210	1085	669	1085	669	1123	210	1123	localized, in appendicitis, 826
14	212	1121	430	1121	430	1154	212	1154	of brain, 1126
15	209	1154	403	1152	403	1185	210	1188	of liver, 882
16	242	1187	749	1185	750	1220	242	1222	malaria and, differentiation, 885
17	207	1220	416	1220	416	1255	207	1255	of lungs, 556
18	209	1254	529	1253	530	1287	209	1288	of mediastinum, 590
19	207	1290	463	1288	464	1319	208	1323	perinephric, 996
20	208	1322	536	1318	536	1354	208	1357	retropharyngeal, 738
21	209	1355	458	1355	458	1386	209	1389	subphrenic, 880
22	208	1388	415	1386	415	1418	208	1419	tonsillar, 728
23	175	1421	373	1421	373	1452	175	1454	Acardia, 689
24	175	1454	475	1453	475	1484	175	1486	Acarus scabiei, 374
25	174	1484	828	1484	828	1520	174	1521	Acetic acid and potassium ferrocyanid

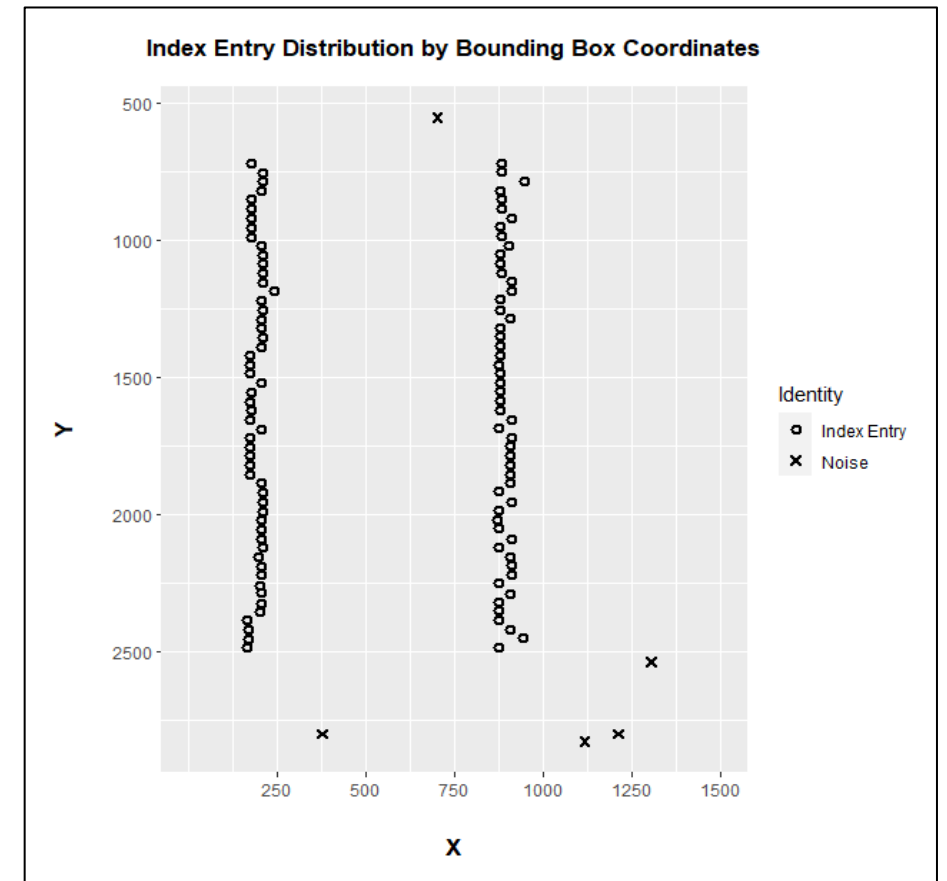


INDEX	
ABDOMINAL aorta, aneurysm of, 707	Adductor paralysis, 1067
reflex, 1028	Adenitis, tuberculous, general, 239
rheumatism, 312	local, 236
typhus, 17	Adenocarcinoma of liver, 890
Abducens nerve, 1055	Adenoid vegetations, 732
Abductor paralysis, 1067	Adenoma of kidney, 999
Abiotrophy, 1107	of liver, 902
Abortive pneumonia, 111	Adenomyces leukemie, 473
Abscess, atheromatous, 693	Adhesive pericarditis, 601
in lobar pneumonia, 97	pleurisy, 580
in pyemia, 168	Adialochocinesia, 1133
localized, in appendicitis, 826	Adiakokinesia, 1133
of brain, 1126	Adiposis cerebri, 1132
of liver, 882	dolorosa, 1195
malaria and, differentiation, 885	tuberosa simplex, 1236
of lungs, 556	Adrenal hemorrhage, 488
of mediastinum, 590	Agnosia, 1026
perinephric, 996	visual, 1053
retropharyngeal, 738	Agoraphobia, 1190
subphrenic, 880	Agraphia, 1025
tonsillar, 728	Ague-cake, 332
Acardia, 689	Ainhum, 1203
Acarus scabiei, 374	Air in pericardium, 604
Acetic acid and potassium ferrocyanid	Albumin, Virchow's, 550
test in ammonium, 945	Albumin in urine, 941
Acetone in urine, 951	Albuminoid degeneration of liver, 875
Acetonuria, 951	Albuminometer, Esbach's, 943
Acholia, 857	Albuminous discharges, excessive, second-
Achylia gastrica, 793	ary anemia from, 464
pancreatica, 911	Albuminuria, 941
Acidosis in diabetes, 410	accidental, 941
Acromegaly, 1192	cyclic, 942
Acroparesthesia, 1200	forms, 941
Actinomyces bovis, 319	orthostatic, 942
Actinomycesis, 318	spurious, 941
bacteriology, 319	tests for, 942
clinical history, 319	Albuminuric neuroretinitis, 980
course, 320	retinitis, 1050
cutaneous, 320	Albumose in urine, 944
diagnosis, 320	Albumosuria, 944
historic note, 318	Alcohol in diphtheria, 161
intestinal, 320	wood, poisoning, 1218
modes of infection, 319	Aleoholic inebriety, 1214, 1215
oral, 319	insanity, acute, 1215
pathology, 318	melancholia, 1215
physical signs, 319	neuritis, 1044
prognosis, 320	Aleoholism, 1214
pulmonary, 319	acute, 1214
symptoms, 320	Aleukemic leukemia, 476
treatment, 320	Alexia, 1024
Acute mercurialization, 1226	Alimentary leucosuria, 895
Adams-Stokes syndrome, 682	tract, syphilis of, 266
Addisonian anemia, 452	tuberculosis of, 269
Addison's disease, 485	Alkaloids, putrefactive, 1227

