HathiTrust Research Center

Welcome to the HathiTrust Research Center!

The HathiTrust Research Center (HTRC) provides research access to the public domain text of the HathiTrust Digital Library. The HTRC is a collaborative research center launched jointly by Indiana University and the University of Illinois, along with the HathiTrust Digital Library, to help meet the technical challenges of dealing with massive amounts of digital text that researchers face by developing cutting-edge software tools and cyberinfrastructure to enable advanced computational access to the growing digital record of human knowledge.

The HTRC provides an infrastructure to search, collect, analyze, and visualize the full text of nearly 3 million public domain works and is intended for nonprofit and educational researchers.

HathiTrust Research Center

Submit Algorithms

HOME ABOUT COLLECTION ALGORITHMS RESULTS HELP LOGOUT

Available Algorithms

- Meandre_Topic_Modeling
- Marc_Downloader
- Simple_Deployable_Word_Count
- Meandre_Spellcheck_Report_Per_Volume
- Meandre_Tag_Cloud_with_Cleaning
- Meandre_OpenNLP_Entity_List
- Meandre_Dunning_Loglikelihood
- HTRC_Meandre_Entity_Social_Network
- Meandre_Simple_Tag_Cloud
- Meandre_OpenNLP_Date_Entities_To_Simile

Algorithm Parameters

Please select an algorithm in the list to display information.

HathiTrust Research Center

Submit Algorithms

HOME ABOUT COLLECTION ALGORITHMS RESULTS HELP LOGOUT

Available Algorithms

- Marc_Downloader
- Simple_Deployable_Word_Count
- Meandre_Spellcheck_Report_Per_Volume
- Meandre_Tag_Cloud_with_Cleaning
- Meandre_OpenNLP_Entity_List
- Meandre_Dunning_Loglikelihood
- Meandre_Simple_Tag_Cloud
- Meandre_OpenNLP_Date_Entities_To_Simile

Algorithm Parameters

Algorithm Name: Meandre_Spellcheck_Report_Per_Volume

Algorithm Description: This spellcheck flow will load data and create several reports with results of spellchecking. This version of spellcheck is set to look for and suggest replacements for OCR errors. Loads each page of each volume from HTRC. For the html report, several spelling statistics are provided at a volume level. For the text file reports, information for each volume is displayed with blank line separating the volumes. There are options to customize the dictionary, token counts, and transformation rules. The token counts data is used to determine if a suggested dictionary word occurs in the token counts data and should be used. There are options for customizing the transformation rules which indicate the types of OCR errors that should be corrected. For instance a known problem is the transformation of an "li" to an "h" and vice a versa, so this is expressed with the transformation rule "li=h" which says for all misspelled words with an "h" check if conversion to "li" forms a correctly spelled word.

Version: 1.0

Algorithm Author: Loretta Auvil;

Please Input Job Name: (required)

dickens-spell-check

Please select a collection for analysis:

Charles_Dickens_Novels

Please provide a text for transformation, e.g. h=li; li=h; rn=m; m=rn; s=f; (default: long list of known OCR errors):

(optional)

Please provide a url that contains a text file of words considered in the dictionary (default: http://repository.seasr.org/Datasets/Text/dict):

easr.org/Datasets/Text/dict (optional)

Please provide a url for token counts that can be used for choosing the best correctly spelled word based on popularity.:

(optional)

Submit

HathiTrust Research Center

Manage Results

HOME ABOUT COLLECTION ALGORITHMS RESULTS HELP LOGOUT

Active Jobs

Job Title	Last Updated	Status	Cancel?
dickens-spell-check	2012-09-09 14:11:04	Running	☐

Cancel

Completed Jobs

Job Title	Last Updated	Status	Delete?
dickens-entities	2012-09-06 10:37:50	Finished	☐

Delete

HathiTrust Research Center

Manage Results

HOME ABOUT COLLECTION ALGORITHMS RESULTS HELP LOGOUT

Active Jobs

Job Title	Last Updated	Status	Cancel?

