ENS 491-492 – Graduation Project

Project Title :

Posted your question to Q&A community websites.
Now what? Just wait?
Towards developing the next step...

Group Members :

23883 Mehmet Utku Eray - 20510 Zeynep Seda Birinci
24142 Deniz Bozkurt - 23914 Selen Özcan

Supervisor: Reyyan Yeniterzi

Sabancı University

May 2020

Overview

Summary

The web has changed how individuals provide, search and offer information. Community Question Answering websites are built to provide the searched information more efficiently by utilizing an interface for users to exchange and share knowledge.

Stack Exchange sites are one of the most popular Q&A community websites to find answers related to pretty much every aspect of coding.

This project aims to use the data retrieved from the Stack Exchange sites to improve them in terms of user experience. In order to demonstrate these improvements, a user-friendly website is implemented.

Problem Statement

- Inquirers do not have the information on whether their question gets an answer or not.
- There is uncertainty regarding when the questions will receive an answer.
- Q&A community websites such as Stack Exchange do not provide users the information of context-based similar questions.
- Some questions do not get an answer even though there are many users actively using the platforms.

Objectives & Tasks

- Inquirer's point of view: "Will my questions get answered?"
- Inquirer's point of view: "When will my question get answered?"
- Inquirer's point of view: "Are there any similar questions asked on Stack Exchange Q&A community websites?"
- Responder's point of view: "Which questions am I more likely to answer correctly?"
- User-friendly website

Methodology

Inquirer's point of view: "Will my questions get answered?"

In this task the goal is to predict whether the question will get an answer or not, at the time of the questions being asked. To make a prediction on answers, Random Forest Classifier has been used. Features that are used to construct training and test sets are shown in the table below:

Features	Definition of a feature
Content	Title of a post +Body of a post
Tags	Related tags of a post
Creation Date	Creation Date of a post
Down Votes	Down Vote count of a user calculated based on the answer he/she gave to a question
Up Votes	Up Vote count of a user calculated based on the answer he/she gave to a question
Reputations	Reputation of a user that is decided accordingly down and up votes counts
Views	Number of views of the user page
User Creation Date	Creation Date of a user in the site
User Last Access Date	Last Access Date of a user to the site
Tag Count	Number of tags that the question has
IsAnswered	This is the label for classification task, If the question has been answered it is 1 otherwise 0

Methodology

Inquirer's point of view: "Will my questions get answered?"

Label encoding was used to convert date format data into numerical values. Time intervals that are used, shown in the table below. Later, one-hot encoding was implemented to allow efficient use of these features.

	Part of Day Classes based on UTC	**Corresponding time intervals in GMT-4**
1	05:00-10:00 →Morning(UTC)	24:00 - 05:00 →NY(GMT-4) Night
2	11:00 16:00 →Midday (UTC)	06:00 - 11:00 →NY(GMT-4) Morning
3	17:00 22:00 →Afternoon(UTC)	12:00 - 17:00 →NY(GMT-4) Midday
4	23:00 04:00 →Night (UTC)	18:00 - 23:00→ NY(GMT-4) Afternoon

Methodology

Inquirer's point of view: "Will my questions get answered?"

- For tags data of posts, the CountVectorizer method from scikit-learn library was used to represent data numerically.
- For (title + body of questions) questions, the Doc2Vec method from gensim library was used to represent text data as vectors.

Classification methods have been tried for prediction on answer. The table shown below indicates the F1 score of each classifier has been used.

Model	F1 Score
Decision Tree Classifier	0.78
K Neighbors Classifier	0.73
Random Forest Classifier	0.80
SVM	0.77
Gradient Boosting Classifier	0.80

Methodology

Inquirer's point of view: "When Will my questions get answered?"

In this task the goal is to predict when the question will get an answer by calculating how much time is expected to elapse after posting a question on the platform. Features that are used in final data sets are shown in the table below.

