Collection Sorting and Dataset Cleaning

In this notebook, we created an algorithm to sort the dataset into collections. This proved difficult, as there was no one way to tell a card from a narrative using the text alone. Thus, we studied the cards in an attempt to find patterns that suggest the type of card a row belonged to. A new author name signified a new collection automatically, but within an author's associated cards, we had to get more creative on how to differentiate between main entries and narrative cards.

Packages used: Pandas and numpy were used for basic data manipulation of the dataset. Glob was used to identify the needed files on the computer. Regex, NLTK, and string were used for natural language processing of the text, in order to determine a card's type. SpaCy is used to identify locations present in the main entry cards. Sklearn was used to predict important keywords in the cards.

In [1]:	_	<pre>con('display.max_colwidth', None) con("display.max_row", None) by as np</pre>		
	Collectio	n Sort		
In [2]:	<pre>def is_main # Check if inde row # Check elif in ret # Calcu main_en # Check if len(nar elif le mai if "Sec mai if "GUI mai if "Check if re.s nar elif (r nar elif re.s nar elif re.s nar elif re.s nar elif re.s mai # Check if re.s mai elif re.s</pre>	<pre>that calculates "points" for a card to b _entry(index, row, df): if previous and subsequent card authors x > 0 and index < len(df)-1 and df.iloc[['Name'] = df.iloc[index-1]['Name'] c for new author -> automatic "main entry dex == 0 or row['Name'] != df.iloc[index carn True thate "points" for classification try_pts, narrative_pts = 0, 0 c length, longer -> narrative, shorter -> rrow['Text']) < 300: crative_pts += 1 m(row('Text']) < 200: .n_entry_pts += 1 c for misc. main entry identifiers ("GUID trition" in row['Text']: .n_entry_pts += 1 DE" in row['Text'].upper() or "GUID" in .n_entry_pts += 3 c for misc. narrative identifiers c INFORMATION ABOUT THIS COLLECTION" in row (rative_pts += 1 re.search(r'\d+", row['Text']): rrative_pts += 1 re.search(r'\d+", row['Text']): rrative_pts += 1 revious card indicates continuation of n "(Continued on next card)" in df.iloc[index-1]['Ter narrative_pts += 1 rerative_pts += 1 revious card indicates continuation of n "(Continued on next card)" in df.iloc[index-1]['Ter marrative_pts += 2 com('Name'].split(",")[0]).upper() in row rrative_pts += 2 c for capitalized last name & card number isearch(r"(card [0-9]+", row['Text']): rrative_pts += 2 com('Name'].split(",")[0]).upper() in row rrative_pts += 3 c for accessioned date (e.g., 13 MAR 97, usearch(r"\d1,2)\s[A-2][3]\s\d[2]", row['Text .n_entry_pts += 1 c.search(r"\d1,2]\s[A-2][3]\s\d[2]", row 'Text .n_entry_pts += 1 c.search(r"\d1,2]\s[A[1,2]\s\d[4]", row</pre>	<pre>match -> fix mispelled name index-1]['Name'] == df.iloc[: " -1]['Name']: main entry E", "Section X", "X items") row['Text'].upper() or "GUDE ow['Text']: 'Text']: dex-1]['Text']) == str: xt'][:50]: arrative onto current card dex-1]['Text'] or "SEE NEXT of at beginning of text (Card of at beginning of at beginning of text (Card o</pre>	<pre>in between index+1]['Name'] and row['Name'] -> main entry " in row['Text'].upper(): CRD" in df.iloc[index-1]['Text'] 2, JONES, Smith, John 3, etc.) -</pre>
	mai	.n_entry_pts += 2	['Text']):	
	if main	<pre>cmine card classification a_entry_pts >= narrative_pts: curn True False</pre>		
	<pre># Check collect for ind</pre>	<pre># Main entries signify a new collection if is_main_entry(index, row,df): collection_num += 1 collection_list.append(collection_num) column to dataframe with collections numb lection'] = collection_list by collection headers as index, with or re_No_Within_Drawer"] = df.index -100 tion_header = [] Mex,row in df.iterrows(): rrow("Collection"]!= curr): collection_header.append(row["Name"]) se: collection_header.append("") rr = row["Collection"] lection_Head"] = collection_header df drawers and combine them name = (glob.glob(r"catalog\file_name_chee : = [gort_collection(i) for i in file_list_name : = [sort_collection(i) for i in file_list_name : = [sort_collection(i) for i in file_list_name to collect (sorted_list) te.to_csv("catalog/csv_iterations/all_sor pd.read_csv(csv_iterations, we can export the new pd.read_csv(csv_iterations, we can export the new pd.read_csv(csv_iterations) as pd. pd. pd. pd. pd. pd. pd. pd. pd. pd.</pre>	<pre>e']) or row['Text'] == "" or ered iginal index cked/*")) t] td ted_before_year_column.csv", sorted_before_year_column.cs / dataset. ow is a collection header</pre>	<pre>row['Name'] == "": index=False) v")</pre>
	elif (n! li.			
	n=i	<pre>append(0) is collection head"] = li</pre>		
	all_text["1 all_text.he	— —		
Out[26]:	Name			Collection_Head Is_collection_head
	0 Davis A and Company	A. B. Davis and Company (Philadelphia, Pa.) See Davis (A. B.) and Company		A. B. Davis and 1 Company 1
		AeHe Roscoe (Firm: Nashville, Tenne)e Journal, 1853, Septe- 1857, Dece 1 item(800 ppe)e Wholesale and retail druggist and dealer in paints, oils, and dyestuffse Summary: Journal (account book) documents the sale of chemical and herbal drugs, paint and painting supplies, dyestuffs, personal and household supplies, and garden seeds to individuals and businessese le DrugstoresTennesseee 2e Paint shops Equipment and supplies Tennesseee 3e Dyes and dyeinge 46 Household suppliesTennesseee 5c Herbs Therapeutic usee 6+ Seed industry and tradeTennessees 7e Nashville ite® nne)—Commercee 20 MAY QO1 23804905 NDHYme	1 1	A.H. Roscoe 1
	2 Abbeville District	Abbeville District (8.C.) See South Carolina. Abbeville District	2 2	Abbeville District 1
	3 Abbott, William B	Abbott, William B. Papers, 1862-1864 Frederick Co., Va. Section A 5-16-57 GUIDE 10 items	3 3	Abbott, William 1 B
		 Abbott, William B. Papers, 1862-1864, Fre- Gerick Co., Va. 10 items. Sketch These are the papers of William B. Abbott, evidently a well-to-do farmer of Frederick Co., Va. There are several documents concerned with the evaluation of damage done to his property by C. S. A. troops in 1862, and.receipts in 1864 for hay bought from Abbott at various times in Aug., 1864 by the C. S. A. Army. 	3 4	NaN 0
	Extracting	years using regular expressions for	r common patterns	
In [27]:	main_entrie	<pre>year column for main entries es = all_text[all_text["Is_collection_heat tries = all_text[all_text["Is_collection]</pre>		
	<pre># patterns first = "(1 second = "() third = "(1 fourth = "() fifth = "(n year_vals =</pre>	<pre>used to extract sequentially if one does .\d{3}[-]\s?\d{2,4})" (1\d{3}.{0,9}1\d{0,3})" .\d{3})" (1\d{3}[-]c(a)?\\.)" n\\.\sd\\.)" #no date = main_entries.Text.str.extract(first)[0]</pre>	not exist .fillna((# if it returns na	, find next pattern
	main_er mai mai	<pre>ntries.Text.str.extract(second)[0])).fill: .n_entries.Text.str.extract(third)[0])).f. .n_entries.Text.str.extract(fourth)[0])). .n_entries.Text.str.extract(fifth)[0])).v.</pre>	na((illna((fillna((