Cancel

Completed Jobs

Job Title	Last Updated	Status	Delete?
dickens-entities	2012-09-06 10:37:50	Finished	☐
dickens-spell-check	2012-09-09 14:28:35	Finished	☐

Delete

HathiTrust Research Center

View Results

Job Details

Job Title: dickens-spell-check

Algorithm Name:
Meandre_Spellcheck_Report_Per_Volume

Last Updated: 2012-09-10 10:33:05

Results:

- stderr.txt
- replacement_rules.txt
- misspellings.txt
- stdout.txt
- misspellings_with_counts.txt
- spellcheck_report.html

Job Parameters:

input_collection: Charles_Dickens_Novels

dictionary: http://repository.seasr.org /Datasets/Text/dict

Job Id:
cd6b2c36-1f09-4ce7-933d-01f5f2928b1f

Status:
Finished

View Results

```
irrepairabel,1
natiirally,1
chequ,1
jem,1
daventry,1
comin,1
kickin,1
muggleton,1
observashuns,1
granby,1
jfc,1
fev,1
cattermole,1
affi,1
visdom,1
rewerse,1
contem,1
craddock,1
curice,1
sentative,1
conweniently,1
municated,1
wolunteers,1
eooms,1
assem,1
palin,1
ner,1
everythin,1
vorldly,1
nev,1
munication,1
stockin,1
dombey,1
reachin,1
nee,1
```

HathiTrust Research Center

View Results

HOME ABOUT COLLECTION ALGORITHMS RESULTS HELP LOGOUT

Job Details

Job Title: dickens-spell-check

Algorithm Name:
Meandre_Spellcheck_Report_Per_Volume

Last Updated: 2012-09-10 10:33:05

Results:

- stderr.txt
- replacement_rules.txt
- misspellings.txt
- stdout.txt
- misspellings_with_counts.txt
- spellcheck_report.html

Job Parameters:

input_collection: Charles_Dickens_Novels

dictionary: http://repository.seasr.org/Datasets/Text/dict

Job Id:
cd6b2c36-1f09-4ce7-933d-01f5f2928b1f

Status:
Finished

View Results

countTotalWords	countMisspelledWords	countCorrectedWords	volume_id
16595	1866	114	uc2.ark:/13960/t5j962x5v
14341	974	107	uc2.ark:/13960/t6m041m4z
7235	399	31	mdp.39015063512019
19773	3887	643	uva.x000983213
19160	4348	516	nyp.33433074954060
19967	3940	273	nyp.33433074954375
15200	1504	48	pst.000062841405
8991	1765	89	uiuo.ark:/13960/t9g45d53j
20098	4032	598	mdp.39015056712683
10814	1176	69	uva.x000892091

HathiTrust Research Center

Submit Algorithms

HOME ABOUT COLLECTION ALGORITHMS RESULTS HELP LOGOUT

Available Algorithms

- Meandre_Topic_Modeling
- Marc_Downloader
- Simple_Deployable_Word_Count
- Meandre_Spellcheck_Report_Per_Volume
- Meandre_Tag_Cloud_with_Cleaning
- Meandre_OpenNLP_Entity_List
- Meandre_Dunning_Loglikelihood
- HTRC_Meandre_Entity_Social_Network
- Meandre_Simple_Tag_Cloud
- Meandre_OpenNLP_Date_Entities_To_Simile

Algorithm Parameters

Algorithm Name: Meandre_Simple_Tag_Cloud

Algorithm Description: Counts the tokens for all volumes and displays the top 200 tokens in a tag cloud. No cleaning of the text is performed.

Version: 1.0

Algorithm Author: Loretta Auvil;

Please Input Job Name: (required)

simple-tag-cloud-dickens

Please select a collection for analysis:

Charles_Dickens_Novels

Submit

HathiTrust Research Center

Manage Results

HOME ABOUT COLLECTION ALGORITHMS RESULTS HELP LOGOUT

Active Jobs

Job Title	Last Updated	Status	Cancel?

Cancel

Completed Jobs

Job Title	Last Updated	Status	Delete?
dickens-entities	2012-09-06 10:37:50	Finished	☐
dickens-spell-check	2012-09-09 14:28:35	Finished	☐
simple-tag-cloud-dickens	2012-09-09 14:30:30	Finished	☐

Delete

HathiTrust Research Center

View Results

HOME ABOUT COLLECTION ALGORITHMS RESULTS HELP LOGOUT

Job Details

Job Title: simple-tag-cloud-dickens
Algorithm Name: Meandre_Simple_Tag_Cloud
Last Updated: 2012-09-09 14:30:30
Results:

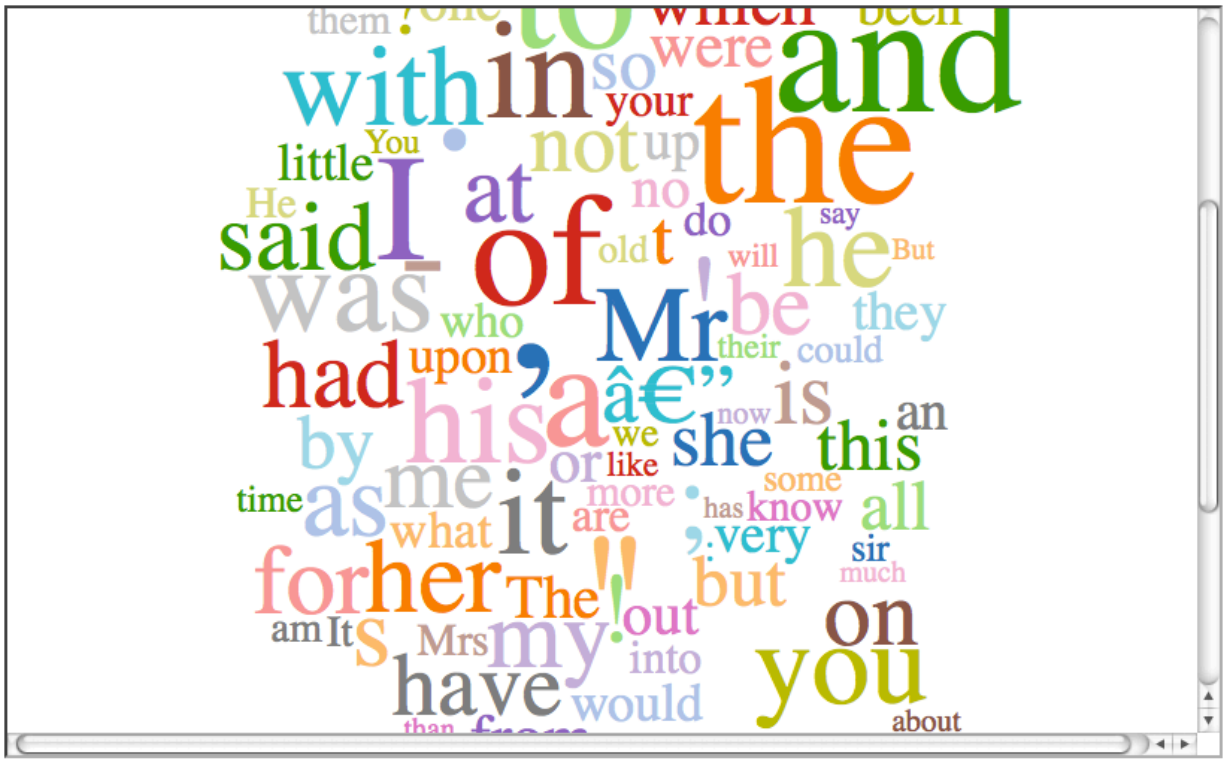
- stdout
- stderr
- skowalcz/de54044f-20de-4d1e-8f09-4148afd3473d/job_results/tagcloudtokencounts.html

Job Parameters:

input_collection: Charles_Dickens_Novels

Job Id:
de54044f-20de-4d1e-8f09-4148afd3473d
Status:
Finished

View Results



HathiTrust Research Center

Submit Algorithms

HOME ABOUT COLLECTION ALGORITHMS RESULTS HELP LOGOUT

Available Algorithms

- Meandre_Topic_Modeling
- Marc_Downloader
- Simple_Deployable_Word_Count
- Meandre_Spellcheck_Report_Per_Volume
- Meandre_Tag_Cloud_with_Cleaning
- Meandre_OpenNLP_Entity_List
- Meandre_Dunning_Loglikelihood
- HTRC_Meandre_Entity_Social_Network
- Meandre_Simple_Tag_Cloud
- Meandre_OpenNLP_Date_Entities_To_Simile

Algorithm Parameters

Algorithm Name: Meandre_Tag_Cloud_with_Cleaning

Algorithm Description: Loads each page of each volume from HTRC. Removes the first and last line of each page. Joins hypenated words that occur at the end of the line. Removes all tokens that don't consist of alphanumeric characters. Uses the replacement rules (learned from our usage of Google Ngrams data) to clean OCR errors, normalize to British spelling and normalize for period spelling. Filters stop words. Counts the tokens remaining for all volumes and displays the top 200 tokens in a tag cloud.