Features	**Definition of a feature**
Content	Title of a post +Body of a post
Tags	Related tags of a post
Creation Date	Creation Date of a post
Down Votes	Down Vote count of a user calculated based on the answer he/she gave to a question
Up Votes	Up Vote count of a user calculated based on the answer he/she gave to a question
Reputations	Reputation of a user that is decided accordingly down and up votes counts
Views	Number of views of the user page
User Creation Date	Creation Date of a user in the site
User Last Access Date	Last Access Date of a user to the site
QuestionLength	Length of the body of the question
isQuestion	Existence of a question mark at the very end of a question title
isQuestionwh	Existence of a "wh" at the very beginning of a question title
isWeekend	Boolean feature that checks if the question was created on weekend or not
Tag Count	Number of tags that the question has
Diff in Minutes	How many minutes has elapsed to answer the question, it will be used as a label in training set

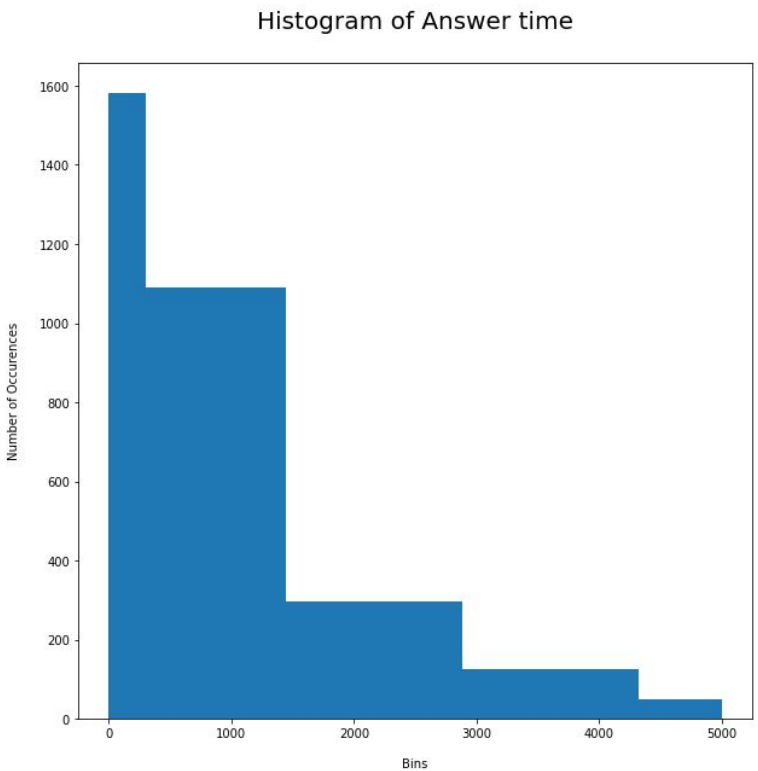
Methodology

Inquirer's point of view: "When Will my questions get answered?"

- For the date format features, label encoding and one-hot encoding were used.
- Tags were represented numerically by the help of CountVectorizer method.
- Questions were represented in a vector format by the help of Doc2Vec method.
- Outliers in response time in minutes have been omitted.

Methodology

Inquirer's point of view: "When Will my questions get answered?"



- Response times were categorized as first 5 hours, 5 hours-1 day, 1-2 days, 2-3 days and 3-4 days.
- The response time data is quite imbalanced. To balance the data, ADASYN oversampling method was used.

After the data distribution has been normalized, classification methods have been implemented.

Methodology

Inquirer's point of view: "Are there any similar questions asked on Stack Exchange Q&A community websites?"

- Data dump generated by Stack Exchange for their Q&A websites and build a xml tree.
- Xml tree is stored in to a pandas dataframe.
- Dataframe is split in to two sets for training and testing.
- There has been several adjustments on different parameters for the Doc2Vec model to obtain the best possible results by training
- The model generated with titles, bodies and tags of questions returns the most similar questions depending on the inferred vector.
- Model generated with TFIDF has the same approach as the Doc2Vec model, however it provides different similarity values for the same test corpus.
- This difference favors the Doc2Vec model.

Methodology

Responder's point of view: "Which questions am I more likely to answer correctly?"

