Final Report

Raj Bunsha — IMT2021010
Pannaga Bhat — IMT2021080
Kadaru Jashwanth Reddy— IMT2021095

Instructor: Prof. Raghuram Bharadwaj *Date:* May 1st, 2024

1 Overview

We have chosen "Blog Recommendation System" as our topic. For implementation purposes, we have specifically limited ourselves to particular types of blogs, so that analysis and evaluation of models becomes easier. We have chosen blogs related/relevant to users who are part of online competitive programming communities. For scraping of data we have specifically chosen 'Codeforces Blog Recommendation System'. The rest of the report contains the description of the dataset that we have constructed, the novel methods/models that we have formulated to recommend blogs to users, and the evaluation metrics that we added later upon feedback.

2 Introduction

Recommendation systems play a vital role in various online platforms, aiding users in discovering relevant content based on their preferences and behaviors. In this project, we focus on developing a recommendation system tailored for users involved in online competitive programming communities. Specifically, we aim to recommend blogs from platforms such as Codeforces, which provide valuable insights, tutorials, and discussions related to competitive programming.

3 Dataset Description

3.1 Dataset Construction

To construct our dataset, we utilized a combination of the Codeforces API and web scraping techniques with Selenium. We collected over 40,000 blogs from Codeforces, focusing on diverse topics within the competitive programming domain.

3.2 Data Overview

The dataset comprises various attributes for each blog entry, including the user, title, tags, and net upvotes/-downvotes. This information forms the basis for our recommendation system.

3.3 Preprocessing & EDA

Prior to analysis, we performed preprocessing steps such as data cleaning and exploratory data analysis (EDA) to gain insights into the distribution and characteristics of the dataset.

4 Methodology & Approaches

4.1 Methodology 1

Our methodology involves leveraging natural language processing (NLP) techniques and machine learning algorithms to develop effective recommendation models.

4.2 Representing blogs as vector

The different ways to get vector from text is as follows:

1. Spacy
2. Gensim model of Doc2Vec
3. BERT model

Once this is done we can move for cosine similarity and find the most similar blogs with the user representation

4.3 PCA for better representation and clustering

As the vector representing the blog can be of very high dimensions or represented in such a way that it cannot be easily clustered we get the PCA of the blog vectors this also helps us in getting the latent features that shape the blogs.

4.4 User representation

We start with trying to get user representation of to do so we started out with making user representation as average of all the blogs that user has liked. However this has a huge problem that average of the blogs just might not accurately represent the user. So we Use KMeans to get centroids of clusters formed from the blogs. They represent the likes of user. Thus the user representation is not a vector but a set of vectors, formed by clustering the blogs.

Now since we don't have a single vector as user representation we can instead get the similarity through all the all the centers and take the maximum which gives us the maximum correlation with all the topics that the user likes.

5 Sample results

5.1 Input:

This are the sample blogs that have been liked

5.2 Bert Image:

We can see blogs with similar tags

5.3 Gensim Image:

We can see blogs with similar tags

5.4 Spacy:

We can see blogs with similar tags

	id	content	tags
1033	65619	i have 2 arrays a[n],b[n].(think of them as co...	['array', 'algorithm complexity', 'comparison']
4830	84887	Be happy be healthyâ™¥	[]
4900	89847	Trying to solvethisproblem. I got the last non...	[]
5095	91171	1389B - Array WalkI have recently started lear...	[]
8960	88238	Recently I was preparing some problems in Poly...	['polygon', 'checker']
9160	86685	Hello and thanks, everyone. I'm new to c++ so ...	['#help me']
10150	119755	Hi, this side Imsfg. I used both map/unordered...	['c++', 'hashmap', 'unordered_map']
10442	114888	As you can see, today's Div2 Rank1syf2008is ac...	['cheater']
12186	49057	BSUIR Open Archives (polygon contests packages...	['bsuir', 'championship', 'open', 'gym']
12506	47337	my code :20895807input : AABCDEFGHIJKLMNOPQRST...	[]
14245	13408	Hello,Can anyone tell me a way to get thelist ...	['problemset']
16123	8317	Infinite SumCan anyone explain the solution.. ...	['yandex.algorithm 2013', 'math']
18313	79163	Hi guys :)Some of you might know that I have Y...	[]
18467	72383	I am new to codeforces .I always see scoring d...	[]
18632	74106	Hi everybody,This Sunday there will be aMoscow...	['622']
18721	73212	TL;DR I am thinking of holding a series of liv...	['livestream', 'videos', 'coaching']
19217	78357	Recently i was giving a competition, specifica...	['#problem solving']
21299	120471	Whenever I try to go into a problem by clickin...	[]
22139	112492	Hello Codeforcers!We are pleased to invite you...	[]
23305	103740	this is my code link which is not accepted:htt...	[]
27889	53385	Hello Codeforcers!We are pleased to invite you...	[]
31250	20985	why my last contset didnt count in rating ? i ...	[]
32930	15377	Hi,I dont know whom to send my query regarding...	[]
33591	1923	Problem 79Eis a challenging problem. I am happ...	['problem 79e', 'geometry']
34058	5867	The USACO 2012 November contest is available N...	['usaco', '2012', 'november contest']

