CS60092 Information Retrieval TERM Project

SCIATICA

Group 4

Ashwani Kumar Kamal - 20CS10011
Hardik Pravin Soni - 20CS30023
Sourabh Soumyakanta Das - 20CS30051
Shiladitya De - 20CS30061

TABLE OF CONTENTS

PROBLEM STATEMENT: QUERY BY EXAMPLE

RETRIEVAL OF SIMILAR PAPERS

User Input: Research paper title and Facet (Background / Method / Result)

Results: Ranked list of most similar papers from a research paper corpus

DATASET USED

CSFCube Dataset 2021
Repository: Github
Paper: Paper

OBJECTIVES

LITERATURE REVIEW

Literature review require spending hours on the internet finding similar papers

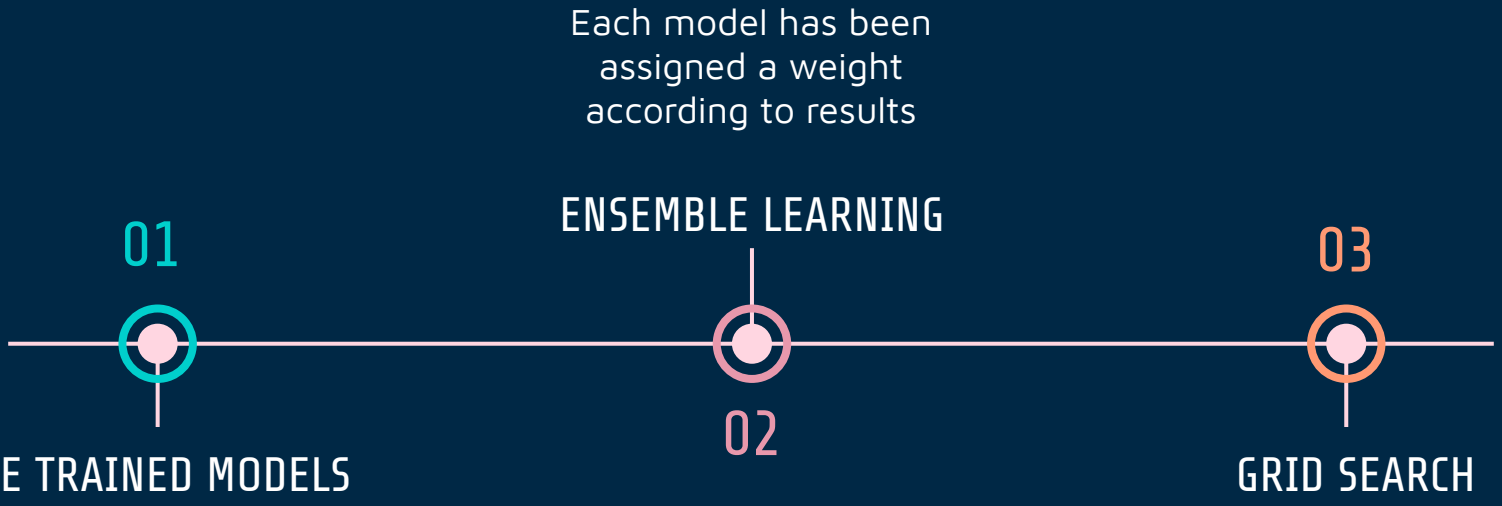
NOVELTY

Check if there exist papers which have used a particular method before

COMPARING RESULTS

comparing results with other papers which worked on same/similar problem statement

MOTIVATION


01
PRE TRAINED MODELS
Few models having better results could sometimes overfit
02
ENSEMBLE LEARNING
Each model has been assigned a weight according to results
03
GRID SEARCH
Hyperparameter tuning through grid search to obtain better results

CSFCube Dataset

Main Features

- 800,000 Computer Science Research Papers
- Sourced Out of 81.1M papers from S2ORC Dataset.
- Used for Training, Testing, Validating and Demonstration of ***SCIATICA***.