Noise Elimination



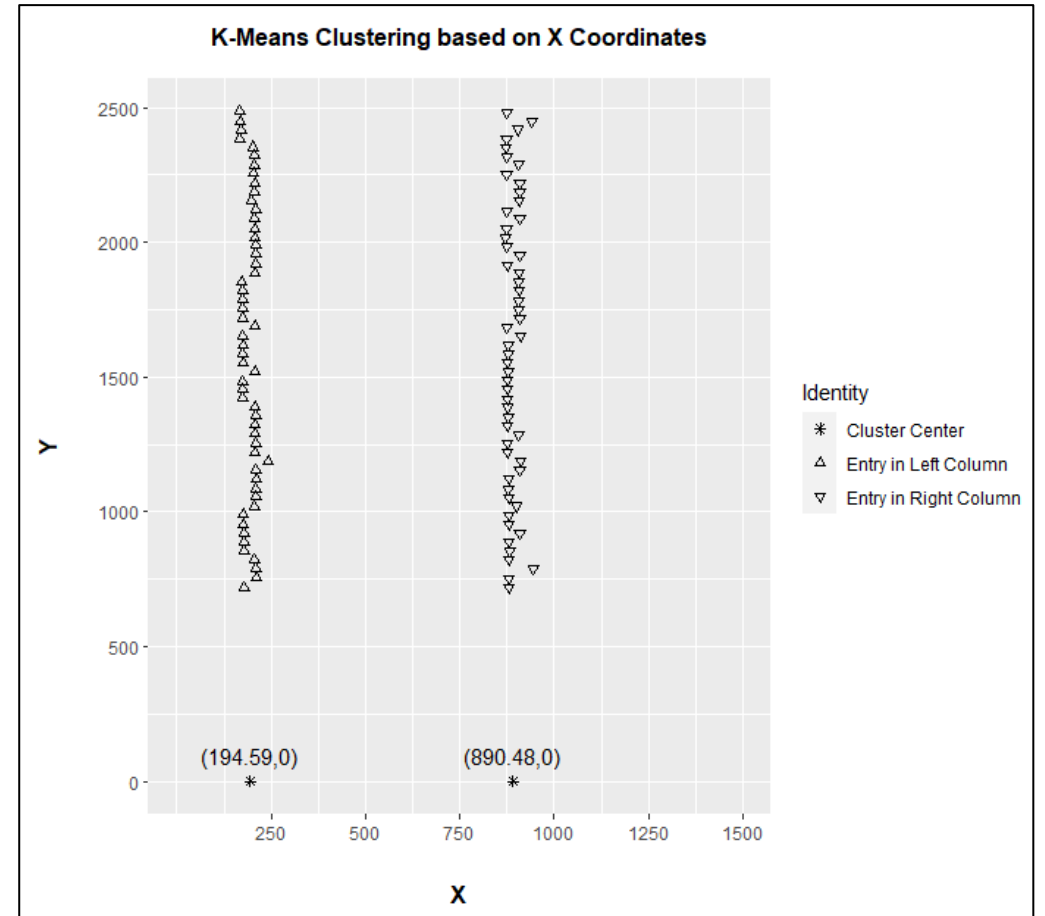
DBSCAN Algorithm



Column Segmentation

INDEX	
<p>ABDOMINAL aorta, aneurysm of, 707 reflex, 1028 rheumatism, 312 typhus, 17 Abducens nerve, 1055 Abductor paralysis, 1067 Abiotrophy, 1107 Abortive pneumonia, 111 Abscess, atheromatous, 693 in lobar pneumonia, 97 in pyemia, 168 localized, in appendicitis, 826 of brain, 1126 of liver, 882 malaria and, differentiation, 885 of lungs, 556 of mediastinum, 590 perinephric, 996 retropharyngeal, 738 subphrenic, 880 tonsillar, 728 Acardia, 689 Acarus scabiei, 374 Acetic acid and potassium ferrocyanid test in albuminuria, 943 Acetone in urine, 951 Acetonuria, 951 Acholia, 857 Achylia gastrica, 793 pancreatica, 911 Acidosis in diabetes, 410 Aeromegaly, 1192 Aeroparesis, 1200 Aetionomyces bovis, 319 Aetionomyces, 318 bacteriology, 319 clinical history, 319 course, 320 cutaneous, 320 diagnosis, 320 historic note, 318 intestinal, 320 modes of infection, 319 oral, 319 pathology, 318 physical signs, 319 prognosis, 320 pulmonary, 319 symptoms, 320 treatment, 320 Acute mercurialization, 1226 Adams-Stokes syndrome, 682 Addisonian anemia, 452 Addison's disease, 485</p>	<p>Adductor paralysis, 1067 Adenitis, tuberculous, general, 239 local, 236 Adenocarcinoma of liver, 899 Adenoid vegetations, 732 Adenoma of kidney, 999 of liver, 902 Adenomyces leukemie, 473 Adhesive pericarditis, 601 pleurisy, 580 Adiadochocinesia, 1133 Adiadokinesia, 1133 Adiposis cerebri, 1132 dolorosa, 1195 tuberosa simplex, 1236 Adrenal hemorrhage, 488 Agnosis, 1026 visual, 1053 Agoraphobia, 1190 Agraphia, 1025 Ague-cake, 332 Ainhum, 1203 Air in pericardium, 604 Albinism, Virchow's, 550 Albumin in urine, 941 Albuminoid degeneration of liver, 875 Albuminometer, Esbach's, 943 Albuminous discharges, excessive, second- ary anemia from, 464 Albuminuria, 941 accidental, 941 cyclic, 942 forms, 941 orthostatic, 942 spurious, 941 tests for, 942 Albuminuric neuroretinitis, 980 retinitis, 1050 Albumose in urine, 944 Albumosuria, 944 Alcohol in diphtheria, 161 wood, poisoning, 1218 Alcoholic inebriety, 1214, 1215 insanity, acute, 1215 melancholia, 1215 neuritis, 1044 Alcoholism, 1214 acute, 1214 Aleukemic leukemia, 476 Alexia, 1024 Alimentary leucosuria, 895 tract, syphilis of, 386 tuberculosis of, 269 Alkaloids, putrefactive, 1227</p>

