AutoLibrary

Jiayi Fan, Yichun Ren & Bingqi Zhou

01 MOTIVATIONS

02 METHODOLOGY

03 EXPERIMENTS

04 RESULTS ANALYSIS

05 CONCLUSION

HOW TO FIND RELATED PAPERS

Recommendations on Digital Libraries

+

Search Engines

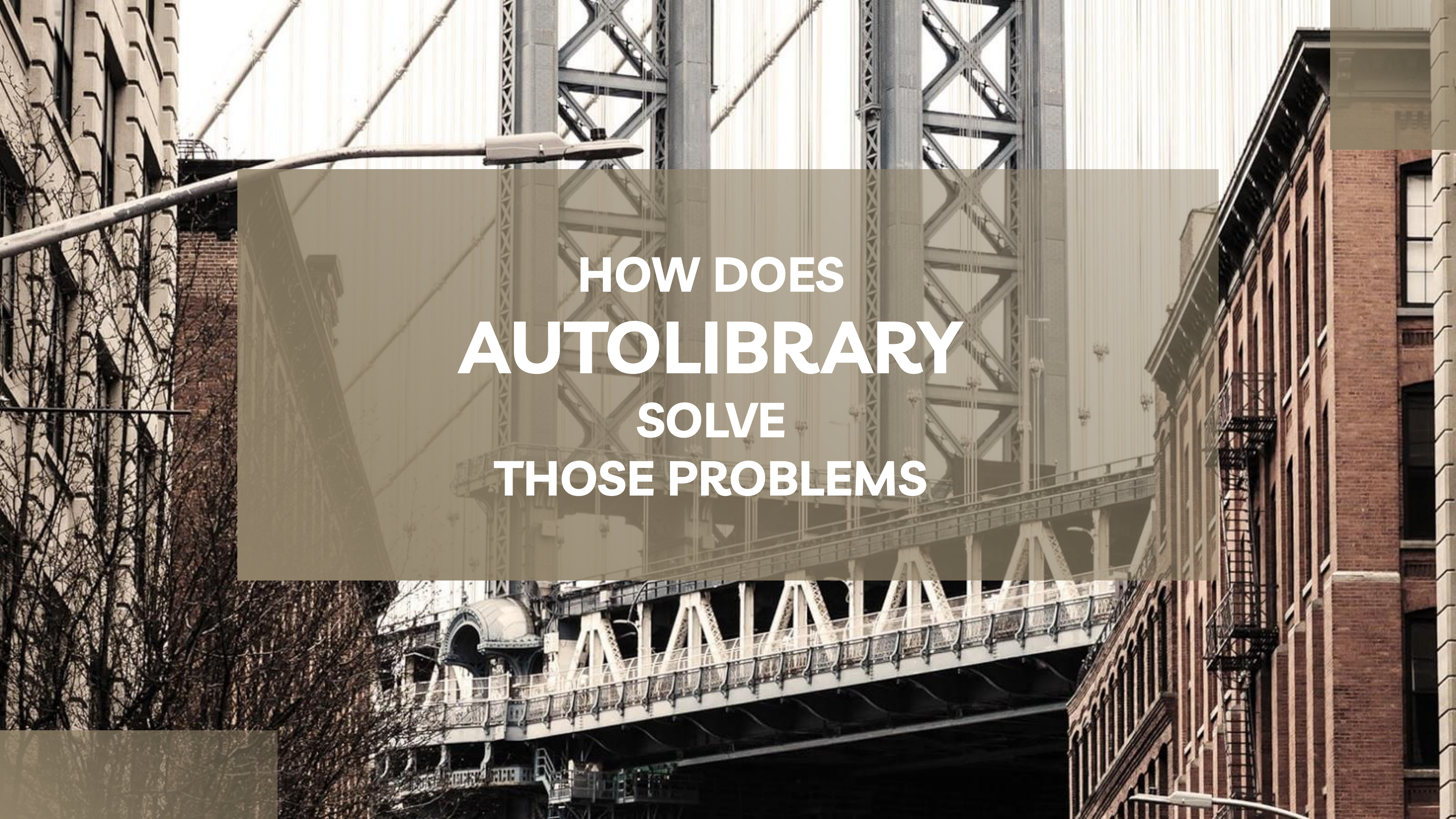
WHAT ARE THE PROBLEMS

POOR RECOMMENDER SYSTEM

LIMITED PAPER DATASETS

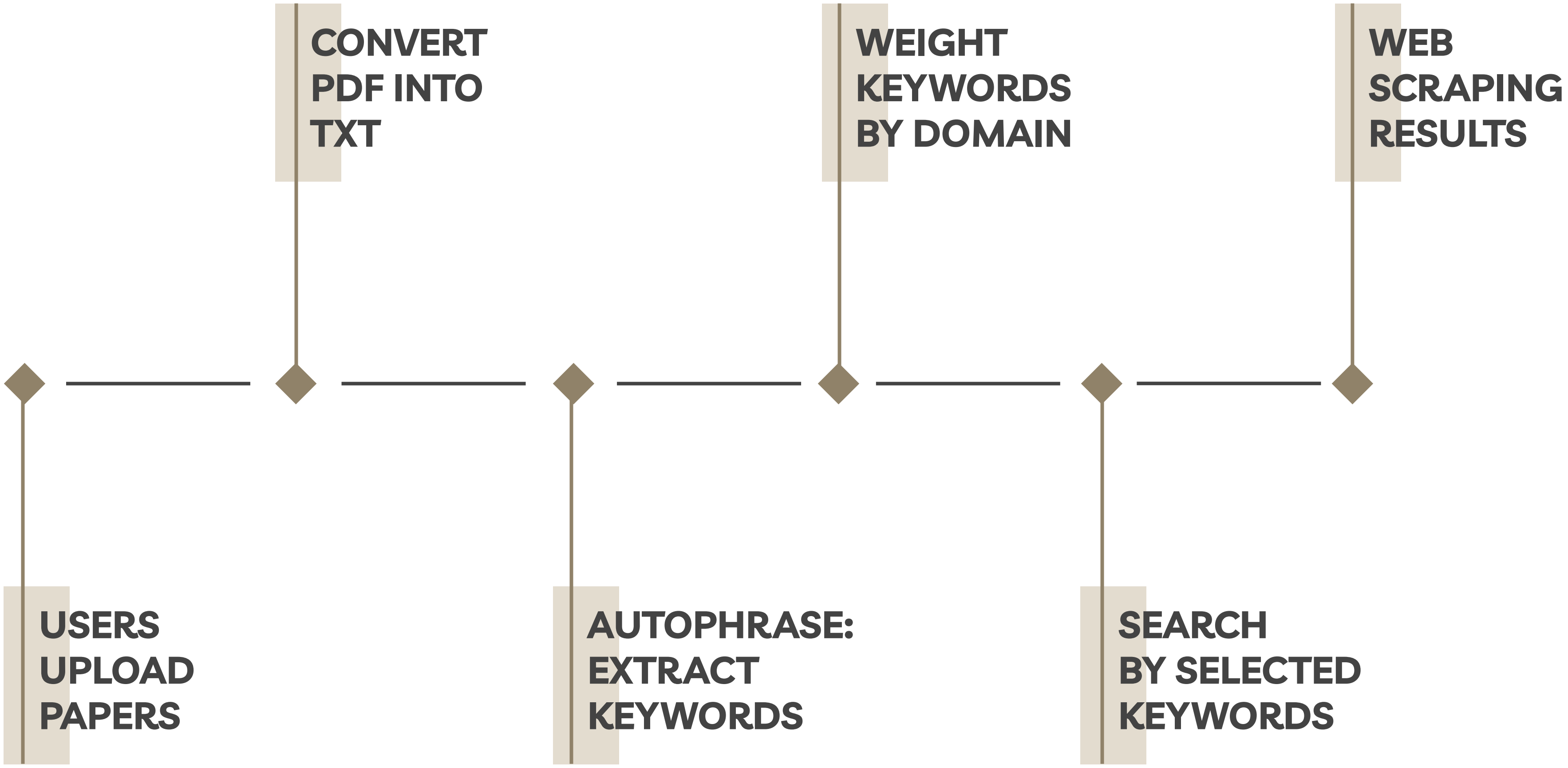
INCORRECT SEARCH KEYWORDS

UNFAMILIAR SCIENTIFIC DOMAINS



HOW DOES AUTOLIBRARY SOLVE THOSE PROBLEMS

WORKFLOW



USERS UPLOAD PAPERS

CONVERT PDF INTO TXT

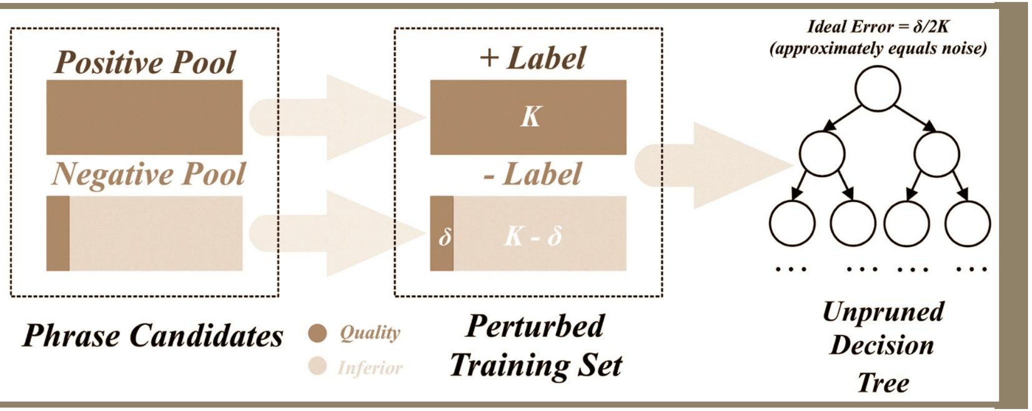