Figure 1: Liked blogs

	id	content	tags
1033	65619	i have 2 arrays a[n],b[n].(think of them as co...	['array', 'algorithm complexity', 'comparison']
4830	84887	Be happy be healthyâ™¥	[]
4900	89847	Trying to solvethisproblem. I got the last non...	[]
5095	91171	1389B - Array WalkI have recently started lear...	[]
8960	88238	Recently I was preparing some problems in Poly...	['polygon', 'checker']
9160	86685	Hello and thanks, everyone. I'm new to c++ so ...	['#help me']
10150	119755	Hi, this side Imsfg. I used both map/unordered...	['c++', 'hashmap', 'unordered_map']
10442	114888	As you can see, today's Div2 Rank1syf2008is ac...	['cheater']
12186	49057	BSUIR Open Archives (polygon contests packages...	['bsuir', 'championship', 'open', 'gym']
12506	47337	my code :20895807input : AABCDEFGHIJKLMNOPQRST...	[]
14245	13408	Hello,Can anyone tell me a way to get thelist ...	['problemset']
16123	8317	Infinite SumCan anyone explain the solution.. ...	['yandex.algorithm 2013', 'math']
18313	79163	Hi guys :)Some of you might know that I have Y...	[]
18467	72383	I am new to codeforces .I always see scoring d...	[]
18632	74106	Hi everybody,This Sunday there will be aMoscow...	['622']
18721	73212	TL;DR I am thinking of holding a series of liv...	['livestream', 'videos', 'coaching']
19217	78357	Recently i was giving a competition, specifica...	['#problem solving']
21299	120471	Whenever I try to go into a problem by clickin...	[]
22139	112492	Hello Codeforcers!We are pleased to invite you...	[]
23305	103740	this is my code link which is not accepted:htt...	[]
27889	53385	Hello Codeforcers!We are pleased to invite you...	[]
31250	20985	why my last contset didnt count in rating ? i ...	[]
32930	15377	Hi,I dont know whom to send my query regarding...	[]
33591	1923	Problem 79Eis a challenging problem. I am happ...	['problem 79e', 'geometry']
34058	5867	The USACO 2012 November contest is available N...	['usaco', '2012', 'november contest']

Figure 2: Top results using bert

6 Methodology 2

Here, we try to group or categorize blogs into several bags, where each bag represents a topic. After we do this, we use MAB to explore and exploit user interests. This is a good strategy to learn the interests of a new user with no representation, in a manner where we simultaneously exploit the information we acquired from the user via implicit/explicit feedback, and also explore further, those topics where the user has a chance of finding reward.