K-Means Algorithm



Skew Correction

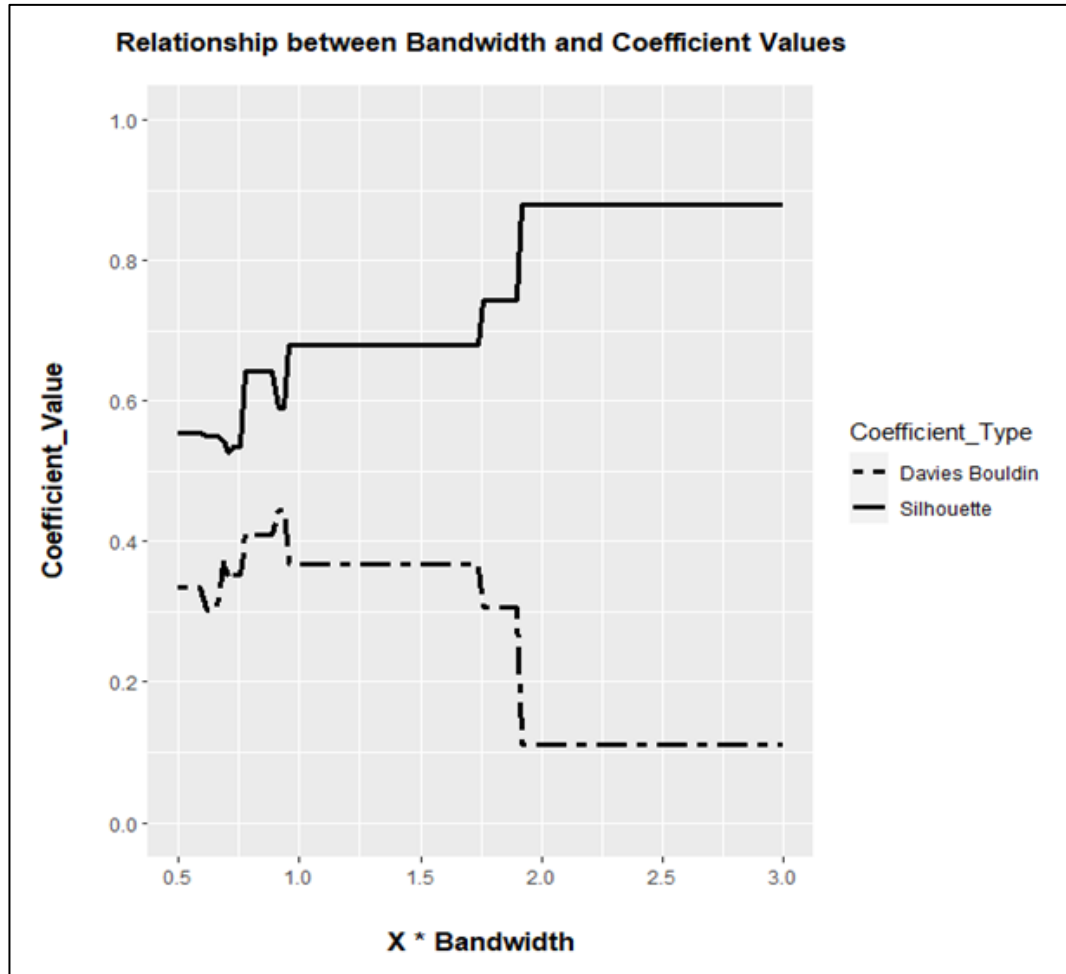
INDEX		1271
Paratyphoid, exanthematous, 735	follicular, 737	herpetic, 735
aneurysmal, 736	sepsis, 736	Pharyngocoele, 744
Pharynx, diseases of, 735	in typhoid fever, 20, 33	tuberculosis of, 209
Phenolphthalein test of renal function, 956	Phlebotomy, 694	Phlegmon of throat, acute infectious, 738
Phlegmonous cystitis, 1001	enteritis, 815	erysipelas, 148
gastritis, 764	Phloridin diabetes, 404	Phloroglucin-vanillin test, 750
Phosphates, excess of, in urine, 952	Phosphaturia, 952	Phosphorus-poisoning, liver in, 880
Phrenic nerve, diseases of, 1073	Phthiasis, 374	pubis, 375
Phthisis, acute, 246	bronchopneumonic, in children, 248	pneumonic, 246
clinical history, 247	pathology, 246	chronic ulcerative, 249. See also Tuberculosis, chronic.
fibroid, 267	complications, 267	course, 268
diagnosis, differential, 267	duration, 268	pathology, 267
symptoms, 267	Bord, 246	stone-cutters', 558
Pia mater, inflammation of, 1079	Pica, 797	in chlorosis, 448
Picric acid test for albuminuria, 943	Pigeon breast in rachitis, 432	Pigmentary retinitis, 1050
Pin-worm, 362	Pituitary gland, rôle of, in diabetes, 403	Plague, 141
bacteriology, 142	clinical history, 142	diagnosis, 143
etiology, 142	historic summary, 141	incubation, 142
modes of transmission and entrance, 142	mortality, 143	pneumonia, 143
predisposing causes, 142	prognosis, 143	prophylaxis, 144