Version: 1.0

Algorithm Author: Loretta Auvil;

Please Input Job Name: (required)

clean-dickens-tag-cloud

Please select a collection for analysis:

Charles_Dickens_Novels

Please provide a url that contains a text file of replacement rules (default: http://repository.seasr.org/Datasets/Text/ngram_corrections.txt):

Text/ngram_corrections.txt (optional)

Please provide a url that contains a text file of stop words to be used (default: http://repository.seasr.org/Datasets/Text/common_words.txt):

ts/Text/common_words.txt (optional)

Please provide the number of tokens to be displayed in the tagcloud (default: 200):

200 (optional)

Submit

HathiTrust Research Center

Manage Results

HOME ABOUT COLLECTION ALGORITHMS RESULTS HELP LOGOUT

Active Jobs

Job Title	Last Updated	Status	Cancel?
clean-dickens-tag-cloud	2012-09-09 14:47:26	Running	☐
simile-entities	2012-09-09 14:39:08	Running	☐

Cancel

Completed Jobs

Job Title	Last Updated	Status	Delete?
dickens-entities	2012-09-06 10:37:50	Finished	☐
dickens-spell-check	2012-09-09 14:28:35	Finished	☐
simple-tag-cloud-dickens	2012-09-09 14:30:30	Finished	☐

Delete

HathiTrust Research Center

View Results

HOME ABOUT COLLECTION ALGORITHMS RESULTS HELP LOGOUT

Job Details

Job Title: clean-dickens-tag-cloud

Algorithm Name:
Meandre_Tag_Cloud_with_Cleaning

Last Updated: 2012-09-10 10:03:36

Results:

- stderr.txt
- stdout.txt
- tagcloudcleantokencounts.html

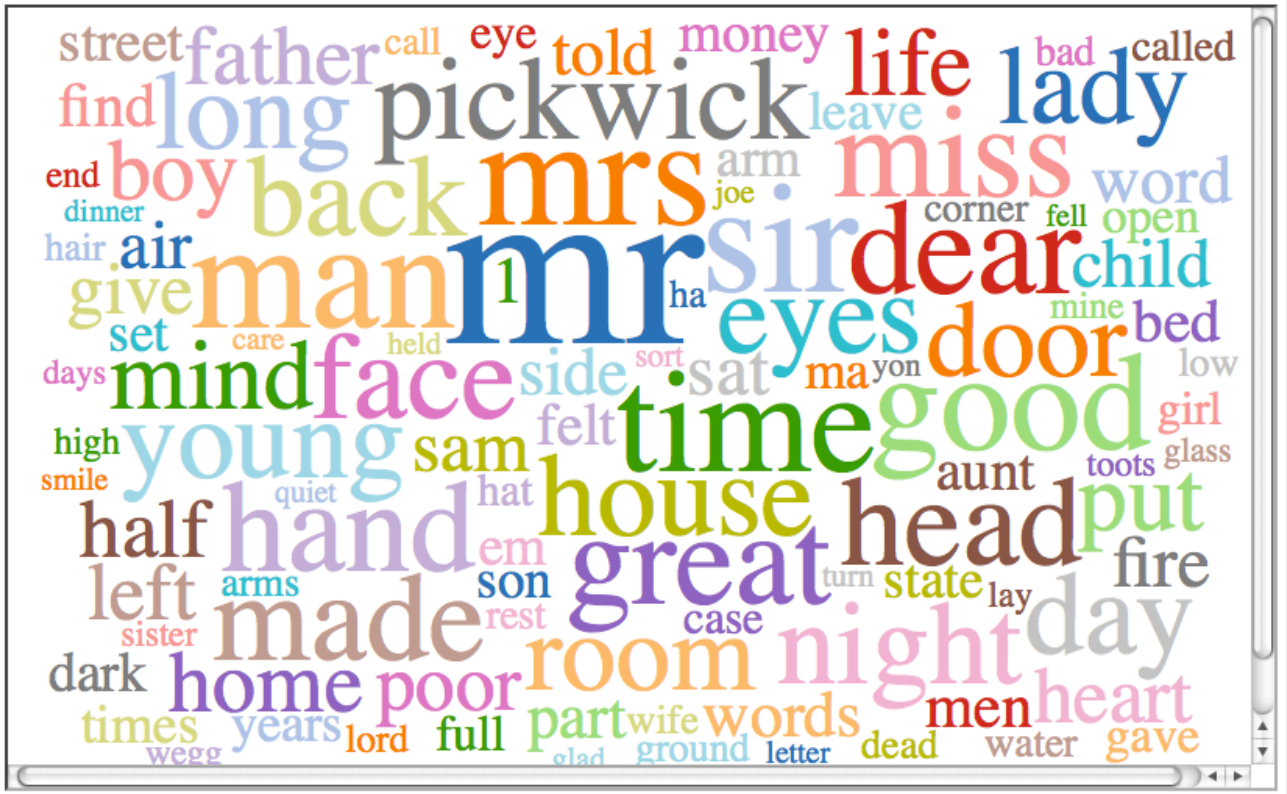
Job Parameters:

- n_top_tokens: 200
- stopwords_list_english_url: http://repository.seasr.org/Datasets/Text/common_words.txt
- input_collection: Charles_Dickens_Novels
- replacement_rules_url: http://repository.seasr.org/Datasets

Job Id:
6e53512c-e8f5-4118-b2ff-81acddc21af8

Status:
Finished

View Results



HathiTrust Research Center

Submit Algorithms

HOME ABOUT COLLECTION ALGORITHMS RESULTS HELP LOGOUT

Available Algorithms

- Meandre_Topic_Modeling
- Marc_Downloader
- Simple_Deployable_Word_Count
- Meandre_Spellcheck_Report_Per_Volume
- Meandre_Tag_Cloud_with_Cleaning
- Meandre_OpenNLP_Entity_List
- Meandre_Dunning_Loglikelihood
- HTRC_Meandre_Entity_Social_Network
- Meandre_Simple_Tag_Cloud
- Meandre_OpenNLP_Date_Entities_To_Simile

Algorithm Parameters

Algorithm Name: Meandre_OpenNLP_Entity_List

Algorithm Description: Loads each page of each volume from HTRC. Removes the first and last line of each page. Joins hypenated words that occur at the end of the line. Extracts entity types specified from the text. Displays each entity with the volume_id, page_id, sentence_id and character position within the sentence.

Version: 1.0

Algorithm Author: Loretta Auvil;

Please Input Job Name: (required)

dickens-entities

Please select a collection for analysis:

Charles_Dickens_Novels

Please provide a comma separated list of entity types to be extracted. Acceptable values are: date, location, money, organization, percentage, person, time. (default: person):

person (optional)

Submit

HathiTrust Research Center

View Results

HOME ABOUT COLLECTION ALGORITHMS RESULTS HELP LOGOUT

Job Details

Job Title: dickens-entities

Algorithm Name:
Meandre_OpenNLP_Entity_List

Last Updated: 2012-09-10 10:06:00

Results:

- stderr.txt
- stdout.txt
- named_entity_list.html

Job Parameters:

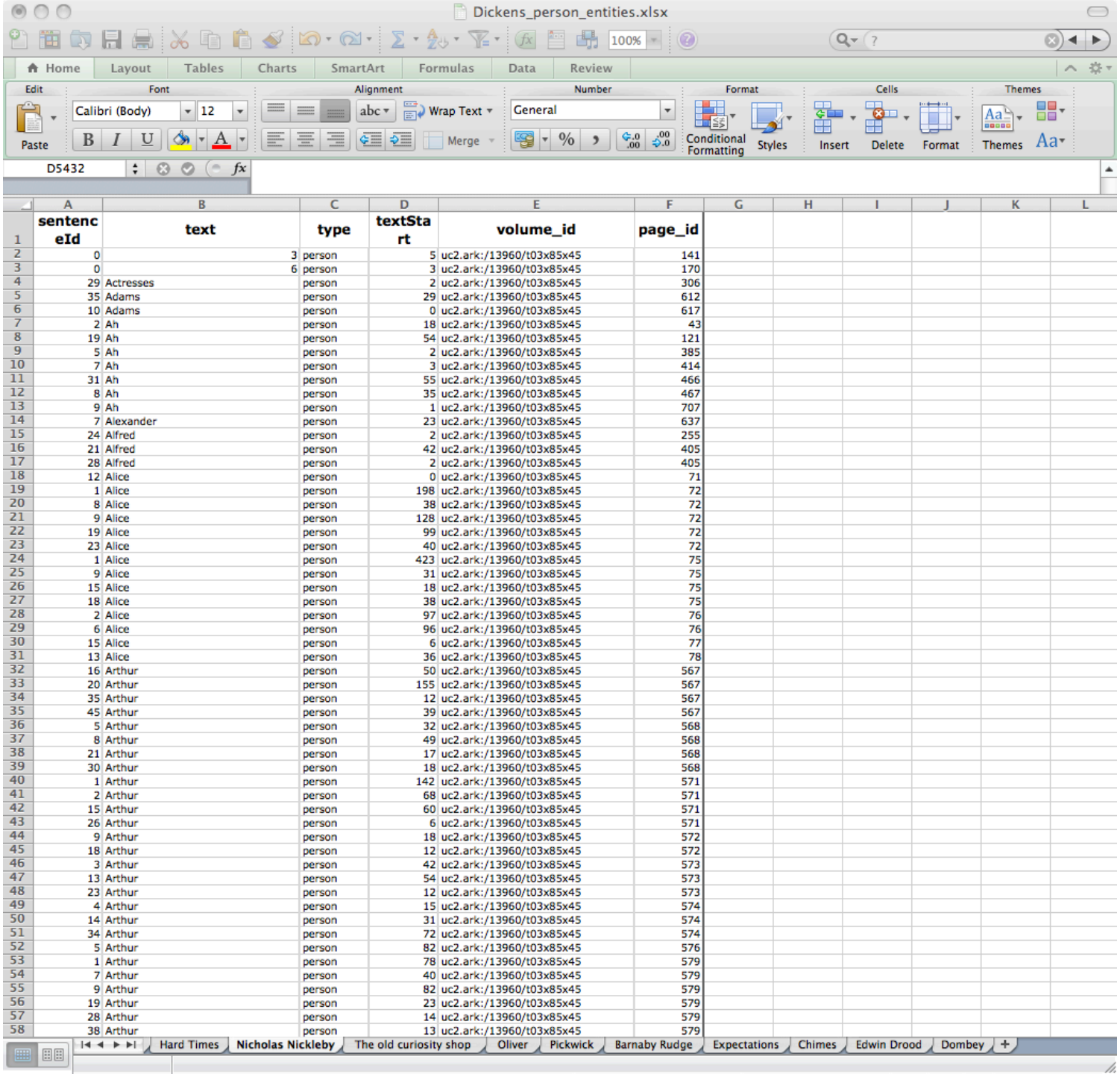
- entity_types: person
- input_collection: Charles_Dickens_Novels