Out[27]:		Name	Text	Collection	Page_No_Within_Drawer	Collection_Head	ls_collection_head	Year
Out[27]:	0	A. B. Davis and Company	A. B. Davis and Company (Philadelphia, Pa.) See Davis (A. B.) and Company	0	0	A. B. Davis and Company	1	NaN
		A.H. Roscoe	AeHe Roscoe (Firm: Nashville, Tenne)e Journal, 1853, Septe-1857, Dece 1 item(800 ppe)e Wholesale and retail druggist and dealer in paints, oils, and dyestuffse Summary: Journal (account book) documents the sale of chemical and herbal drugs, paint and painting supplies, dyestuffs, personal and household supplies, and garden seeds to individuals and businessese le Drugstores Tennesseee 2e Paint shopsEquipment and supplies Tennesseee 3e Dyes and dyeinge 46 Household suppliesTennesseee 5c Herbs Therapeutic usee 6+ Seed industry and trade Tennessees 7e Nashville ite ® nne)—Commercee 20 MAY QO1 23804905 NDHYme	1	1	A.H. Roscoe	1	1853, Septe- 1857
	2	Abbeville District	Abbeville District (8.C.) See South Carolina. Abbeville District	2	2	Abbeville District	1	NaN
2 3 4 5 6 7	3	Abbott, William B	Abbott, William B. Papers, 1862-1864 Frederick Co., Va. Section A 5-16-57 GUIDE 10 items	3	3	Abbott, William B	1	1862- 1864
	4	Abbott, William B	Abbott, William B. Papers, 1862-1864, Fre- Gerick Co., Va. 10 items. Sketch These are the papers of William B. Abbott, evidently a well-to-do farmer of Frederick Co., Va. There are several documents concerned with the evaluation of damage done to his property by C. S. A. troops in 1862, and receipts in 1864 for hay bought from Abbott at various times in Aug., 1864 by the C. S. A. Army.	3	4	NaN	0	NaN
	5	Abbott & Company	Abbott & Company Papers, 1856-1871 Philadelphia, Pennsylvania a Section A 66 items OCT 10 "49	4	5	Abbott & Company	1	1856- 1871
	6	Abbott & Company	Abderhalden, Emil Papers, 1919 Halle, Germany Josiah C. Trent Collection in the History of Medicine, Div, 4-8-60	5	6	Abbott & Company	1	1919
	7	Abbott & Company	Abbott & Company. Papers 1856-1871 Philadelphia, Pennsylvania, € Miscellaneous letters concerning scales sold by Abbott & Company.	6	7	Abbott & Company	1	1853, 1 Septe- 1857 1 NaN 1 1862- 1864 NaN 1 1862- 1864 1864 1864 1864
	8	Abel, Ernest L.	Abel, Ernest L. Papers, 1922-1952 West Palm Beach, Palm Beach Co., Fla. Saat; 550 items & 8 vols. Box 1 of Labor Archives	7	8	Abel, Ernest L.	1	
	9	Abel, Ernest L.	Abel, Ernest L. Papers. West Palm Beach, Palm Beach Co., Fla' : Ernest L. Abel was involved in postal union activity at least from the mid-1920s to the early 1950s. He served from 1925-1926 as Secretary-Treasurer for Local Union No. 749 of the National Federation of Post Office Clerks (AFL) and was State Representative for the National Federation at the same time. In 1927, Abel became the first president of the Florida Federation of Post Office Clerks, His union service included being President Union No. 749 in 1945, and State Legislative Representative of the' ® Florida	7	9	NaN	0	NaN