sequels, 143	treatment, 144	varieties, 142
Plantar reflex, 1021	Plaques à surface réticulée, 18	jaunes, 1121
Plaques opales, 725	Plasmodium malaris, 332	vivax, 336
Plastic bronchitis, 532	pericarditis, acute, 593	pleurisy, acute, 563
in influenza, 134	pleura, carcinoma of, 589	diseases of, 562
dropsy of, 587	new growths of, 589	sarcoma of, 589
leureisy, 562	acute plastic, 563	adhesive, 580
bacteriology, 563	blocked, 576	chronic, 580
dry, 581	kilopathie, 581	treatment, 581
with effusion, 580	diaphragmatic, 571	dry fibrous, 363
encysted, 571	hemorrhagic, 572	in lobar pneumonia, 107
in typhoid fever, 20	interlobar, 571	plastic, in influenza, 134
pulsating, 578	serofibrinous, 565	diagnosis, 573
duration, 574	etiology, 566	Grocco's sign in, 569
Laennee's epiphony in, 569	lobar pneumonia and, differentiation, 575	pathology, 565
prognosis, 574	Röntgen rays in, 570	Skoda's resonance in, 569
special forms, 570	subacute, 565	tuberculous, 271, 570
varieties, 562	with effusion, 565	pleuritis, 562
purulent, 576. See also Empyema.	retahens, 579	pleurodynia, 312
flexus, brachial, diseases of, 1073	cervical, diseases of, 1073	lumbar, diseases of, 1076
sacral, diseases of, 1076	flexion polonica, 375	Rumbson, 1222
Pneumatism, 955	Pneumatism, 955	Pneumococci infections, 95
Pneumococcus septicaemia, 107	Pneumogastric nerve, diseases of, 1065	Pneumonia, abortive, 111
bilious, 111		