AUTOPHRASE: EXTRACT KEYWORDS

WEIGHT KEYWORDS BY DOMAIN

SEARCH BY SELECTED KEYWORDS

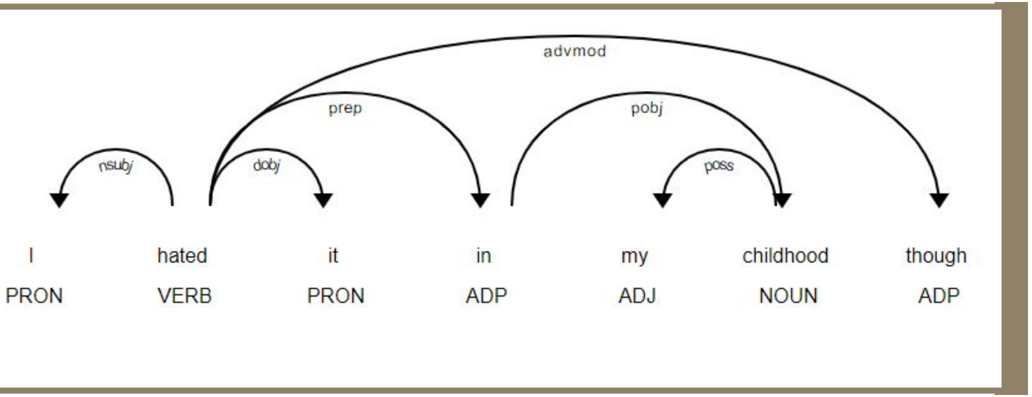
WEB SCRAPING RESULTS

AUTOPHRASE



TECHNIQUE 1

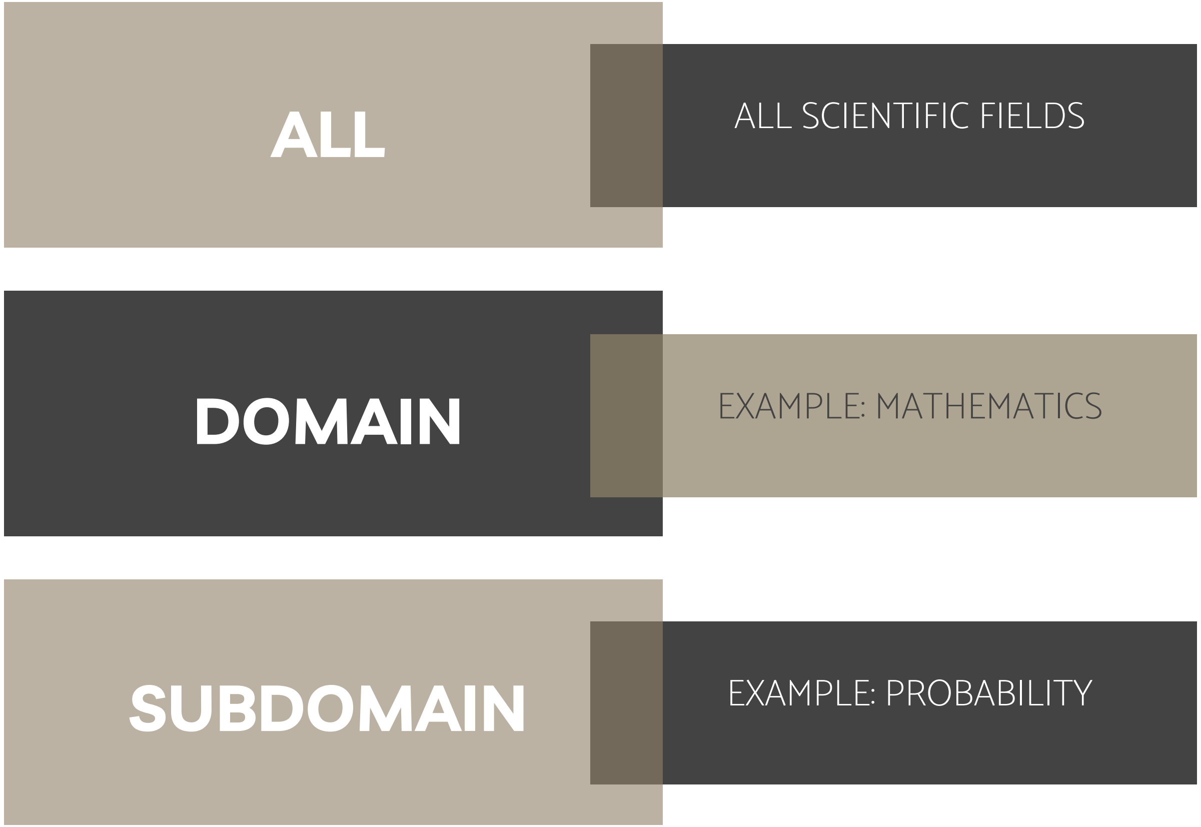
Robust Positive-Only Distant Training



TECHNIQUE 2

POS-Guided Phrasal Segmentation

PRETRAINED DOMAIN DATASETS



ALL

ALL SCIENTIFIC FIELDS

DOMAIN

EXAMPLE: MATHEMATICS

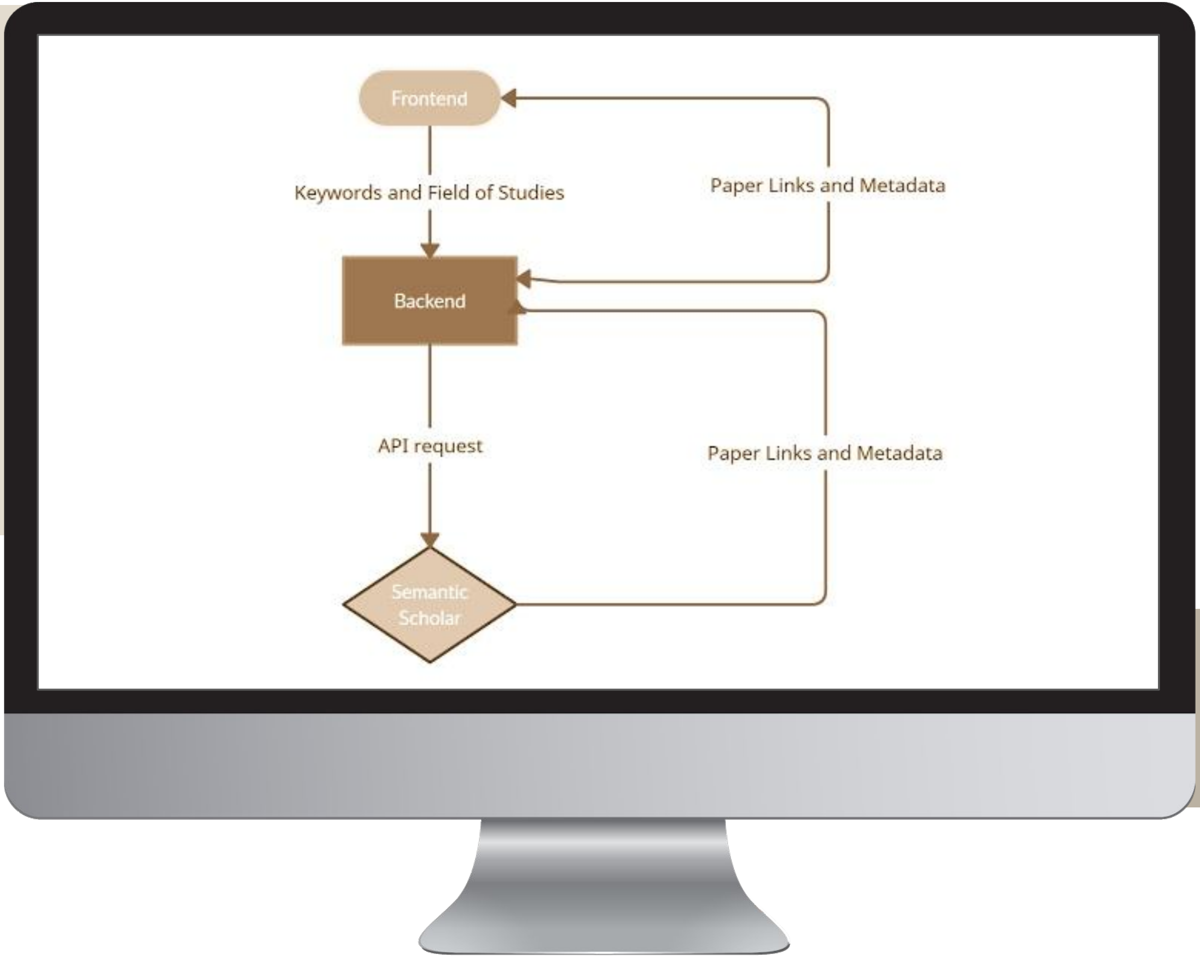
SUBDOMAIN

EXAMPLE: PROBABILITY

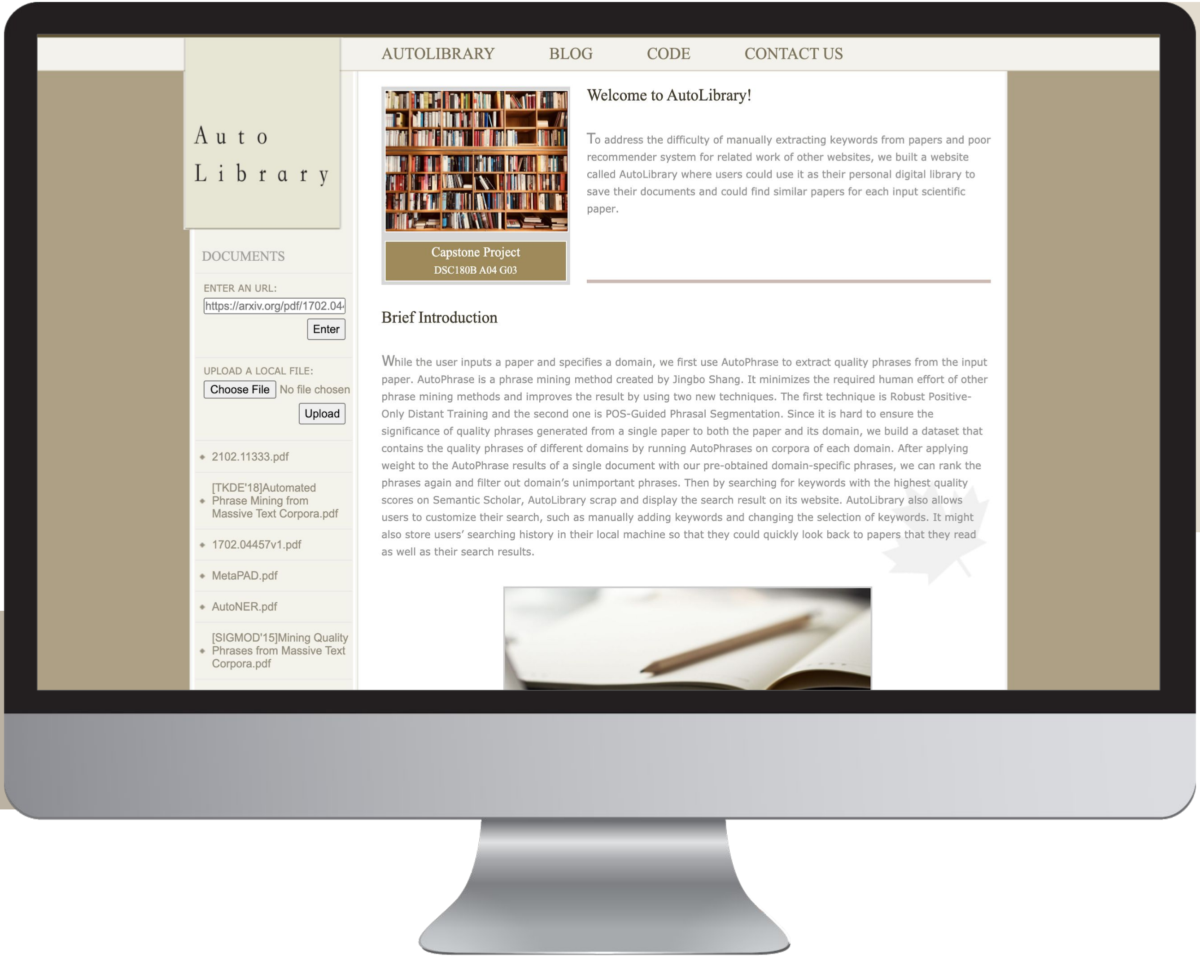
APPLY WEIGHT

Weighted Quality = Quality * Domain Quality