	ID	Title	Content	Tags	Point	Closest Cluster	Similarity
0	73212	Solving problems on first sight + streaming	TL;DR I am thinking of holding a series of liv...	['livestream', 'videos', 'coaching']	[-2.6193121755473867, -1.3790123710790896, 3.7...	1	1.999813
1	49057	A little bit BSUIR Open into gym	BSUIR Open Archives (polygon contests packages...	['bsuir', 'championship', 'open', 'gym']	[-1.4790922973966387, 1.1760136585521195, -1.9...	3	1.998623
2	81358	I'm ecnerwala. Ask Me Anything!	Hi everyone!code_warriorrecently asked "How is...	['ama', 'interview', 'social', 'ecnerwala']	[-1.2109257036676127, -1.3922421517603145, 2.1...	1	1.738015
3	20489	What's slower than a segment tree and needs so...	It'ssqrt decomposition. It didn't rime. Bummer...	[]	[5.318040170927215, 0.4252801709976638, 3.6447...	0	1.705204
4	46323	AtCoder Grand Contest 002	AtCoder Grand Contest 002will be held on Sunda...	['atcoder', 'agc', 'agc002']	[-0.4557780972498998, 0.5157800955597956, -0.8...	3	1.690436
5	78815	(Maybe) A Bug Of Codeforces Interactive Proble...	In this problem:1364E - X-OR,the discription s...	[]	[0.2923337398530138, -1.6323641414485264, 0.02...	4	1.668582
6	20276	I will be doing 24h livestream during Marathon...	Updated on 07/12 for the last timePostmortem--...	['livestream', 'marathon24', 'psyhogonewild']	[-3.010387476689212, 0.3709521356919836, 3.811...	1	1.663143
7	7087	Codeforces Round #175 (Div. 2)	Good day, friends)Welcome to regular Codeforce...	['codeforces', 'round', '175']	[-0.35446753696755834, 1.1836930535402639, -0....	3	1.652974
8	61239	Blogewoosh #2	Hello, codeforces!All signs in the sky and on ...	['blogewoosh']	[5.526446065698168, 2.273668987526779, 2.13987...	0	1.652061
9	74017	Screencast collection	I will update this blog when I upload new scre...	['screencast']	[-0.8996442776611158, 0.18249550113377527, -0....	3	1.646925
10	49748	Google Chrome extension for automatic sample c...	Hello all, I'd like to share with you a google...	['tester']	[-3.1936102897983623, -2.7541294829990286, 3.7...	1	1.643370
11	58207	MathMash â€" Round 2	Hello Codeforces!We will be hosting thesecond ...	['mathmash', 'contest', 'math']	[-1.988142368815055, 1.0795740116178048, 1.738...	1	1.642629
12	83511	[Gym] XXI Open Cup. Grand Prix of Korea	Hello! I'm happy to announce XXI Open Cup. Gra...	['opencup', 'korea']	[-2.1858114388101124, 1.557129689520686, -1.70...	3	1.637138
13	83248	[Tutorial] Square root decomposition and appli...	A brief introduction into the applications of ...	['#data structures']	[8.898522637390693, 0.0975905733292326, 3.4547...	0	1.628477