INDEX		1244
Alkapton in urine, 956	Alkapturia, 956	Allen diet in diabetes, 413
Allothymia, 679	Alveolar ectasis, 548	Alveolitis, 721
Amanita muscaria, 1230	Amaturosis, 1050	hysterie, 1179
uremie, 962	Amaturotic family idiocy, 1196	dropsy of, 587
Ambyopia, toxic, 1050	Amebiasis, 322	Amebic dysentery, 322. See also Dysentery, amebic.
American trypanosomiasis, 328	Amimia, 1025, 1132	Amoeba dysentery, 322
Amphistomum hominis, 351	Amusia, 1025	Amorphous degeneration of heart, 670
of spleen, 906	infiltration of liver, 875	kidney, 965
Amyotonia congenita, 1212	Amyotrophia spinalis progressiva, 1104	Amyotrophic lateral sclerosis, 1105
Anæmia infantum pseudoleukæmicum, 483	Anal reflex, 1030	Anaphylactogen, 1230
Anaphylatoxin, 1231	Anaphylaxis, 1230	in diphtheria, 164
Anæmia Addisonian, 452	aplastic, 460	brickmaker's, 364
hemolytic, of pregnancy and puerperium, 460	idiopathic, 452	mountain, 364
of brain, 1112	of liver, 871	of spinal cord, 1084
primary or essential, 447	splenomegaly with, 461	progressive pernicious, 452
blood-examination in, 455	diagnosis, 456	differential, 456
laboratory findings in, 455	obscure gastric carcinoma and, differentiation, 457	pathology, 452
predisposing causes, 453	prognosis, 457	splenectomy in, 459
symptoms, 454	gastro-intestinal, 454	nervous, 455
respiratory, 454	treatment, 457	secondary, 463
blood in, 463	diagnosis, 465	from excessive albuminous discharges, 464
from inanition, 464	from toxic agents, 464	
Anæmia, secondary, hemorrhage from, 463	symptoms, 465	treatment, 465
splenic, 461	Anæsthesia dolorosa, 1007	olfactory, 1049
segmental, 1179	Anæsthetic form of leprosy, 293	Aneurysm, 699
arteriovenous, 699, 709	axial, 699	congenital, 710
dissecting, 699	etiology, 699	false, 699
miliary, 699	mycotic, 699	of abdominal aorta, 707
of brain, 1114	of celiac axis, 708	of heart, 671
valvular, 671	of hepatic artery, 709	of pulmonary artery, 708
of renal arteries, 709	of splenic artery, 709	of superior mesenteric artery, 709
of thoracic aorta, 700	diagnosis, 704	differential, 704
prognosis, 705	symptoms, 700	treatment, 705
pathology, 699	peripheral, 699	varicose, 699
aneurysms aortæ, 700	aneurysmal varix, 699	Angina abdominalis, 696
Ludwig's, 724	major, 687	maligna, 151. See also Diphtheria.
membranous, 159	minor, 687	pectoris, 686
pseudo-, 687	scarlatinal, 201	Vincent's, 159, 731
Angiocheilitis ulcerative, 863	Angioma of brain, 1130	of kidney, 999
of liver, 992	Angioneurotic edema, 1107	intermittent, 1107
Angiosclerosis, 694	Angiospastic dilatation of heart, 657	Anguilla stevoralis, 373
Achlo-clonus, 1028	Achyllostoma duodenale, 363	Achyllostomiasis, 363
Abomia, 1025	Achilles claviger, 333	crucians, 333
febrifer, 333		