Adding start and end date columns for a collection, if applicable

In [28]:

all_text["Year"] =np.nan

all_text.loc[ind_main_entries,"Year"]= year_vals
all_text.head(10)

Start and End date column start,end=[],[] for i in ind_main_entries: year = all_text.Year[i] if type(year) == float and pd.isna(year): start.append(np.nan) end.append(np.nan) continue if(re.match(first,year)): start.append(int(year.split("-")[0])) if(len(year.split("-")[1])==2): end.append(int(year.split("-")[0][:2]+ year.split("-")[1])) else: end.append(int(year.split("-")[1])) elif(re.match(second, year)): # print("Second",year) lt =re.findall("1\d{3}",year) **if**(len(lt)==2): start.append(int(lt[0])) end.append(int(lt[1])) else: start.append(np.nan) end.append(np.nan) elif(re.match(third,year)): # print("Third",year) start.append(int(year)) end.append(np.nan) else: start.append(np.nan) end.append(np.nan) all_text.loc[ind_main_entries, "Start"] = start
all_text.loc[ind_main_entries, "End"] = end all_text.head()

]:	Name	Text	Collection	Page_No_Within_Drawer	Collection_Head	ls_collection_head	Year	NaN NaN 853, pte- 1853.0 1857.0 1853.0 1857.0 NaN NaN NaN NaN NaN NaN 1862- 1864.0 1864.0	End
0	A. B. Davis and Company	A. B. Davis and Company (Philadelphia, Pa.) See Davis (A. B.) and Company	0	0	A. B. Davis and Company	1	NaN	NaN	NaN
1	A.H. Roscoe	AeHe Roscoe (Firm: Nashville, Tenne)e Journal, 1853, Septe- 1857, Dece 1 item(800 ppe)e Wholesale and retail druggist and dealer in paints, oils, and dyestuffse Summary: Journal (account book) documents the sale of chemical and herbal drugs, paint and painting supplies, dyestuffs, personal and household supplies, and garden seeds to individuals and businessese le Drugstores Tennesseee 2e Paint shops Equipment and supplies Tennesseee 3e Dyes and dyeinge 46 Household supplies Therapeutic usee 6+ Seed industry and tradeTennessees 7e Nashville ite ® nne)— Commercee 20 MAY QO1 23804905 NDHYme	1	1	A.H. Roscoe	1	1853, Septe- 1857	1853.0	1857.0
2	Abbeville District	Abbeville District (8.C.) See South Carolina. Abbeville District	2	2	Abbeville District	1	NaN	NaN	NaN
3	Abbott, William B	Abbott, William B. Papers, 1862- 1864 Frederick Co., Va. Section A 5-16-57 GUIDE 10 items	3	3	Abbott, William B	1	1862- 1864	1862.0	1864.0
4	Abbott, William B	Abbott, William B. Papers, 1862- 1864, Fre- Gerick Co., Va. 10 items. Sketch These are the papers of William B. Abbott, evidently a well-to-do farmer of Frederick Co., Va. There are several documents concerned with the evaluation of damage done to his property by C. S. A. troops in 1862, and.receipts in 1864 for hay bought from Abbott at various times in Aug., 1864 by	3	4	NaN	0	NaN	NaN	NaN