Figure 3: Top results using gensim

	ID	Title	Content	Tags	Point	Closest Cluster	Similarity
0	49057	A little bit BSUIR Open into gym	BSUIR Open Archives (polygon contests packages...	['bsuir', 'championship', 'open', 'gym']	[0.6721071451312359, 1.6719665755113329, -0.85...	2	1.998623
1	47060	Helping Contestants Help Us All or Competition...	During my competition history I've collected a...	['problemsetting']	[-0.9754051311150096, -0.5132965739437328, 0.0...	0	1.933445
2	112856	If I can do it, anyone can	I'm writing this blog as a brief account of my...	[]	[-0.9408107488894119, -0.4487585443743908, 0.0...	0	1.932874
3	98621	Self-deception: maybe why you're still grey af...	I generally don't like to give much advice on ...	[]	[-0.8238899417379001, -0.47182073317265444, 0....	0	1.930273
4	83895	My thoughts on clarifications and on the round...	Alright, I'm done. I wanted to write a blog li...	[]	[-0.7155926381190538, -0.41588837692613895, 0....	0	1.923100
5	94765	Changes in CodeChef Problemsetting	Hello Codeforces!For the last half a year I ha...	['codechef']	[-0.7683515448516817, -0.37817487944130584, 0....	0	1.921006
6	110245	Red Isn't Impossible	I waited for more than a year to write this bl...	[]	[-0.7225991509366844, -0.3836328887848367, 0.0...	0	1.913017
7	6928	Judges adding tests against submitted solution...	In the discussion athttp://codeforces.com/blog...	[]	[-0.775915036464599, -0.4463080483494599, -0.0...	0	1.912728
8	103077	You can do it, too!	IntroductionSo, I set myselfthat challengeto r...	[]	[-0.7464074945562664, -0.328822340655177, 0.08...	0	1.911110
9	96313	False Positive in Codeforces' Anti-Cheat Detec...	On the recent round 750 (div. 2), I originally...	['cheating', 'anti-cheat']	[-0.823031863329896, -0.4447989281660075, 0.03...	0	1.904396
10	124515	Does Polygon automaticly send statements and t...	Recently Polygon got enhanced with AI-based fe...	['polygon', 'ai']	[-0.8907374379333005, -0.499032022851545, -0.0...	0	1.904386
11	79151	UpSolve.me | Efficient Codeforces Practice	Hello, Codeforces community!First of all, I wo...	[]	[-0.7785181893594785, -0.45844553870306365, 0....	0	1.902181
12	73212	Solving problems on first sight + streaming	TL;DR I am thinking of holding a series of liv...	['livestream', 'videos', 'coaching']	[-0.8299719856534992, -0.3633904615821784, -0....	0	1.901773
13	22889	My sad story	Teams advancing to 2016 WF Phuket in Asia have...	['acm icpc 2015', 'acm icpc', 'acm icpc region...	[-0.8632620900073557, -0.4581740873341882, -0....	0	1.900844
14	49346	Revoting for comments: dreams come true?	They do! In some way.For eager oneslink to use...	['codeforces', 'comments', 'userscript']	[-0.7390193329340394, -0.3147566023091068, -0....	0	1.900513
15	71899	Fast Walsh Hadamard Transforms and it's inner ...	One of my favorite algorithms out there is FWH...	['#tutorial', 'fwht']	[-0.6186067757119754, -0.3153192005780671, 0.0...	0	1.898926

Figure 4: Top results using spacy

6.1 Topic Modelling of Blogs using N-grams + LDA

Documents (in this case blogs) are represented as a bag of words. In this case, for better representation and better results, we try to represent the documents as a bag of n-grams (a bag of words/n-grams, is basically a collection of tokens/strings) & then pass the processed blogs to LDA model which then fits and labels the blogs with respective topic numbers.

Each topic is modelled as a distribution over tokens or words. We have tried bi, tri, quad, and pentagrams, and obtained results. Tri-grams gave better probabilities for most weighted word in each topic.

6.2 MAB - Upper Confidence Bound for Recommendations

We then used multi-arm bandit to make batch recommendations to users and simulated user interaction programmatically by programming predefined stochastic user behaviour. The MAB model uses the Upper Confidence Bound (UCB) method to calculate the reward associated with each arm. This method takes into account the number of trials made per arm, and follows the principle of "optimism in the face of uncertainty". If a particular topic is not tried or recommended as many times as other topics, then the MAB chooses that topic with a high probability initially, to compose the next batch of recommendations. The topics that have a high average reward, are also picked with a high probability later in the recommendation process.