Job Id:
089a2b19-463f-4fcd-a1cd-2fbc55e4f972

Status:
Finished

View Results

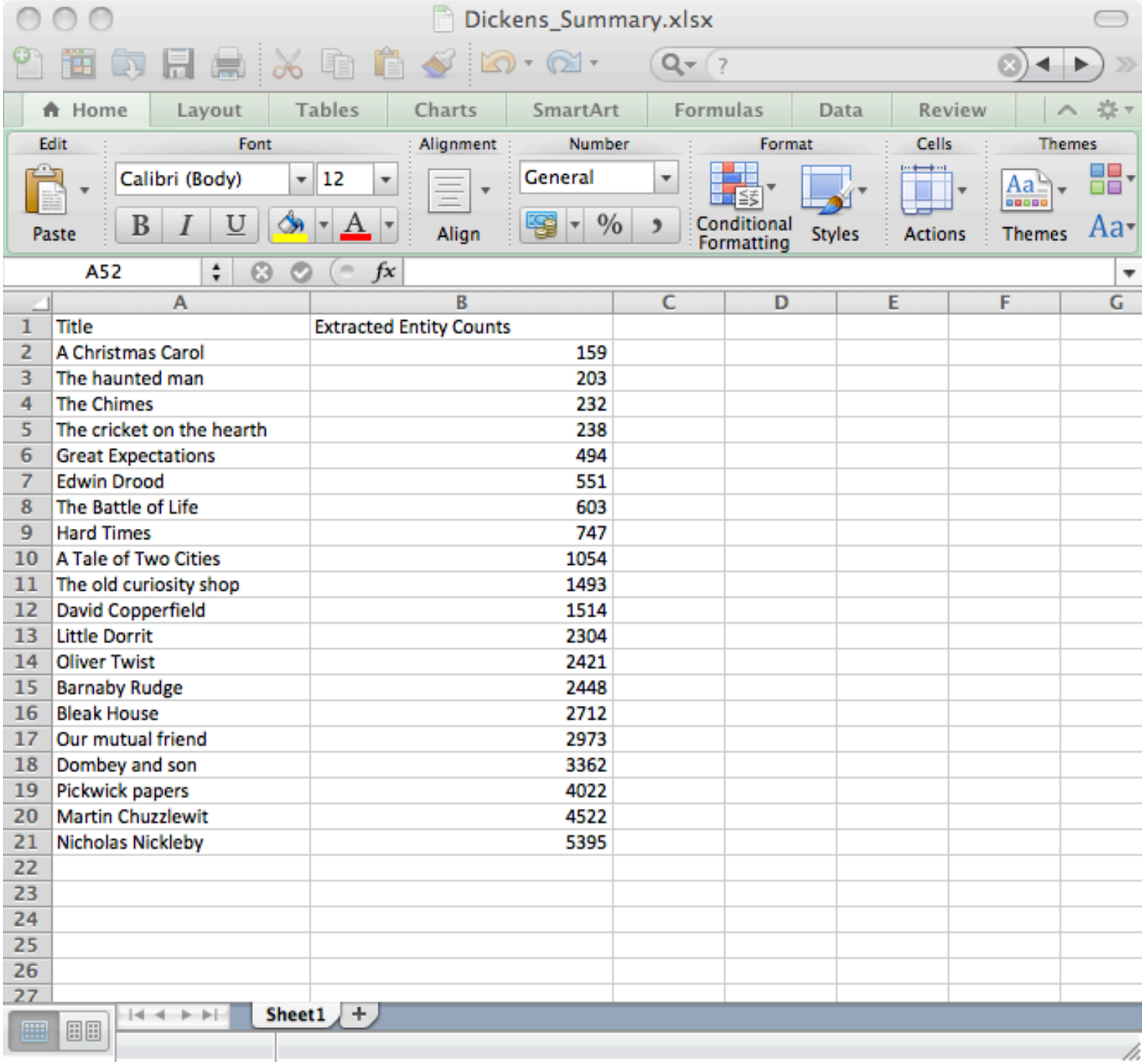
sentenceId	text	type	textStart	volume_id	page_id
43	Charles Dickens	person	90	uc2.ark:/13960/t5j962x5v	6
8	EDWARD CHAPMAN	person	0	uc2.ark:/13960/t5j962x5v	14
18	EDWARD CHAPMAN	person	0	uc2.ark:/13960/t5j962x5v	15
20	EDWARD CHAPMAN	person	0	uc2.ark:/13960/t5j962x5v	15
1	Moses	person	27	uc2.ark:/13960/t5j962x5v	16
6	Ebenezer Chapel	person	300	uc2.ark:/13960/t5j962x5v	16
3	Laws	person	235	uc2.ark:/13960/t5j962x5v	17
7	MRS.	person	22	uc2.ark:/13960/t5j962x5v	19
4	Joseph Smiggers	person	0	uc2.ark:/13960/t5j962x5v	21
5	Samuel Pickwick	person	191	uc2.ark:/13960/t5j962x5v	21
5	Samuel Pickwick	person	427	uc2.ark:/13960/t5j962x5v	21
6	Samuel Pickwick	person	215	uc2.ark:/13960/t5j962x5v	21
1	Samuel Pickwick	person	139	uc2.ark:/13960/t5j962x5v	22
3	Samuel Pickwick	person	99	uc2.ark:/13960/t5j962x5v	22
3	G.C.M.P.C.	person	122	uc2.ark:/13960/t5j962x5v	22
3	Tracy Tupman	person	136	uc2.ark:/13960/t5j962x5v	22
3	Nathaniel Winkle	person	204	uc2.ark:/13960/t5j962x5v	22
2	Mr. Tracy	person	22	uc2.ark:/13960/t5j962x5v	25
3	Ho	person	0	uc2.ark:/13960/t5j962x5v	26
40	He	person	0	uc2.ark:/13960/t5j962x5v	26
17	Tommy	person	22	uc2.ark:/13960/t5j962x5v	27
28	Sam	person	16	uc2.ark:/13960/t5j962x5v	28
29	Sam	person	18	uc2.ark:/13960/t5j962x5v	28
4	Brown	person	8	uc2.ark:/13960/t5j962x5v	33
15	Ah	person	2	uc2.ark:/13960/t5j962x5v	33
31	Ah	person	2	uc2.ark:/13960/t5j962x5v	33