$$\tan(\theta^\circ) = 1/N \cdot \sum (Y_R - Y_L)/(X_R - X_L) \quad (1)$$

$$X_i' = X_i + \tan(\theta^\circ) \cdot (Y_i - Y_1) \quad (2)$$

Entry Classification



Mean Shift Bandwidth Estimation



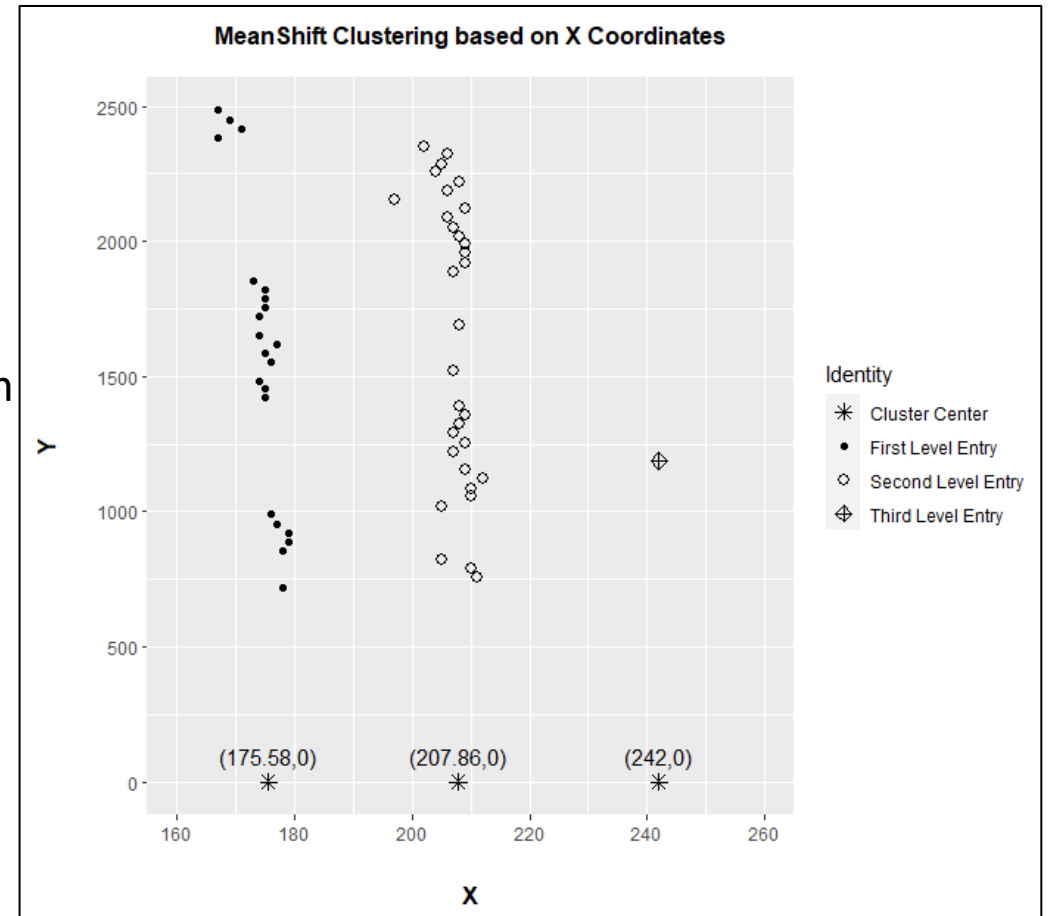
Mean Shift clustering performance evaluation:

- Silhouette Coefficient
- Davies-Bouldin Index

Recovering Typographical Layout Logic with

INDEX	
ABDOMINAL aorta, aneurysm of, 707	Adductor paralysis, 1067
reflex, 1028	Adenitis, tuberculous, general, 239
rheumatism, 312	local, 236
typhus, 17	Adenocarcinoma of liver, 899
Abducens nerve, 1055	Adenoid vegetations, 732
Abductor paralysis, 1067	Adenoma of kidney, 999
Abiotrophy, 1107	of liver, 902
Abortive pneumonia, 111	Adenomyces leukemie, 473
Abscess, atheromatous, 693	Adhesive pericarditis, 601
in lobar pneumonia, 97	pleurisy, 580
in pyemia, 168	Adiadochocinesia, 1133
localized, in appendicitis, 826	Adiokokinesia, 1133
of brain, 1126	Adiposis cerebri, 1132
of liver, 882	dolorosa, 1195
malaria and, differentiation, 885	tuberosa simplex, 1236
of lungs, 556	Adrenal hemorrhage, 488
of mediastinum, 590	Agnosia, 1026
perinephric, 996	visual, 1053
retropharyngeal, 738	Agoraphobia, 1190
subphrenic, 880	Agraphia, 1025
tonsillar, 728	Ague-cake, 332
Acardia, 689	Ainhum, 1203
Acarus scabiei, 374	Air in pericardium, 604
Acetic acid and potassium ferrocyanid	Albinism, Virchow's, 550
test in albuminuria, 943	Albumin in urine, 941
Acetone in urine, 951	Albuminoid degeneration of liver, 875
Acetonuria, 951	Albuminometer, Esbach's, 943
Acholia, 857	Albuminous discharges, excessive, second-
Achylia gastrica, 793	ary anemia from, 464
pancreatica, 911	Albuminuria, 941
Acidosis in diabetes, 410	accidental, 941
Acromegaly, 1192	cyclic, 942
Acroparesthesia, 1200	forms, 941
Actinomyces bovis, 319	orthostatic, 942
Actinomyces, 318	spurious, 941
bacteriology, 319	tests for, 942
clinical history, 319	Albuminuric neuroretinitis, 980
course, 320	retinitis, 1050
cutaneous, 320	Albumose in urine, 944
diagnosis, 320	Albumosuria, 944
historic note, 318	Alcohol in diphtheria, 161
intestinal, 320	wood, poisoning, 1218
modes of infection, 319	Alcoholic inebriety, 1214, 1215
oral, 319	insanity, acute, 1215
pathology, 318	melancholia, 1215
physical signs, 319	neuritis, 1044
prognosis, 320	Alcoholism, 1214
pulmonary, 319	acute, 1214
symptoms, 320	Aleukemic leukemia, 476
treatment, 320	Alexia, 1024
Acute mercurialization, 1226	Alimentary leucosuria, 895
Adams-Stokes syndrome, 682	tract, syphilis of, 386
Addisonian anemia, 452	tuberculosis of, 269
Addison's disease, 485	Alkaloids, putrefactive, 1227