6.3 Sample Results:

Below are pics of Topic modelling output, dataset after topic labelling, and changing probabilities over topics. (Topics cannot be described in well-defined english terms, as they are generated and are represented as distribution over tokens.)

```
Name: 4918, dtype: object
Topic 0: 0.108*"! !" + 0.018*"problem ." + 0.017*"wrong answer" + 0.014*"binary search" + 0.010*"help !" + 0.008*"open contest" + 0.008*"contest (" + 0.008*"contest #" + 0.007
*"... ..." + 0.007*"# 3"
Topic 1: 0.021*"– round" + 0.019*"dynamic programming" + 0.013*"help !" + 0.011*"number theory" + 0.010*"need help" + 0.008*": (" + 0.008*"online judge" + 0.007*"? !" + 0.007*"r
ound 1" + 0.006*") problem"
Topic 2: 0.020*"round 2" + 0.018*"code jam" + 0.015*"( rated" + 0.015*"google code" + 0.014*"– editorial" + 0.013*"( div.2" + 0.013*") ?" + 0.012*"round 3" + 0.008*"problem –" +
0.007*"contest –"
Topic 3: 0.014*"vk cup" + 0.014*"hacker cup" + 0.012*"facebook hacker" + 0.010*"round 1" + 0.009*"solution ?" + 0.007*"2021 (" + 0.007*"open cup" + 0.006*"contest ?" + 0.006*"ti
me limit" + 0.006*"tle ?"
Topic 4: 0.028*"help needed" + 0.026*"programming contest" + 0.026*"problem ?" + 0.017*"div 2" + 0.016*"topcoder srm" + 0.016*", b" + 0.016*", c" + 0.015*"b ," + 0.014*"solve pr
oblem" + 0.012*", ,"
Topic 5: 0.124*"round #" + 0.093*"codeforces round" + 0.067*"div ." + 0.055*"( div" + 0.050*". 2" + 0.046*"2 )" + 0.032*"need help" + 0.032*"beta round" + 0.030*"codeforces bet
a" + 0.012*"3 )"
Topic 6: 0.031*"segment tree" + 0.031*"runtime error" + 0.017*"( )" + 0.011*"test case" + 0.011*"qualification round" + 0.011*"time complexity" + 0.010*"graph problem" + 0.009
*"problem b" + 0.008*": )" + 0.007*"problems ?"
Topic 7: 0.018*") –" + 0.013*"# 2" + 0.012*"getting tle" + 0.010*"+ div" + 0.009*"problem e" + 0.009*"1 +" + 0.009*"test cases" + 0.008*"help please" + 0.007*"2 ," + 0.007*"? ("
Topic 8: 0.063*"? ?" + 0.016*"problem c" + 0.015*"( n" + 0.012*"spoj problem" + 0.011*"n )" + 0.010*"'s algorithm" + 0.008*"interesting problem" + 0.008*"topcoder open" + 0.007
*"help solve" + 0.006*"anyone help"
Topic 9: 0.031*"competitive programming" + 0.029*"invitation codechef" + 0.018*"new year" + 0.010*"tree problem" + 0.009*"long challenge" + 0.008*"codeforces :" + 0.007*"program
ming ?" + 0.006*"help :" + 0.005*"happy new" + 0.005*"codeforces rating"
Topic 10: 0.017*"beginner contest" + 0.017*"world finals" + 0.017*"c ," + 0.013*"atcoder beginner" + 0.013*"1 )" + 0.010*"data structure" + 0.010*") )" + 0.009*"editorial codefo
rces" + 0.009*"olympiad informatics" + 0.008*"segment trees"
Topic 11: 0.031*"educational codeforces" + 0.019*"rated div" + 0.017*"tutorial ]" + 0.016*"ca n't" + 0.015*"[ tutorial" + 0.014*"acm icpc" + 0.012*"[ help" + 0.011*"problem :" +
0.010*"2 ]" + 0.010*"codeforces ?"
                                              title  topic
0             "ICPC India Online Round 2019 Discussion"      3
1                          "Need help in this problem"      5
2                             "mobil balance problem"      9
3      "Need help in the problem 'The least Round Way'."      5
4                                    "Help with DFS"      3
...                                               ...    ...
41458                       "Best way to String matching"      6
41459  "Japanese Olympiad in Informatics (JOI) 2016 –...      9
41460  "Junior Programming Camp BUET 2016 -Day3 conte...     10
41461  "No of ways to express N as a sum of k partiti...      1
41462                    "Educational Codeforces Round 7"     11

[41462 rows x 2 columns]
```