WEB SCRAPING



WEBSITE OUTLOOK



Welcome to AutoLibrary!

To address the difficulty of manually extracting keywords from papers and poor recommender system for related work of other websites, we built a website called AutoLibrary where users could use it as their personal digital library to save their documents and could find similar papers for each input scientific paper.

Brief Introduction

While the user inputs a paper and specifies a domain, we first use AutoPhrase to extract quality phrases from the input paper. AutoPhrase is a phrase mining method created by Jingbo Shang. It minimizes the required human effort of other phrase mining methods and improves the result by using two new techniques. The first technique is Robust Positive-Only Distant Training and the second one is POS-Guided Phrasal Segmentation. Since it is hard to ensure the significance of quality phrases generated from a single paper to both the paper and its domain, we build a dataset that contains the quality phrases of different domains by running AutoPhrases on corpora of each domain. After applying weight to the AutoPhrase results of a single document with our pre-obtained domain-specific phrases, we can rank the phrases again and filter out domain's unimportant phrases. Then by searching for keywords with the highest quality scores on Semantic Scholar, AutoLibrary scrap and display the search result on its website. AutoLibrary also allows users to customize their search, such as manually adding keywords and changing the selection of keywords. It might also store users' searching history in their local machine so that they could quickly look back to papers that they read as well as their search results.