Using SpaCy to glean the locations from the main entry card for each collection

In [29]: *# Get locations* import spacy import en_core_web_sm nlp = en_core_web_sm.load() loc_vals = [] for i in main_entries.index: doc = nlp(str(main_entries.Text[i])) l = list() st= set() for ent in doc.ents: if(ent.text in st): continue if (ent.label_ == 'GPE'): l.append(ent.text) st.add(ent.text) elif (ent.label_ == 'LOC'): l.append(ent.text) st.add(ent.text) loc_vals.append(", ".join(l)) all_text["Loc"] =np.nan all_text.loc[ind_main_entries,"Loc"] = loc_vals all_text.head()

the C. S. A. Army.

t[29]:	Name	Text	Collection	Page_No_Within_Drawer	Collection_Head	ls_collection_head	Year	Start	End	Loc
C	A. B. Davis and Company	A. B. Davis and Company (Philadelphia, Pa.) See Davis (A. B.) and Company	0	0	A. B. Davis and Company	1	NaN	NaN	NaN	Pa.
1	A.H. Roscoe	AeHe Roscoe (Firm: Nashville, Tenne)e Journal, 1853, Septe- 1857, Dece 1 item(800 ppe)e Wholesale and retail druggist and dealer in paints, oils, and dyestuffse Summary: Journal (account book) documents the sale of chemical and herbal drugs, paint and painting supplies, dyestuffs, personal and household supplies, and garden seeds to individuals and businessese le Drugstores Tennesseee 2e Paint shopsEquipment and supplies Tennesseee 3e Dyes and dyeinge 46 Household supplies Tennesseee 5c Herbs- -Therapeutic usee 6+ Seed industry and tradeTennessees 7e Nashville ite® nne) Commercee 20 MAY QO1 23804905	1	1	A.H. Roscoe	1	1853, Septe- 1857	1853.0	1857.0	Nashville, Tenne

