Comp8380 Final Project Report

Author X, Author Y

1 Overview

Table 1 summarizes the features we have implemented. (please modify the table according to your actual implementation.)

2 Search Engine 1.0

Our search engine can be accessed and evaluated at url . It supports spelling correction and ...They are implemented using Lucene API....

3 Embedding Evaluation

3.1 Models and Embeddings

We evaluate 5 models, including BERT, SBERT, LLAMA3..... Embeddings are obtained using (Gensim / Word2Vec /Doc2Vec / Glove / HuggingFace / SBERT / SentenceTransformer... elaborate depending what you actually do and add links to your code). For BERT, we take the CLS/last layer for embedding, for llama, we ...(continue with your hyper-parameters)

3.2 Evaluation in classification task

We evaluate those 4 models on the classification task using data MR and ICSE/SIGMOD. The F1 results are tabulated in Table **??** and plotted in Figure 1. From the figure we can see that model A outperforms B consistently. ....

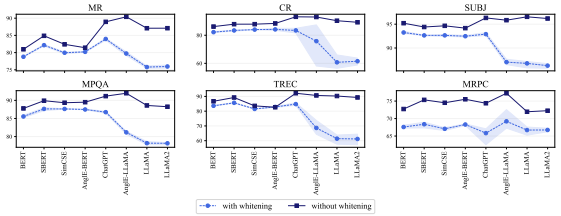

Figure 1: Comparison of models A and B on classification task

3.3 Evaluation in STS task

We evaluate those models on STS datasets ??? and ??? (and ...). Fig 2 shows the Spearman's correlation. We observe that model A outperforms B ... . This observation is consistent with the results reported in [][].

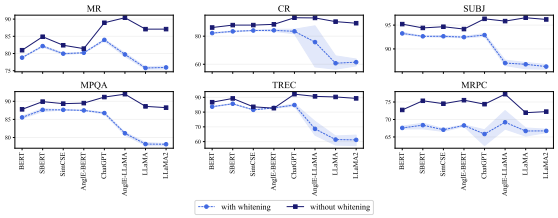

Figure 2: Comparison of models A and B on STS task

3.4 Fine-tuning embeddings

We fine-tune embeddings using both unsupervised and supervised methods. The data is ... the loss functions are .... Fig ? shows that method A outperforms other methods.... . The colab code is at ...

4 Search Engine 2.0

4.1 PageRank

PageRank algorithm is implemented using .... and experimented on the SIGMOD graph and Microsoft graph.

4.2 Vector Search

Vector index and search is implemented using .... (url) demonstrates that our search engine can find relevant documents that are not syntactically related. (here is a concrete example). We experimented with HNSM and (or others), and find that ....

4.3 ....

5 Acknowledgements

We want to give credits to ?? for feature A and ?? for feature B.

Table 1: Self Evaluation of the Project. ((Fill in `My Marks` column according to your actual work. The weight of the Phase II is 30%. There are 18 bonus marks for some 'deeper' topics, and 10 bonus marks for 'reusability'.)).

Sub-tasks	Marks	Break-downs	Marks	My Marks
Search Engine 1.0	5	Functional search engine	1	1
		2 features (e.g. fault tolerant, excerpt)	2	2
		large data	2	2 (used 5M microsoft data)
		Crawling/data addition	2 bonus	1 (crawled but not present in SE)
		Web interface	3 bonus	2 (localhost only, not accessible by others)
Embedding	15	Get embedding using 3+ methods	3	3 (tfidf, w2v, bert, sbert, llama ...)
		Eval in STS	4	3 (have table but no quality plot)
		Eval in classification	4	3 (F1 table only)
		LLM vs NB	4	2 (implemented NB. not compared with emb methods)
		Fine-tuning	4 bonus	1(run FT), 2(compared with non-FT), 1 (explore FT optio
Search Engine 2.0	10	PageRank	4	2 (implemented but not integrated)
		Vector search	4	2 (can index and search vectors, but not integrated with SI
		Clustering	2	1 (run clustering, not integrated with SE)
		Co-occurrence matrix and SVD	3 bonus	2 (run co-occurrence matrix and SVD, not compared with
		Terminology extraction	3 bonus	1 (tackled but not using MI, not integrated into SE)
		Near duplicate	3 bonus	
		...	...	
Adoption bonus	10	Adoption of feature 1 by Joe	1	
		Adoption of feature 2 by John	1	
		Adoption of feature 1 by Emily	1	
	...	...		
Total	30 +10		30+18+10	27