- The methodology used for this task is similar to the one used in "Are there any similar questions asked on Stack Exchange Q&A community websites?" task.
- There are three parts to this task, which are suggesting responder's questions based on their overall profile, displaying similar questions individually for each question that the responder have answered and finally displaying questions based on responder's preferred tags.
- Both suggesting questions based on responder's profile and suggesting questions based on each question that the responder has answered utilizes the same model.
- Regarding searching questions based on preferred tags, the aim is to display questions by generating results using the responder's profile and filtering out the questions that do not have preferred tags.

Results & Discussion

Inquirer's point of view: "Will my questions get answered?"

In classification methods the highest accuracy score has been obtained from Random Forest Classifier with 80% F1 score. The confusion matrix of each method demonstrated in the table below:

Models	True positive	False Positive	False Negative	True Negative	Support values
Decision Tree Classifier	35	83	29	358	118 -true 387-false
K Neighbors Classifier	64	54	84	303	118 -true 387-false
SVM Classifier	0	118	0	387	118 -true 387-false
Random Forest Classifier	30	88	15	372	118 -true 387-false
Gradient Boosting Classifier	30	88	15	372	118 -true 387-false

Results & Discussion

Inquirer's point of view: "When Will my questions get answered?"

The task had been identified as a regression task at the beginning. However, the mean square error of the regression model was quite high to predict significantly.

Model	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Linear Regression	356.194	193332.928	439.696

Thus, the problem has been transformed into a classification task. Classification report of each classifier that has been tried are shown in the table below:

Model	F1 score of 1st class 0-5 hours	F1 Score of 2nd class 5hours-1day	F1 Score of 3rd class 1-2 days	F1 Score of 4th class 2-3 days	F1 Score of 5th class 3-4 days	Overall F1 Score
Random Forest Classifier	0.67	0.27	0.36	0.24	0.22	0.38
Logistic Regression	0.02	0.20	0.00	0.00	0.31	0.19
KNN Classifier	0.23	0.25	0.12	0.25	0.19	0.21
Support Vector Machine	0.09	0.29	0.00	0.30	0.14	0.22
Gradient Boosting Classifier	0.63	0.29	0.32	0.32	0.10	0.36

Results & Discussion

Inquirer's point of view: "When Will my questions get answered?"

The table shown below is the confusion matrix of the final model of this task. The model still can be improved since F1 scores of each class and overall F1 score is low.

1st class 0-5 hours	2nd class 5 hours-1 day	3rd class 1-2 days	4th class 2-3 days	5th class 3-4 days	Support values
275	71	1	0	0	347
145	103	68	49	10	375
44	52	141	75	35	347
35	56	169	85	6	351
3	66	172	44	65	350