	Abbeville	3804905 NDHYme e District ee South Abbeville 2	2 Abbeville District	1 NaN	Abbeville NaN NaN South
	Abbott, W Abbott, William B GUIDE	62-1864 c Co., Va. 3	Abbott, William 3 B	1862- 1 1864 1864	Carolina 1862.0 1864.0 Abbott, Va.
	Abbott, W Papers, 186 Fre- Gerick Co items. Sketo are the p William B. evidently a we	52-1864, b., Va. 10 ch These apers of Abbott,			
	farmer of F Co., Va. T Abbott, several doo William B concerned evaluation of done to his p	rederick here are cuments with the damage property	4 NaN	0 NaN	NaN NaN NaN
	by C. S. A. t 1862, and.red 1864 for hay from A various times 1864 by the	ceipts in bought bbott at in Aug.,			
In [30]:	# Get linke identif:	iers, size of each dra	orresponding card in Inte wer and drawer numbers atalog_A_to_Amer', 'rubenstei		Ames to Armh', 'rubens
In [31]:	<pre>real_size = [838, 68 drawer_no = ['157', '209', '210', '211' # create link list n</pre>	30, 935, 720, 639, 787 '158', '159', '160', ', '212', '213', '214' for the dataset	, 765, 776, 730, 681, 707, 77 '161', '162', '163', '164', ' , '215', '217', '218', '219',	77, 798, 645, 790, 7 '165', '166', '167', , '220', '221', '222	41, 654, 760, 581, 658 '169', '170', '171',
	<pre>link_list = [] for i in range(len() curr_size = real curr_ident = lin # print(curr_size for i in range(corr_size))</pre>	<pre>link_identifier_list)) L_size[i] hk_identifier_list[i] ze,curr_ident) curr_size):</pre>		identifier_list]	
	<pre>drawer_no_list = [] for i in range(len(n for j in range(n))</pre>				
In [32]:	<pre>all_text["Link"] = 1 all_text["Drawer_No" all_text.head()</pre>				
Out[32]:	A. B. Davis and A. B. Company Davis (Philadelphia, and Pa.) See	Collection Page_No_Within	_ Drawer Collection_Head Is_collection _ O A. B. Davis and Company	on_head Year Start	End Loc NaN Pa. https://arch
	Company Davis (A. B.) and Company AeHe Roscoe (Firm: Nashville,				
	Tenne)e Journal, 1853, Septe- 1857, Dece 1 item(800 ppe)e Wholesale				
	and retail druggist and dealer in paints, oils, and dyestuffse Summary:				
	Journal (account book) documents the sale of chemical and herbal drugs,				
	paint and painting supplies, dyestuffs, personal and household A.H. supplies, and	1	1 A.H. Roscoe	1853, 1 Septe- 1853.0 1	1857.0 Nashville, https://arch
	Roscoe garden seeds to individuals and businessese le Drugstores Tennesseee	·		1857	Tenne
	2e Paint shops Equipment and supplies Tennesseee 3e Dyes and				
	dyeinge 46 Household supplies Tennesseee 5c Herbs Therapeutic usee 6+				
	Seed industry and trade Tennessees 7e Nashville ite® nne)—- -Commercee				
	20 MAY QO1 23804905 NDHYme Abbeville District (8.C.) See South	2	2 Abbeville District	1 NaN NaN	Abbeville District, NaN Conthe https://arch
	District Carolina. Abbeville District Abbott, William B. Papers, 1862-1864	-			Carolina
	Abbott, William B William B Section A 5- 16-57 GUIDE 10 items Abbott,	3	Abbott, William 3 B	1 1862- 1862.0 1 1864	Abbott, https://arch Va.
	William B. Papers, 1862-1864, Fre- Gerick Co., Va. 10 items. Sketch These are the				
	papers of William B. Abbott, evidently a well-to-do farmer of Frederick				
	4 Abbott, William B Co., Va. There are several documents concerned with the evaluation of	3	4 NaN	0 NaN NaN	NaN NaN https://arch
	damage done to his property by C. S. A. troops in 1862, and.receipts				
	in 1864 for hay bought from Abbott at various times in Aug., 1864 by the C. S.				
In [35]:	all_text.rename(colu	<pre>'Collection_Head",inpl umns = {"Page_No_Withis</pre>	ace =True) n_Drawer":"Page_drawer","Is_c ,"Loc","Start","End", "Page_c		
In [36]: Out[36]:	<pre>all_text.head(2) Name</pre>	Text Year	Loc Start End Page_drawer D	rawer_No	
-	Collection_Head A. B. A. B. Davis and Company and Company	A. B. Davis and Company (Philadelphia, Pa.) See Davis (A. B.)	Pa. NaN NaN O	157 https://archive.or	rg/details/rubensteinmanuscript
		and Company AeHe Roscoe (Firm: Nashville, Tenne)e			
		Journal, 1853, Septe- 1857, Dece 1 item(800 ppe)e Wholesale and retail druggist and			
		dealer in paints, oils, and dyestuffse Summary: Journal (account			
		book) documents the sale of chemical and herbal drugs, paint and painting			
	A.H. A.H. Roscoe Roscoe	supplies, dyestuffs, personal and household supplies, and garden seeds to individuals	/ille, 1853.0 1857.0 1 nne	157 https://archive.or	rg/details/rubensteinmanuscript
		and businessese le Drugstores Tennesseee 2e Paint shops			
		Equipment and supplies Tennesseee 3e Dyes and dyeinge 46 Household			
		supplies Tennesseee 5c Herbs Therapeutic usee 6+ Seed industry and			