EXPERIMENTS

Compare & Contrast
with
Similar Web Applications

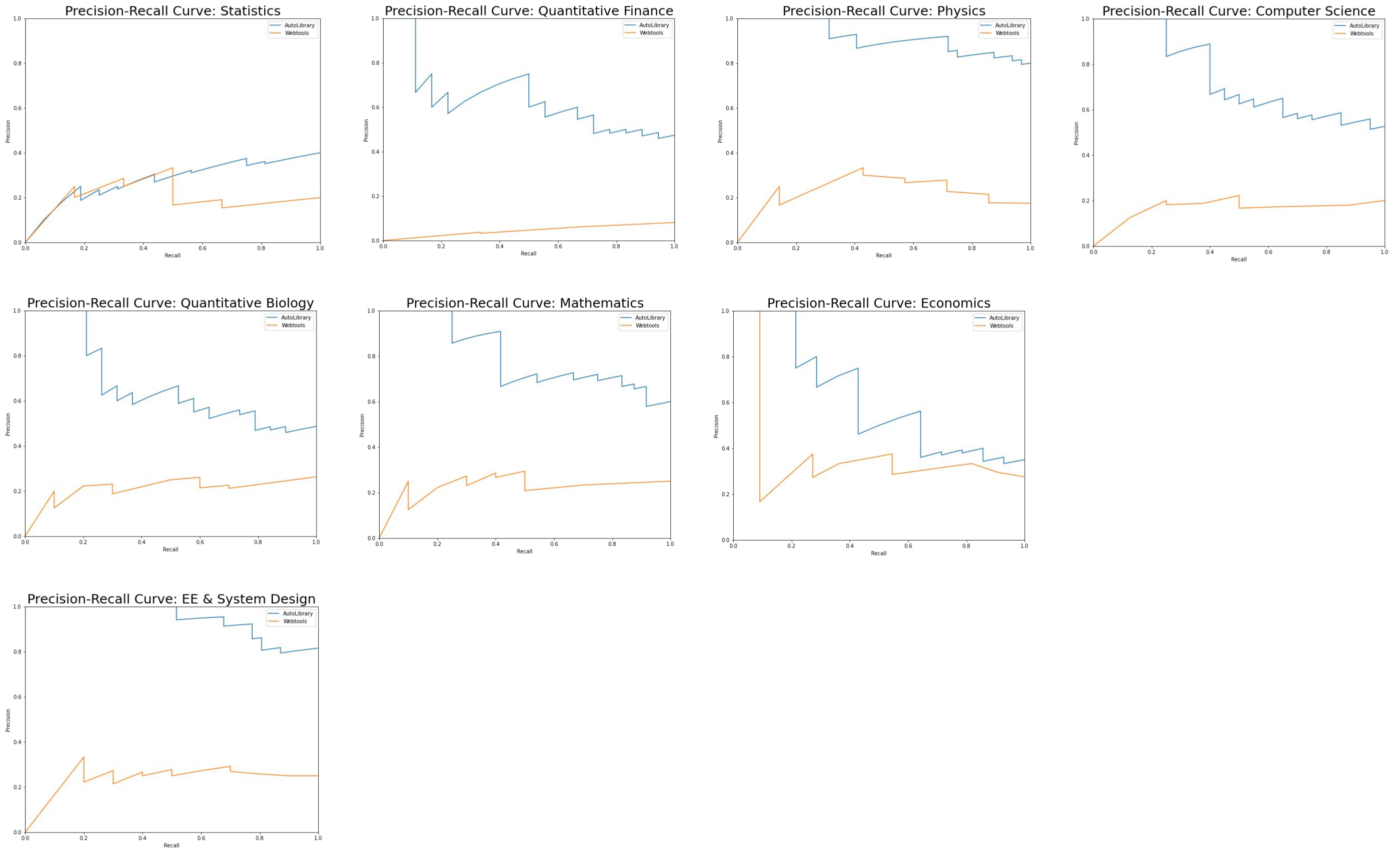
AutoLibrary

Jstor

Webtools

MonkeyLearn

AutoLibrary vs. Webtools



Precision-Recall Curve: Statistics

Precision-Recall Curve: Quantitative Finance

Precision-Recall Curve: Physics

Precision-Recall Curve: Computer Science

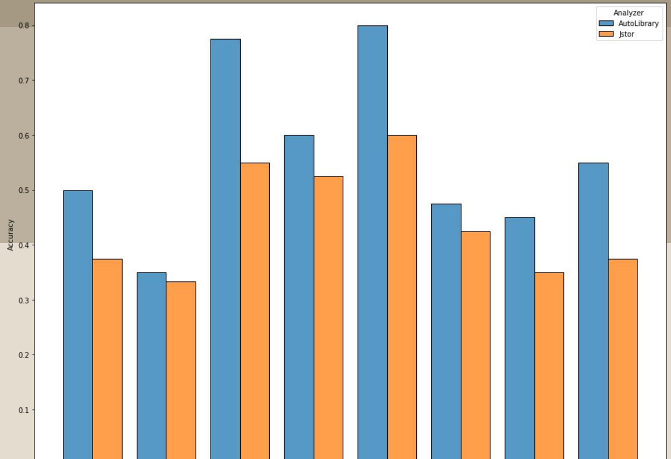
Precision-Recall Curve: Quantitative Biology

Precision-Recall Curve: Mathematics

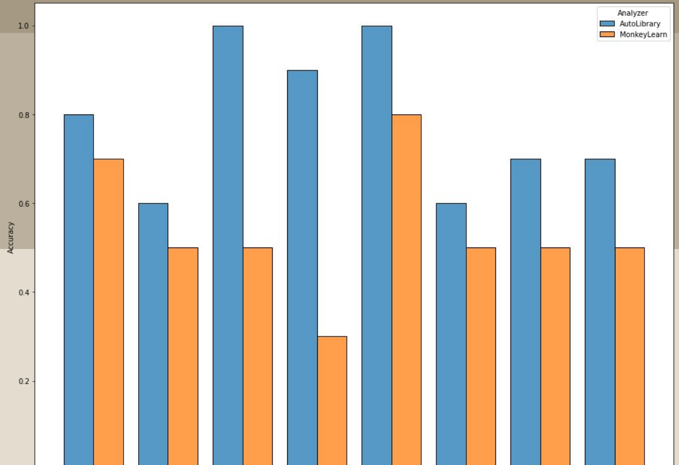
Precision-Recall Curve: Economics

Precision-Recall Curve: EE & System Design

	Computer Science	Economics	Electrical Engineering	Math	Physics	Quantitative Biology	Quantitative Finance	Statistics
Auto Library	50	35	78	60	80	48	45	55
Jstor	38	33	55	53	60	43	35	38



	Computer Science	Economics	Electrical Engineering	Math	Physics	Quantitative Biology	Quantitative Finance	Statistics
Auto Library	80	60	100	90	100	60	70	70
Monkey Learn	70	50	50	30	80	50	50	50



5 Papers Published by Professor Shang

Article	Publish Year	Domain
CrossWeigh	2019	Computer Science
AutoPhrase	2018	Computer Science
LM-LSTM-CRF	2018	Computer Science
AutoNER	2018	Computer Science
SetExpan	2017	Computer Science