Salient Features

- A Human Annotated Relevance Score between 0-3 provided for each query-pair document.
- Each Query Pool comprises of 100-200 Research Papers.
- Currently contains contains only abstract and title of the paper.
- Contains 16 Background Queries, 17 Method Queries and 17 Result Queries for Validation.
- A Total of 6,244 Query-Candidate Pairs

EXPERIMENTS

BASE MODELS

Generated encodings from pretrained sentence transformer models

GRID SEARCH

Standard grid search technique is applied to tune the hyper parameters of DistilBertModel of AutoTransformers.

ENSEMBLE LEARNING

Results of all the models are combined using weights according to the quality of results produced by each model

RESULTS

	Background					*Method*				
	RP	**P@20**	**R@20**	**$NDCG_{\%100}$**	**$NDCG_{\%20}$**	**RP**	**P@20**	**R@20**	**$NDCG_{\%100}$**	**$NDCG_{\%20}$**
BERT_NLI	0.2004	0.2750	0.4328	0.7735	0.5781	0.1656	0.1028	0.3265	0.6056	0.3393
BERT_PP	0.2332	0.3109	0.5024	0.7760	0.5974	0.1826	0.0998	0.3388	0.6350	0.3865
SPECTER	0.2353	0.3125	0.4936	0.7994	0.6407	0.1843	0.1097	0.4107	0.6269	0.3744
DISTILBERT (pretrained)	0.3249	0.3781	0.6224	0.8544	0.7264	0.1416	0.1490	0.4753	0.6731	0.4518
ALL_MPNET	0.2797	0.3469	0.5750	0.8536	0.7166	0.2005	0.1641	0.4734	0.6633	0.4544
TF-IDF	0.1777	0.2266	0.3789	0.7262	0.4795	0.0892	0.0748	0.2434	0.5439	0.2440
ALBERT	0.2510	0.2828	0.4119	0.7809	0.5951	0.1285	0.1045	0.3568	0.5994	0.3346
ENSEMBLED MODEL	0.2959	0.3594	0.5829	0.8583	0.7187	0.1981	0.1630	0.4940	0.6769	0.4656

RESULTS

	Result					*Aggregated*				
	RP	**P@20**	**R@20**	**$NDCG_{\%100}$**	**$NDCG_{\%20}$**	**RP**	**P@20**	**R@20**	**$NDCG_{\%100}$**	**$NDCG_{\%20}$**
BERT_NLI	0.1278	0.1826	0.4023	0.6538	0.4072	0.1643	0.1859	0.3866	0.6768	0.4404
BERT_PP	0.1548	0.2273	0.5484	0.7048	0.5183	0.1898	0.2119	0.4631	0.7043	0.4995
SPECTER	0.1904	0.2856	0.6814	0.7649	0.6022	0.2030	0.2353	0.5286	0.7296	0.5379
DISTILBERT (pretrained)	0.2357	0.2818	0.6258	0.7817	0.6246	0.2336	0.2688	0.5745	0.7688	0.5996
ALL_MPNET	0.2498	0.3113	0.7276	0.7899	0.6497	0.2429	0.2733	0.5919	0.7680	0.6055
TF-IDF	0.1083	0.1333	0.3067	0.6425	0.3851	0.1247	0.1437	0.3084	0.6361	0.3676
ALBERT	0.1603	0.2109	0.4837	0.6909	0.4804	0.1795	0.1986	0.4174	0.6895	0.4687
ENSEMBLED MODEL	0.2564	0.3085	0.7178	0.7916	0.6610	0.2496	0.2762	0.5984	0.7748	0.6139

FUTURE WORKS

- Applying Semantic Similarity and Intent.
- Improving results through user feedback.
- Scraping the H-index to provide an option to rank the data based on the H-index of the authors.
- Expanding to full S2ORC corpus.
- Implementing multiprocessing to improve the query search time by putting different models into different cores.

Any questions?

THANKS