		trade Tennessees 7e Nashville ite® nne)—- -Commercee 20 MAY QO1 23804905			
In [37]:		NDHYme rds column using T L_sorted_collection.cs			
In [4]:	<pre>df = df[~df.Text.isr from nltk.corpus imp from nltk.tokenize i stop_words = stopword</pre>	<pre>port stopwords import word_tokenize rds.words('english')</pre>	sv")		
	<pre>stop_words.extend([' "december","january' def clean(row): if(type(row))==f return row</pre>	ollection","family","cd 'dec', 'jan', 'feb', 'n ',"febraury","march",", float:	o","also","added","see","sket mar', 'apr', 'may', 'jun', 'j april","may","june","july","a	jul', 'aug', 'sep',	'oct', 'nov',
	<pre>full_no_punc = ' # Tokenize text_tokens = wo # # Change to log</pre>	e.sub('[%s]' % re.escap ' ".join([i for i in f ord_tokenize(full_no_p ower case and select o	nly alphabet)>1])	
Out[4]:	<pre>df["Clean"] = df.Tex df = df.iloc[:,[1,2, df.head() Name</pre>	(pre_process) kt.apply(clean)	t_tokens if i.isalpha() and i Text	I. IOWEI () HOL IN Sto	p_words j Clean
 -	A. B. Davis and Company AeHe Roscoe ppe)e	(Firm: Nashville, Tenne)e Journa Wholesale and retail druggist an	a, Pa.) See Davis (A. B.) and Company II, 1853, Septe-1857, Dece 1 item(800 d dealer in paints, oils, and dyestuffse	aehe roscoe firm nashville ter wholesale retail druggist d	adelphia pa davis company nne journal septe dece ppe lealer paints oils dyestuffse
	A.H. Roscoe shopsEquipr supplies	ry: Journal (account book) docu and painting supplies, dyestuffs, eds to individuals and businesses ment and supplies-—- Tennessee -Tennesseee 5c HerbsTherapeu	ments the sale of chemical and herbal personal and household supplies, and se le DrugstoresTennesseee 2e Paint he 3e Dyes and dyeinge 46 Household	summary journal accoun herbal drugs paint painting s ousehold supplies garden see le drugstorestenness tennesseee dyes dyeinge hou	nt documents sale chemical supplies dyestuffs personal eds individuals businessese seee paint shopsequipment
	³ William B Abbott, Williar	m B. Papers, 1862-1864 Frederic m B. Papers, 1862-1864, Fre- Ger	See South Carolina. Abbeville District k Co., Va. Section A 5-16-57 GUIDE 10 items ick Co., Va. 10 items. Sketch These are		abbott william frederick va
In [5]:	4 William B proper	reral documents concerned with ty by C. S. A. troops in 1862, and Abbott at various tir	well-to-do farmer of Frederick Co., Va. the evaluation of damage done to his .receipts in 1864 for hay bought from nes in Aug., 1864 by the C. S. A. Army. rt CountVectorizer, TfidfVect	welltodo farmer fred concerned evaluation dan andreceipts hay bought	erick va several documents nage done property troops abbott various times army
	<pre>#get the text column docs=df.Clean.tolist #create a vocabulary</pre>	n c() y of words, appear in 85% of docum			
	<pre>cv=CountVectorizer(n word_count_vector=cv tfidf_transformer=Tf</pre>	<pre>nax_df=0.85,max_feature .fit_transform(docs)</pre>	es=10000) _idf =True, use_idf =True)		
Out[5]: In [6]:	<pre>TfidfTransformer() docs_test=df['Clean' feature_names=cv.get</pre>				
In [9]:	<pre>return sorted(tu def extract_topn_from</pre>	o_matrix.col, coo_matr uples, key =lambda x: (:	<pre>x[1], x[0]), reverse=True) s, sorted_items, topn=10):</pre>		
	<pre>sorted_items = s score_vals = [] feature_vals = </pre>				
	<pre>for idx, score i #keep track score_vals.a feature_vals</pre>	of feature name and i append(round(score, 3) s.append(feature_names	ts corresponding score		
	<pre>#results = zip() results= {} for idx in range</pre>	s of feature,score feature_vals,score_val e(len(feature_vals)): cure_vals[idx]]=score_			
	<pre>def get_keywords(idx #generate tf-idn tf_idf_vector=tf #sort the tf-idn</pre>	f for the given docume fidf_transformer.trans f vectors by descendin	<pre>form(cv.transform([docs[idx]] g order of scores</pre>]))	
	sorted_items=son #extract only th keywords=extract	<pre>ht_coo(tf_idf_vector.to he top n; n here is 10 t_topn_from_vector(feat words.keys())[:5]</pre>			
	<pre>#generate tf-idi tf_idf_vector=tf #sort the tf-idi sorted_items=sor</pre>	f for the given docume fidf_transformer.trans f vectors by descendin rt_coo(tf_idf_vector.te	<pre>form(cv.transform([row])) g order of scores pcoo())</pre>		
In [22]:	<pre>keywords=extract return list(keyw li = df.Clean.apply d = pd.read_csv("all </pre>	<pre>words.keys()) (get_top_words) L_sorted_collection.cs</pre>	ture_names,sorted_items,10)		
	<pre>d["Key_word"] = np.r ind = d[~d.Text.isnu d.loc[ind,"Key_word" d = d.iloc[:,[0,1,2,</pre>	nan 111()].index			
In [23]:	d.head(20)				

