NDLI - UNESCO
INTERNATIONAL SYMPOSIUM
ON KNOWLEDGE ENGINEERING
FOR DIGITAL LIBRARY DESIGN 2019
IN ASSOCIATION WITH IIT DELHI

SMART AND OPEN TECHNOLOGIES
FOR DIGITAL LIBRARY

Citta : A Lite Semantic Recommendation Framework for Digital Libraries

Rajaswa P.[1] and Siddhant M.[2]

Birla Institute of Technology & Science Pilani {1} ; Vellore Institute of Technology {2}

BITS Pilani
K K Birla Goa Campus

VIT

ABSTRACT

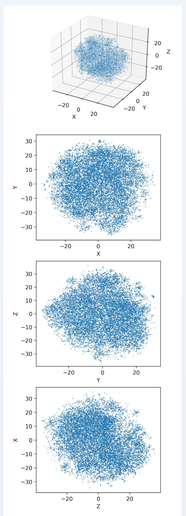
In this work we propose the idea of a content-sentiment aware semantic search based reading recommendation framework : **Citta** (named after the Sanskrit term for consciousness : '*Chitta'*). The framework solely works on pure text-content data and makes no use of any metadata, user-reading history or user-interactions. We analyse the performance of our proposed framework with respect to processing time and precision with respect to semantic relevance.

INTRODUCTION

1. Recommending new books without using metadata, user history or user interactions
2. Recommending semantically related books instead of keyword based search
3. Using Deep Learning based Language Models to provide accurate reading suggestions

- Recent advances in large Transformer based Deep Language Models have achieved state-of-the-art results for various NLP tasks
- Pre-trained language models (BERT : 768 dimensions) can be used to extract high-dimensional contextual representations of textual data for semantic search related tasks
- Practically infeasible to use pre-trained language models over large databases due to computational cost and query-time involved
- Filtering the database by ranking the documents based on a predefined content-sentiment aware heuristic

DATASET



t-SNE Plots of the Book Embedding Space

- *CMU Book Summary Dataset*
- *Contains **plot summaries** for **16,559** books extracted from **Wikipedia***
- *Aligned metadata from **Freebase**, including book **author**, **title**, and **genre***

Original Dataset (2013)

WIKIPEDIA ID	1166383
FREEBASE ID	/m/04cvx9
BOOK TITLE	White Noise
BOOK AUTHOR	Don DeLillo
PUBLICATION DATE	1985-01-21
GENRES	Novel, Postmodernism, Speculative fiction, Fiction

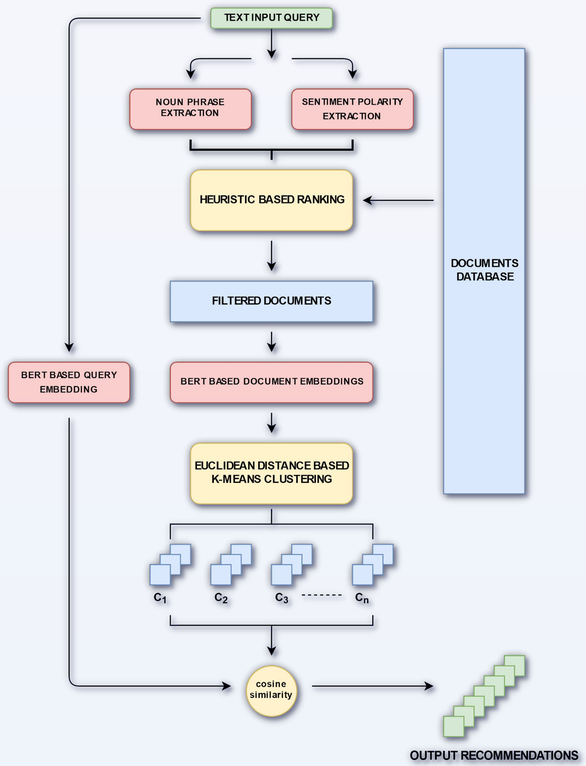
Metadata

- *Extracted Noun-Phrases*
- *Extracted Sentiment-Polarities*

ENTITY	NOUN PHRASES	SENTIMENT	WORDS
MEAN	37.042515	-0.208488	429.201159
STD. DEVIATION	34.820679	0.796597	500.339352

Statistics for Extracted Data

FRAMEWORK



$$similarity_{NOUN-PHRASE} = \frac{|QUERY_{NOUN-PHRASE} \cap DOCUMENT_{NOUN-PHRASE}|}{\sqrt{|QUERY_{TEXT}| \; X \; |DOCUMENT_{TEXT}|}}$$

$$similarity_{SENTIMENT} = |sentiment_{QUERY} - sentiment_{DOCUMENT}|$$

$$Heuristic\ Score = (\alpha \; x \; similarity_{NOUN-PHRASE}) + (1 - \alpha) \; x \; (similarity_{SENTIMENT})$$