Mean Shift Algorithm



Evaluation of Pipeline Effectiveness: Accuracy Testing

Entry_Text	Entry_Level
1 ABDOMINAL aorta, aneurysm of, 707	1
2 reflex, 1028	2
3 rheumatism, 312	2
4 typhus, 17	2
5 Abducens nerve, 1055	1
6 Abductor paralysis, 1067	1
7 Abiotrophy, 1107	1
8 Abortive pneumonia, 111	1
9 Abscess, atheromatous, 693	1
10 in lobar pneumonia, 97	2
11 in pvemia, 168	2
12 localized, in appendicitis, 826	2
13 of brain, 1126	2
14 of liver, 882	2
15 malaria and, differentiation, 885	3
16 of lungs, 556	2
17 of mediastinum, 590	2
18 perinephric, 996	2
19 retropharyngeal, 738	2
20 subphrenic, 880	2
21 tonsillar, 728	2
22 Acardia, 689	1
23 Acarus scabiei, 374	1
24 Acetic acid and potassium ferrocyanid	1
25 test in albuminuria, 943	2

Accuracy Testing



INDEX	
<p>ABDOMINAL aorta, aneurysm of, 707 reflex, 1028 rheumatism, 312 typhus, 17 Abducens nerve, 1055 Abductor paralysis, 1067 Abiotrophy, 1107 Abortive pneumonia, 111 Abscess, atheromatous, 693 in lobar pneumonia, 97 in pyemia, 168 localized, in appendicitis, 826 of brain, 1126 of liver, 882 malaria and, differentiation, 885 of lungs, 556 of mediastinum, 590 perinephric, 996 retropharyngeal, 738 subphrenic, 880 tonsillar, 728 Acardia, 689 Acarus scabiei, 374 Acetic acid and potassium ferrocyanid test in albuminuria, 943 Acetone in urine, 951 Acetonuria, 951 Achohia, 857 Achyia gastrica, 793 pancreatica, 911 Acidosis in diabetes, 410 Acromegaly, 1192 Acroparesthesia, 1200 Actinomyces bovis, 319 Actinomycesis, 318 bacteriology, 319 clinical history, 319 course, 320 cutaneous, 320 diagnosis, 320 historic note, 318 intestinal, 320 modes of infection, 319 oral, 319 pathology, 318 physical signs, 319 prognosis, 320 pulmonary, 319 symptoms, 320 treatment, 320 Acute mercurialization, 1226 Adams-Stokes syndrome, 682 Addisonian anemia, 452 Addison's disease, 485</p>	<p>Adductor paralysis, 1067 Adenitis, tuberculous, general, 239 local, 236 Adenocarcinoma of liver, 899 Adenoid vegetations, 732 Adenoma of kidney, 999 of liver, 902 Adenomyces leuquemis, 473 Adhesive pericarditis, 601 pleurisy, 580 Adiadochocinesia, 1133 Adiadokinesia, 1133 Adiposis cerebri, 1132 dolorosa, 1195 tuberosa simplex, 1236 Adrenal hemorrhage, 488 Agnosia, 1026 visual, 1053 Agoraphobia, 1190 Agraphia, 1025 Ague-cake, 332 Ainhum, 1203 Air in pericardium, 604 Albinism, Virchow's, 550 Albumin in urine, 941 Albuminoid degeneration of liver, 875 Albuminometer, Esbach's, 943 Albuminous discharges, excessive, second- ary anemia from, 464 Albuminuria, 941 accidental, 941 cyclic, 942 forms, 941 orthoatic, 942 spurious, 941 tests for, 942 Albuminuric neuroretinitis, 980 retinitis, 1050 Albumose in urine, 944 Albumosuria, 944 Alcohol in diphtheria, 161 wood, poisoning, 1215 Alcoholic inebriety, 1214, 1215 insanity, acute, 1215 melancholia, 1215 neuritis, 1044 Alcoholism, 1214 acute, 1214 Aleukemic leukemia, 476 Alexia, 1024 Alimentary levulosuria, 895 tract, syphilis of, 269 tuberculosis of, 269 Alkaloids, putrefactive, 1227</p>

Evaluation of Pipeline Effectiveness: ML Procedures

Procedure	Algorithm	Accuracy
Noise Elimination	DBSCAN	95.4 %
Column Segmentation	K-Means	98.2 %
Entry Classification	Mean Shift	93.8 %

Recovered Typographical Layout Logic to Graph (1)

Acariasis, 533, 626, 631 (Index entry from HathiTrust item t54f2j58c) (Q11494)

From Wikibase.slis.ua.edu



From: Modern medicine : its theory and practice, in original contributions by American and foreign authors (1907-1910.). Volume 1.





[wikit](#)

[In more languages](#) [Configure](#)

Language	Label	Description	Also known as
English	Acariasis, 533, 626, 631 (Index entry from HathiTrust item t54f2j58c)	From: Modern medicine : its theory and practice, in original contributions by American and foreign authors (1907-1910.). Volume 1.	

Statements

instance of	 Book index main heading  0 references
-------------	--

has part	 => demodectic, 630 (Index entry from HathiTrust item t54f2j58c)  0 references
	 =>sarcoptic, 627 (Index entry from HathiTrust item t54f2j58c)  0 references

Recovered Typographical Layout Logic to Graph (2)

==> demodectic, 630 (Index entry from HathiTrust item t54f2j58c) (Q11495)

From Wikibase.slis.ua.edu

From: Modern medicine : its theory and practice, in original contributions by American and foreign authors (1907-1910.). Volume 1.