Dickens_person_entities.xlsx

sentenceId	text	type	textStart	volume_id	page_id
0	3	person	5	uc2.ark:/13960/t03x85x45	141
0	6	person	3	uc2.ark:/13960/t03x85x45	170
29	Actresses	person	2	uc2.ark:/13960/t03x85x45	306
35	Adams	person	29	uc2.ark:/13960/t03x85x45	612
10	Adams	person	0	uc2.ark:/13960/t03x85x45	617
2	Ah	person	18	uc2.ark:/13960/t03x85x45	43
19	Ah	person	54	uc2.ark:/13960/t03x85x45	121
5	Ah	person	2	uc2.ark:/13960/t03x85x45	385
7	Ah	person	3	uc2.ark:/13960/t03x85x45	414
31	Ah	person	55	uc2.ark:/13960/t03x85x45	466
8	Ah	person	35	uc2.ark:/13960/t03x85x45	467
9	Ah	person	1	uc2.ark:/13960/t03x85x45	707
7	Alexander	person	23	uc2.ark:/13960/t03x85x45	637
24	Alfred	person	2	uc2.ark:/13960/t03x85x45	255
21	Alfred	person	42	uc2.ark:/13960/t03x85x45	405
28	Alfred	person	2	uc2.ark:/13960/t03x85x45	405
12	Alice	person	0	uc2.ark:/13960/t03x85x45	71
1	Alice	person	198	uc2.ark:/13960/t03x85x45	72
8	Alice	person	38	uc2.ark:/13960/t03x85x45	72
9	Alice	person	128	uc2.ark:/13960/t03x85x45	72
19	Alice	person	99	uc2.ark:/13960/t03x85x45	72
23	Alice	person	40	uc2.ark:/13960/t03x85x45	72
1	Alice	person	423	uc2.ark:/13960/t03x85x45	75
9	Alice	person	31	uc2.ark:/13960/t03x85x45	75
15	Alice	person	18	uc2.ark:/13960/t03x85x45	75
18	Alice	person	38	uc2.ark:/13960/t03x85x45	75
2	Alice	person	97	uc2.ark:/13960/t03x85x45	76
6	Alice	person	96	uc2.ark:/13960/t03x85x45	76
15	Alice	person	6	uc2.ark:/13960/t03x85x45	77
13	Alice	person	36	uc2.ark:/13960/t03x85x45	78
16	Arthur	person	50	uc2.ark:/13960/t03x85x45	567
20	Arthur	person	155	uc2.ark:/13960/t03x85x45	567
35	Arthur	person	12	uc2.ark:/13960/t03x85x45	567
45	Arthur	person	39	uc2.ark:/13960/t03x85x45	567
5	Arthur	person	32	uc2.ark:/13960/t03x85x45	568
8	Arthur	person	49	uc2.ark:/13960/t03x85x45	568
21	Arthur	person	17	uc2.ark:/13960/t03x85x45	568
30	Arthur	person	18	uc2.ark:/13960/t03x85x45	568
1	Arthur	person	142	uc2.ark:/13960/t03x85x45	571
2	Arthur	person	68	uc2.ark:/13960/t03x85x45	571
15	Arthur	person	60	uc2.ark:/13960/t03x85x45	571
26	Arthur	person	6	uc2.ark:/13960/t03x85x45	571
9	Arthur	person	18	uc2.ark:/13960/t03x85x45	572
18	Arthur	person	12	uc2.ark:/13960/t03x85x45	572
3	Arthur	person	42	uc2.ark:/13960/t03x85x45	573
13	Arthur	person	54	uc2.ark:/13960/t03x85x45	573
23	Arthur	person	12	uc2.ark:/13960/t03x85x45	573
4	Arthur	person	15	uc2.ark:/13960/t03x85x45	574
14	Arthur	person	31	uc2.ark:/13960/t03x85x45	574
34	Arthur	person	72	uc2.ark:/13960/t03x85x45	574
5	Arthur	person	82	uc2.ark:/13960/t03x85x45	576
1	Arthur	person	78	uc2.ark:/13960/t03x85x45	579
7	Arthur	person	40	uc2.ark:/13960/t03x85x45	579
9	Arthur	person	82	uc2.ark:/13960/t03x85x45	579
19	Arthur	person	23	uc2.ark:/13960/t03x85x45	579
28	Arthur	person	14	uc2.ark:/13960/t03x85x45	579
38	Arthur	person	13	uc2.ark:/13960/t03x85x45	579

Hard Times | Nicholas Nickleby | The old curiosity shop | Oliver | Pickwick | Barnaby Rudge | Expectations | Chimes | Edwin Drood | Dombey



	A	B	C	D	E	F	G
1	Title	Extracted Entity Counts					
2	A Christmas Carol	159					
3	The haunted man	203					
4	The Chimes	232					
5	The cricket on the hearth	238					
6	Great Expectations	494					
7	Edwin Drood	551					
8	The Battle of Life	603					
9	Hard Times	747					
10	A Tale of Two Cities	1054					
11	The old curiosity shop	1493					
12	David Copperfield	1514					
13	Little Dorrit	2304					
14	Oliver Twist	2421					
15	Barnaby Rudge	2448					
16	Bleak House	2712					
17	Our mutual friend	2973					
18	Dombey and son	3362					
19	Pickwick papers	4022					
20	Martin Chuzzlewit	4522					
21	Nicholas Nickleby	5395					
22							
23							
24							
25							
26							
27							