RESULT ANALYSIS

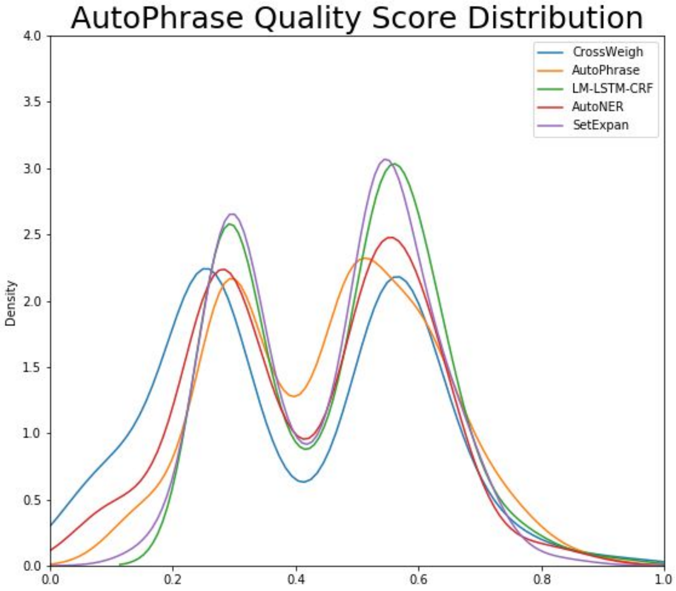
Papers With:

Overlapped Topics
+
Different Topics

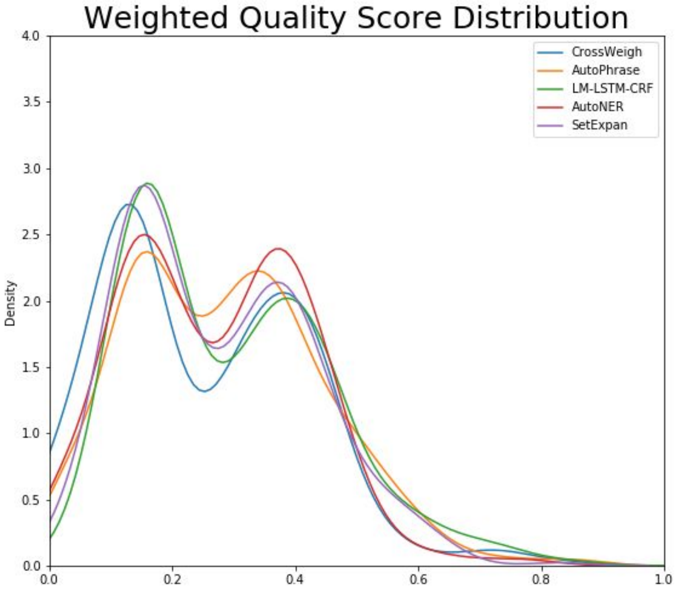
Top 10 Quality Phrases from 5 Papers

Rank	CrossWeigh	AutoPhrase	LM-LSTM-CRF	AutoNER	SetExpan
1	natural language processing	knowledge base	neural networks	natural language	bipartite graph
2	natural language	information extraction	pos tagging	domain specific	skip gram
3	computational linguistics	domain specific	bi lstm	named entity	ranked lists
4	cross validation	text corpora	sequence labeling	distant supervision	semantic drift
5	named entity recognition	keyphrase extraction	word embedding	lstm crf	text corpora
6	pos tagging	pos tagger	transfer learning	ablation experiments	texas
7	lstm crf	natural language	language model	distantly supervised	coarse grained
8	chicago	massive text corpora	word embeddings	ncbi	california
9	japan	cn	lstm crf	ner	skip grams
10	f1	auc	conditional random	ram	ranked list