Out[23]:	Collection_Head A. B. Davis and Company	Name A. B. Davis and	Text A. B. Davis and Company (Philadelphia, Pa.) See Davis	Key_word [davis, company, philadelphia, pa]	Year	Loc Pa.	Start NaN	End NaN	Page_drawer Drawe	er_No 157	https://archive.or
		Company	(A. B.) and Company AeHe Roscoe (Firm: Nashville, Tenne)e Journal, 1853, Septe-1857, Dece 1 item(800 ppe)e Wholesale and retail druggist and dealer in paints, oils, and dyestuffse Summary: Journal (account book) documents the sale of chemical and herbal drugs, paint and painting supplies, dyestuffs, personal and household supplies, and	[paint, household,	1853,						
	A.H. Roscoe	A.H. Roscoe	garden seeds to individuals and businessese le Drugstores Tennesseee 2e Paint shops Equipment and supplies Tennesseee 3e Dyes and dyeinge 46 Household supplies Tennesseee 5c Herbs Therapeutic usee 6+ Seed industry and trade Tennessees 7e Nashville ite ® nne) Commercee 20 MAY QO1 23804905 NDHYme	nashville, supplies, journal, dyes, roscoe, seed, qo, seeds]	Septe- 1857	Nashville, Tenne	1853.0	1857.0	1	157	https://archive.oi
	Abbeville District	Abbeville District	Abbeville District (8.C.) See South Carolina. Abbeville District Abbott, William	[abbeville, district, south, carolina]	NaN	Abbeville District, South Carolina	NaN	NaN	2	157	https://archive.or
	Abbott, William B	Abbott, William B	B. Papers, 1862- 1864 Frederick Co., Va. Section A 5-16-57 GUIDE 10 items Abbott, William B. Papers, 1862-	[abbott, frederick, va, william]	1862- 1864	Abbott, Va.	1862.0	1864.0	3	157	https://archive.or
	NaN	Abbott, William B	1864, Fre- Gerick Co., Va. 10 items. Sketch These are the papers of William B. Abbott, evidently a well- to-do farmer of Frederick Co., Va. There are several documents concerned with the evaluation of damage done to his property by C. S. A. troops in 1862, and.receipts in 1864 for hay bought from	[abbott, evaluation, damage, fre, hay, va, bought, done, farmer, william]	NaN	NaN	NaN	NaN	4	157	https://archive.oi
			Abbott at various times in Aug., 1864 by the C. S. A. Army. Abbott & Company Papers, 1856- 1871	[abbott,							
	Abbott & Company	Abbott & Company	Philadelphia, Pennsylvania a Section A 66 items OCT 10 "49 Abderhalden, Emil Papers, 1919 Halle, Germany Josiah	pennsylvania, philadelphia, company] [emil, germany,	1856- 1871	Philadelphia, Pennsylvania Abderhalden,	1856.0	1871.0	5	157	https://archive.or
	Abbott & Company	Abbott & Company	C. Trent Collection in the History of Medicine, Div, 4-8-60 Abbott & Company. Papers 1856-	div, medicine, trent, josiah, history]	1919	Halle, Germany	1919.0	NaN	6	157	https://archive.oi
	Abbott & Company	Abbott & Company	1871 Philadelphia, Pennsylvania, € Miscellaneous letters concerning scales sold by Abbott & Company.	[abbott, company, scales, sold, pennsylvania, philadelphia, miscellaneous, concerning]	1856- 1871	Philadelphia, Pennsylvania	1856.0	1871.0	7	157	https://archive.or
	Abel, Ernest L.	Abel, Ernest L.	Abel, Ernest L. Papers, 1922- 1952 West Palm Beach, Palm Beach Co., Fla. Saat; 550 items & 8 vols. Box 1 of Labor Archives	[palm, beach, abel, ernest, fla, box, archives, labor, west, vols]	1922- 1952	West Palm Beach, Palm Beach Co., Fla.	1922.0	1952.0	8	157	https://archive.or
	NaN	Abel, Ernest L.	Abel, Ernest L. Papers. West Palm Beach, Palm Beach Co., Fla' : Ernest L. Abel was involved in postal union activity at least from the mid- 1920s to the early 1950s. He served from 1925-1926 as Secretary- Treasurer for Local Union No. 749 of the National Federation of Post Office Clerks (AFL) and was State Representative for the National Federation at the same time. In 1927, Abel became the first president of the Florida Federation of Post Office Clerks, His union service included being President Union No. 749 in 1945, and State Legislative Representative of the' ® Florida	[abel, federation, palm, union, beach, clerks, ernest, representative, florida, post]	NaN	NaN	NaN	NaN	9	157	https://archive.oi
	NaN	Abel, Ernest L.	Abel, Ernest L. Federation at least in 1952. The correspondence covers the years 1925-1928, with a small amount of material for 1945 to 1952. Abel's letters reveal a variety of topics, such as charters and organizing efforts, finances and the per capita tax, disaster relief for Post Office employees by the Red Cross after a hurricane, conventions, and legislation. The later correspondence includes a communication in 1945 from Leo George, President of the National Federa- tion of Post Office Clerks, about an editorial in the local union paper. In 1952, there are letters from Congressmen Spessard Holland and George	[post, office, hurricane, disaster, leo, communication, abel, congressmen, conventions, clerks]	NaN	NaN	NaN	NaN	10	157	https://archive.org