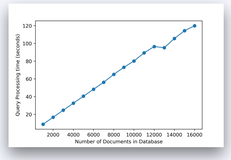
$$\alpha \in (0, 1)$$

bert_embedding	To extract contextual embeddings for the text-query & documents
textblob	To extract Noun-Phrases from the text-query & documents
nltk.VADER	To extract the sentiment-polarity score for the text-query & documents
scikit-learn	To perform K-means clustering over the contextual embeddings
regex	To pre-process the text-query

ANALYSIS

- Performed analysis by measuring **PRECISION** over recommendation outputs in terms of User Satisfaction
- Tested the Framework with **20 USERS**

- **HYPERPARAMETERS**:
1. Number of Filtered documents = 500
2. Number of Clusters = 10
3. alpha = 0.95



	PRECISION	QUERY-LENGTH	NP-COUNT	NP-COUNT-NORM	QUERY-SENTIMENT
PRECISION		0.329321	0.330093	-0.036580	-0.087509
QUERY-LENGTH	0.329321		0.483500	-0.201662	0.201573
NP-COUNT	0.330093	0.483500		0.701180	0.270583
NP-COUNT-NORM	-0.036580	-0.201662	0.701180		0.185872
QUERY-SENTIMENT	-0.087509	0.201573	0.270583	0.185872	

Correlation Table

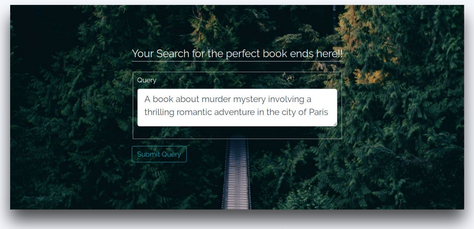
DISCUSSION

1. The Framework follows a **Near-Linear** time-complexity against the size of the database. [NEEDS IMPROVEMENT]
2. The Framework **Precision** follows a **considerable correlation** with **Query-Lengths**. [NEEDS IMPROVEMENT]
3. The Framework **Precision** follows a **low correlation** with **Normalized Noun-Phrase counts**.
4. The Framework **Precision** follows a **low correlation** with **Query Sentiment** .

FUTURE WORK

1. Adding multimodality to the framework by including image & speech based searches
2. Defining better heuristic to reduce query-time and better scaling over larger databases
3. Modeling user reading history to recommend new reading suggestions
4. Defining new empirical evaluation metrics for unsupervised text-based recommendation tasks
5. Expanding the framework for STEM & Academic literature databases

WEB APPLICATION



Text Query Input Interface

Books Matching Your Query

Book Title	Book Author	Book Summary
You Can't Go Home Again	Thomas Wolfe	George Webber has written a successful novel about his family and hometown. When he returns to that town, he is shaken by the force of outrage and hatred that greets him. Family and life-long friends feel naked and exposed by what they have seen in his books, and their fury drives him from his home. Outcast, George Webber begins a search for his own identity. It takes him to New York and a hectic social whirl; to Paris with an uninhibited group of ex-patriots; to Berlin, lying cold and sinister under Hitlers shadow. The journey comes full circle when Webber returns to America and rediscovers it with love, sorrow, and hope.
The Liquidator	nan	In Paris in 1944 Tank Corps Sergeant Boysie Oakes kills two Germans attempting to assassinate an Intelligence Corps officer named Mostyn. Twenty years later Mostyns memories have transformed Oakes (who is in reality cowardly and hedonistic) into a fearless master assassin though nothing could be further from the truth. Mostyn recruits Oakes into the Secret Service where after a training course he is given an enviable lifestyle. Oakes function is to "liquidate" security risks for the State. Oakes hires a mild mannered professional assassin to do his dirty work for him. Going for a "dirty weekend" leads to Boysie being

Output Recommendations from the database

Scan for our Open-Sourced Code

Scan for a video demonstration

REFERENCES & ACKNOWLEDGEMENTS

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (J Devlin et al.)(2018)
- CMU Book Summary Dataset (D Bamman & N Smith)(2013)
- Semantic Search on Text and Knowledge Bases (H Bast et al.)(2016)
- VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text (C Hutto & E Gilbert)(2014)
- NLTK: The Natural Language Toolkit (E Loper & S Bird)(2002)

We would like to thank the reviewers for their invaluable feedback on this work and suggestions for this poster presentation.