DISTRIBUTION CHANGE



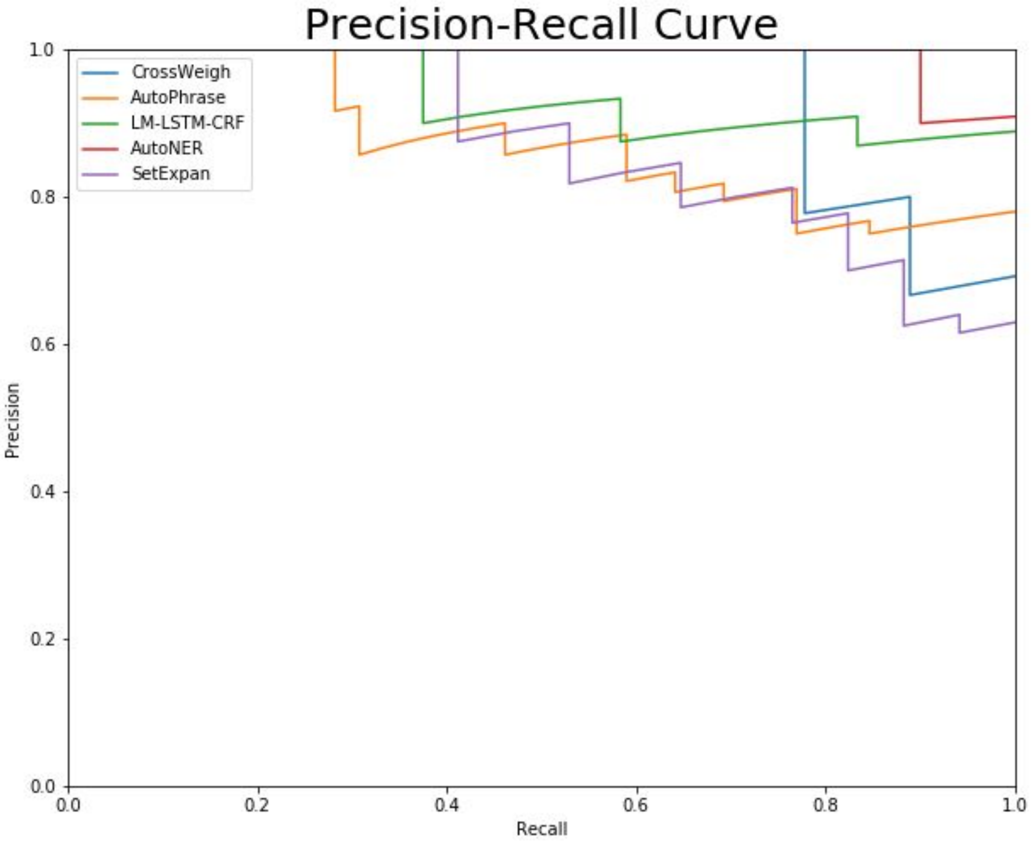
AutoPhrase Quality Score Distribution



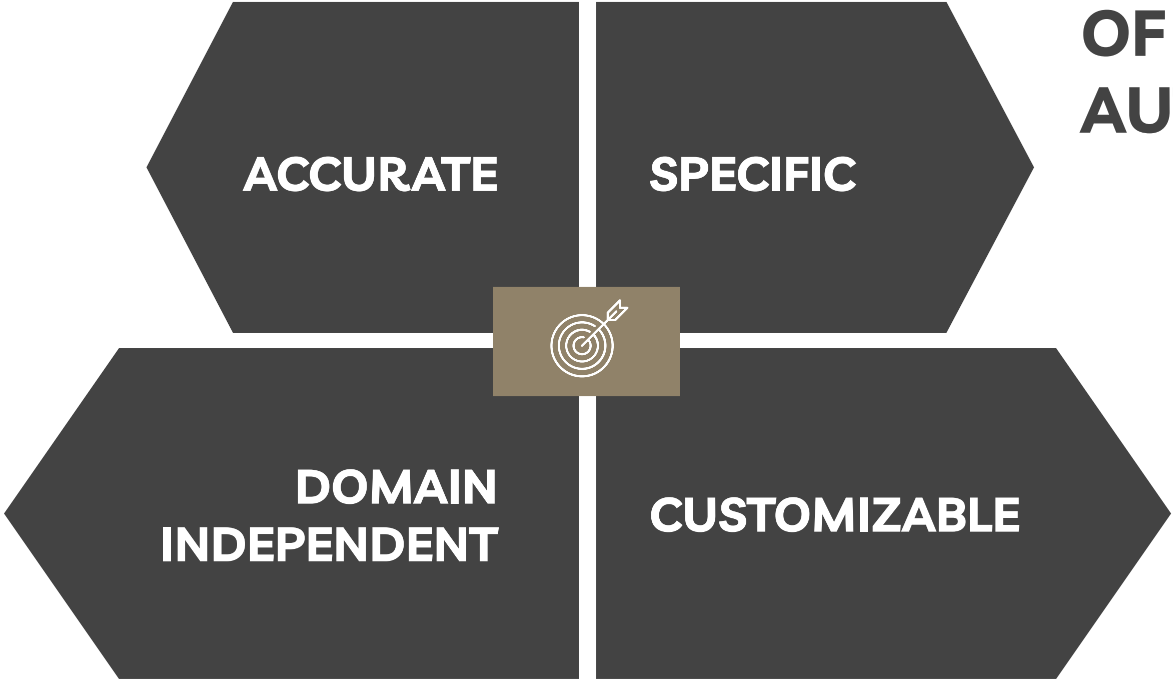
Weighted Quality Score Distribution

Accuracy Compared to Manual Labeling

Article	Accuracy		
	Quality Score > 0.5	Quality Score > 0.6	Quality Score > 0.7
CrossWeigh	0.6429	1.0	1.0
AutoPhrase	0.7800	0.8571	1.0
LM-LSTM-CRF	0.8889	0.9231	1.0
AutoNER	0.8333	1.0	1.0
SetExpan	0.6296	0.8889	1.0


Precision-Recall Curve
CrossWeigh
AutoPhrase
LM-LSTM-CRF
AutoNER
SetExpan
Precision
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Recall

ADVANTAGES OF AUTOLIBRARY



ACCURATE

SPECIFIC

DOMAIN INDEPENDENT

CUSTOMIZABLE

FUTURE WORKS

1. IMPROVE RUNTIME
2. ADD MORE DOMAINS
3. HAVE USER MANAGEMENT SYSTEM

THANKS

Feel free to use our projects:
https://yichunren.pythonanywhere.com/autolibrary