Figure 5: Topic Modelling using n-grams + LDA

6.4 Topic labelled dataset:

Dataset after topic labelling

```
df_new.head(10)
```

	id	title	content	authorHandle	authorID	rating	tags	title_trigrams	topic	custom rating score
0	70681	"ICPC India Online Round 2019 Discussion"	"This blog is to discuss the problems of ICPC ...	kjain1810	1	113	[]	['"icpc india online', 'india online round', '...	2	1.978457
1	70261	"Need help in this problem"	"This is the problemhttps://codeforces.com/con...	prabalsingh24	2	0	[]	['"need help in', 'help in this', 'in this pro...	6	1.000000
2	68226	"mobil balance problem"	"When multiple objects of weight are given,Obt...	bluespace	3	-17	[]	['"mobil balance problem"']	9	0.672523
3	68713	"Need help in the problem 'The least Round Way'."	"I am getting TLE in TC 11 in the problem,The ...	suvro_coder	4	0	['#dp']	['"need help in', 'help in the', 'in the probl...	6	1.000000
4	70585	"Help with DFS"	"i've been solving this problem for a day nowh...	tensaida	5	-3	[]	['"help with dfs"']	6	0.940072
5	70358	"Codeforces Round 591 (and Technocup 2020 — El...	"1223A - CMEIdea:RomsTutorialTutorial is loadi...	Roms	6	112	[]	['"codeforces round 591', 'round 591 (and', '5...	4	1.977587
6	70730	"Problem — any help?"	"We have a two strings:AandB. We need to find ...	imaking	7	4	['#interview', '#strings']	['"problem — any', '— any help?"']	0	1.079830
7	71563	"Tricks or best approaches for greedy(ad-hoc) ...	"Hi! I'm looking for the best approaches to fi...	baka_	8	-3	['#greedy', '#adhoc']	['"tricks or best', 'or best approaches', 'bes...	5	0.940072
8	67292	"Purpose of Codeforces "	"Hi all, I am new to this website.However I wo...	I_love_GOT	9	0	[]	['"purpose of codeforces', 'of codeforces "']	4	1.000000
9	67452	"Changing 'int' to 'Integer' gives significant...	"I tried to solvethisproblem. Refer to my solu...	vivekiitian	10	24	[]	['"changing \'int\' to', "'int' to 'integer'",...	6	1.446244

Figure 6: dataset df.head

6.5 MAB changing probabilities of rewards associated with Blogs for particular user:

We can see randomly assigned probabilities moving towards actual probability scores.

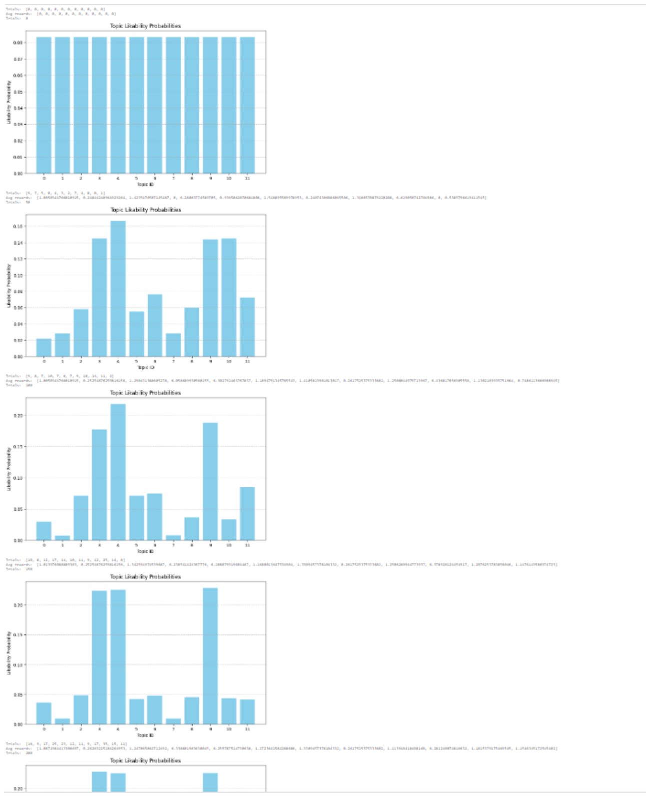

Figure 7: MAB-UCB results

7 Evaluation of Models

7.1 Evaluation Metrics

We evaluated the performance of our models using metrics such as precision, recall, and mean average precision (MAP), considering both overall performance and user-specific recommendations.

7.2 Results

Our results indicate promising performance of the recommendation models, with high precision and recall scores. Additionally, user feedback and interaction data further refine the recommendations over time.

8 Conclusion

In conclusion, our project demonstrates the feasibility and effectiveness of developing a blog recommendation system tailored for competitive programming communities. By leveraging advanced NLP techniques and machine learning algorithms, we have successfully provided users with personalized and relevant blog suggestions, thereby enhancing their overall experience on platforms like Codeforces.