	NaN	Abel, Ernest L.	Abel, Ernest L. pending legislation. In the printed material are programs for various Florida postal organizations' conventions, spanning the twenty-year period, 1927 to 1947. Included are the Florida State Convention of the National Association of Letter Carriers and National Federation of Post Office Clerks, the Florida Postal Groups, the Joint Convention of Florida Postal Organizations, the Florida Federation of Post Office Clerks, and the Florida State Convention of the National Federation of Post Office Clerks, and the Florida State Convention of the National Federation of post Office Clerks, and the Florida State Convention of the National Federation of post Office Clerks. Since this collection has not been cataloged in detail, the entries made for the Autograph File may be	[florida, clerks, postal, federation, convention, post, office, national, organizations, pending]	NaN	NaN	NaN	NaN	11	157	https://archive.orc
	NaN	Abel, Ernest L.	Abel, Ernest L. incomplete. Correspondence usually was indexed when there was a whole folder of correspondence for a person or a significant number of letters for a person through- out the collection. Only general' subject entries have been made.	[person, abel, indexed, correspondence, ernest, incomplete, significant, whole, usually, folder]	NaN	NaN	NaN	NaN	12	157	https://archive.orc
	Abercrombie, John	Abercrombie, John	John Papers, 1839 Edenburgh, Scotland Josiah C, Trent Collection in the History of	[scotland, div, medicine, trent, josiah, history, john]	1839	Scotland	1839.0	NaN	13	157	https://archive.orc
	Abercrombie, John Joseph	Abercrombie, John Joseph	Medicine, Div. 4-8-60 5 items Abercrombie, John Josephe Papers, 1771- 1907. 131 items. Ue Se brigadier general who served in Second Seminole War and Mexican Ware Family papers date from 1839 to 1879-6 The earliest one is an unsigned letter by Abercrombie in Flae in diary form relating an expedition against the Indianse Early correspondence pertains to Abercrombie's career in the militarye The Civil War correspondence is mainly to and from family members, but a few are from other officers about military matterse One wartime letter concerns politics in Pae in 1863 and Lincolne Much of the correspondence is to and rro@- "Sallie" Patterson, Abercrombie's sister and wife of 29 NOV 94	[correspondence, josephe, seminole, matterse, flae, wartime, pae, unsigned, sallie, pertains]	1771- 1907	NaN	1771.0	1907.0		157	https://archive.org
	NaN	Abercrombie, John Joseph	Abercrombie, John Josephe Papers, cece (Card 2) General Robert Pattersone Abercrombie was brevetted brigadier general at the end of the war and retired from military service on June 12, 1865-4 Unprocessed collectione Cataloged from accession recorde PJ le Lincoln, Abraham, 1809- 1865. 2. Patterson, Salliee 3e United States HistoryCivil War, 1861-1865. 4e Indians of North America Government relationse 5e Pennsylvania Politics and government Civil War, 1861- 1865-e 6e« SoldiersUnited States ia hiaee ncee 29 NOV 94	[war, josephe, states, relationse, historycivil, general, ia, retired, brigadier, pj]	NaN	NaN	NaN	NaN	15	157	https://archive.orc
	Abercrombie, Lascelles	Abercrombie, Lascelles	Abercrombie, Lascelles Pavers, Nn. d. Pp. I Section A oct 10 "49 GUIDE 1	[nn, pavers]	n. d.	NaN	NaN	NaN	16	157	https://archive.orc
			item Abercrombie, Lascelles. Papers, n. d. l item. Sketch Letter to "Ivy".from Lascelles								
	NaN	Abercrombie, Lascelles	Abercrombie (1881-1938) Eng. poet and critic, born in Cheshire. Prof. of Eng. Lit. Leeds (1922-29) Unive of London (1929- 1935), Author of Gnter- ludes:& Poems (1908), Emblems of Love (1918) etc, Letter concerns injuries Abercrombie re- ceived in an accident and it was taken from Apercrombie's Emblems of Love in D. U. L,	[love, eng, unive, accident, leeds, ceived, lit, critic, cheshire, prof]	NaN	NaN	NaN	NaN	17	157	https://archive.orc
	Abercromby, James, Baron Dunfermline	Abercromby, James, Baron Dunfermline	James, First Baron Dunfermline Papers, 1840- 1851 "Colinton House," Midlothian Co.,	[abercromby, scotland, baron, house, first, james]	1840- 1851	Scotland	1840.0	1851.0	18	157	https://archive.orc
In [24]:	NaN d.to_csv("al	Abercromby, James, Baron Dunfermline	Scotland XVIII- =z 20 items 10- 28-58 Abercromby, James, Baron Dunfermline. Papers, 1840- 1851. Midlothian Co., Scotland. 20 items Sketch. James Abercroihy, Baron Dunfermline (1776-1858), first entered Parliament in 1807, and in 1835 was elected Speaker of the House of Commons with the assistance of Lord Russell. Upon his retirement in 1839 he was created a baron. There was an earlier set of Harls Dunfermline which died out in the mid- seven= teenth century. The letters in this collection are address- ed to James Loch, also_a Scottish Whig and	[baron, james, abercromby,	NaN	NaN	NaN	NaN	19	157	https://archive.org
1:	co_csv("al	_Jurted_ke	tu.CSV")								