▼ In more languages [Configure](#)

Language	Label	Description	Also known as
English	==> demodectic, 630 (Index entry from HathiTrust item t54f2j58c)	From: Modern medicine : its theory and practice, in original contributions by American and foreign authors (1907-1910.). Volume 1.	

Statements

instance of	 Book index subheading ▼ 0 references
-------------	---

part of	 Acariasis, 533, 626, 631 (Index entry from HathiTrust item t54f2j58c) ▼ 0 references
---------	---

has part	 ====>parasite of, 630 (Index entry from HathiTrust item t54f2j58c) ▼ 0 references
	 ====>symptoms of, 630 (Index entry from HathiTrust item t54f2j58c) ▼ 0 references

Displaying Typographical Logic Layout: SPARQL Query

- The transitive operator (*) allows for queries that follow our partOf (P22) property path, which captures the multiple levels of index subentries under the Acariasis (Q11494) main entry in book index.
- Transitive SPARQL query example (<https://tinyurl.com/ydv6rq2y>):

```
1 # Book index entry for "Acariasis": Example query with transitive operator demonstrating property path search
2 SELECT ?IndexEntryLabel ?URL
3 WHERE
4 {
5   wd:Q11494 wdt:P22* ?IndexEntry.
6   ?IndexEntry wdt:P181 ?URL.
7   SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
8 }
```

Displaying Typographical Logic Layout: SPARQL Results

Arrows indicate restored indentations. Wikibase SPARQL query results exportable as HTML for overlay index.

IndexEntryLabel	URL
Acariasis, 533, 626, 631 (Index entry from HathiTrust item t54f2j58c)	https://babel.hathitrust.org/cgi/pt?id=nnc2.ark:/13960/t54f2j58c&view=1up&seq=573
Acariasis, 533, 626, 631 (Index entry from HathiTrust item t54f2j58c)	https://babel.hathitrust.org/cgi/pt?id=nnc2.ark:/13960/t54f2j58c&view=1up&seq=668
Acariasis, 533, 626, 631 (Index entry from HathiTrust item t54f2j58c)	https://babel.hathitrust.org/cgi/pt?id=nnc2.ark:/13960/t54f2j58c&view=1up&seq=673
==> demodectic, 630 (Index entry from HathiTrust item t54f2j58c)	https://babel.hathitrust.org/cgi/pt?id=nnc2.ark:/13960/t54f2j58c&view=1up&seq=672
====>parasite of, 630 (Index entry from HathiTrust item t54f2j58c)	https://babel.hathitrust.org/cgi/pt?id=nnc2.ark:/13960/t54f2j58c&view=1up&seq=672
====>symptoms of, 630 (Index entry from HathiTrust item t54f2j58c)	https://babel.hathitrust.org/cgi/pt?id=nnc2.ark:/13960/t54f2j58c&view=1up&seq=672
==>sarcoptic, 627 (Index entry from HathiTrust item t54f2j58c)	https://babel.hathitrust.org/cgi/pt?id=nnc2.ark:/13960/t54f2j58c&view=1up&seq=669
====>diagnosis of, 629 (Index entry from HathiTrust item t54f2j58c)	https://babel.hathitrust.org/cgi/pt?id=nnc2.ark:/13960/t54f2j58c&view=1up&seq=671
====>frequency of, 628 (Index entry from HathiTrust item t54f2j58c)	https://babel.hathitrust.org/cgi/pt?id=nnc2.ark:/13960/t54f2j58c&view=1up&seq=670
====>parasites of, 627 (Index entry from HathiTrust item t54f2j58c)	https://babel.hathitrust.org/cgi/pt?id=nnc2.ark:/13960/t54f2j58c&view=1up&seq=669
====>symptoms of, 628 (Index entry from HathiTrust item t54f2j58c)	https://babel.hathitrust.org/cgi/pt?id=nnc2.ark:/13960/t54f2j58c&view=1up&seq=670
====>treatment of, 629 (Index entry from HathiTrust item t54f2j58c)	https://babel.hathitrust.org/cgi/pt?id=nnc2.ark:/13960/t54f2j58c&view=1up&seq=671

Conclusion and Future Work

- Conclusion”
 - Pipeline development: From theory to optimizing for scale up
 - Evaluation
- Future work:
 - Layout variations
 - Overlay index interface work in the context of a larger expression of navigation paratext theory and philology graphs

References

- Birke, Dorothee and Christ, Birte. 2013. Paratext and Digitized Narrative: Mapping the Field. *Narrative* 21(1), 65-87.
- Genette, Gerard. 1997. *Paratexts: thresholds of interpretation*. Cambridge University Press.
- Klement, Susan. 2002. Open-system versus closed-system indexing. *The Indexer*, 23(1), 23-31.
- Mayer-Schönberger, Viktor and Cukier, Kenneth. 2013. Chapter 5. Datafication. In *Big Data : A Revolution That Will Transform How We Live, Work, and Think* (pp. 73-97). Houghton, Mifflin, Harcourt.