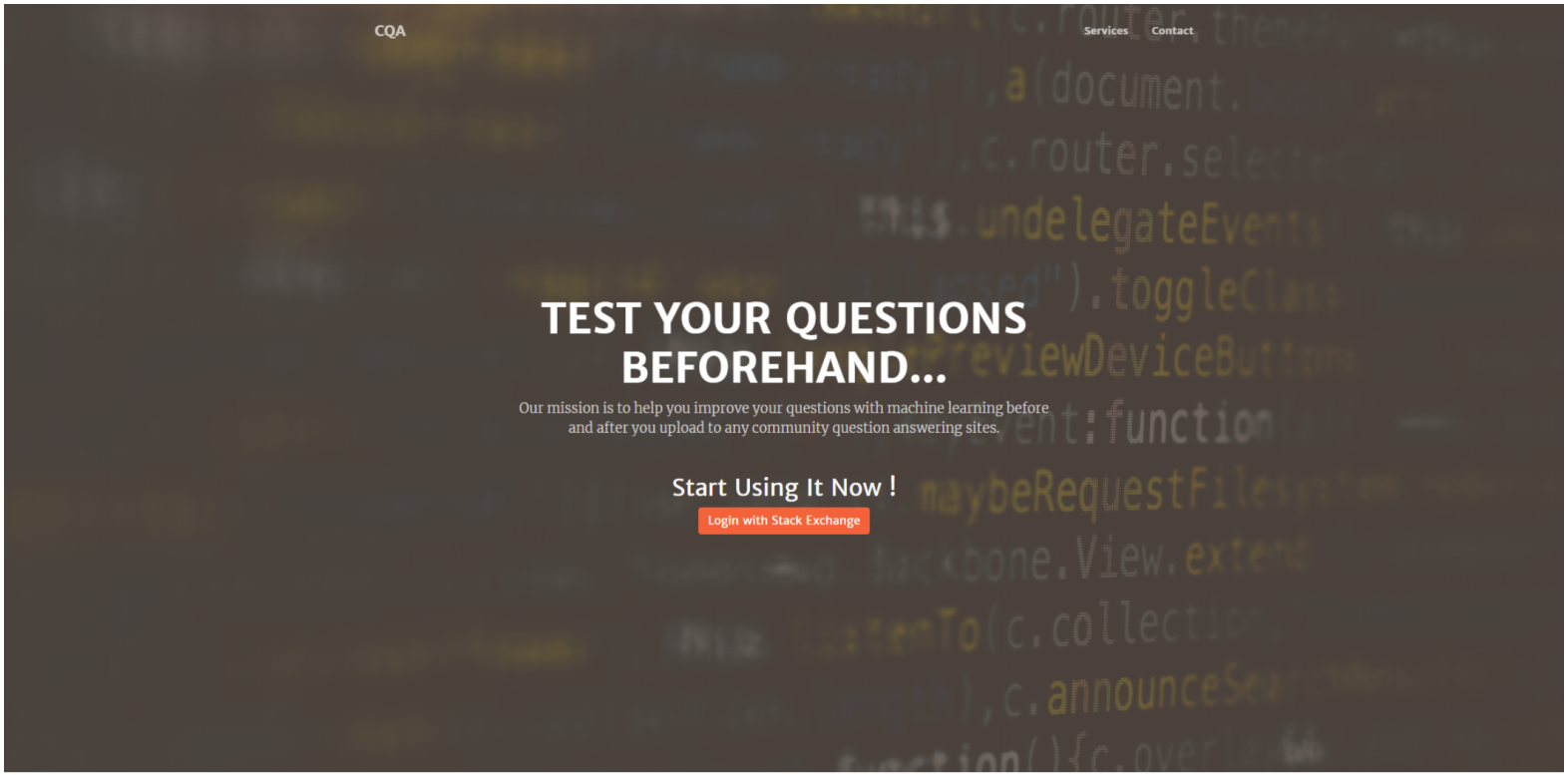
Results & Discussion

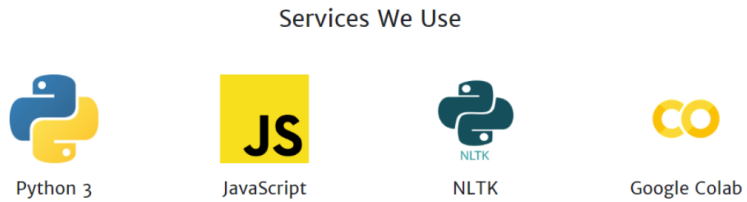
- One of the tasks that was written in the progress report which is "Inquirer's point of view: If I receive multiple answers to my question, which are/is better solution/s?" has been changed.
- The changed task was a feature for the inquirer which determined the better answer for his/her question by comparing the answers for the question.
- This task was replaced by finding similar questions on the website task which was referred to in Tasks section and Methodology section.

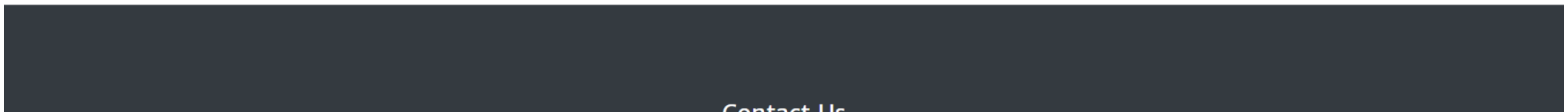
Results & Discussion

- "Are there any similar questions asked on Stack Exchange QA community websites?" task has been completed. Our model is capable of finding the questions that are similar to the given question.
- On average, Doc2Vec performed better than TFIDF since its results matched the related questions listed by the Stack Exchange itself more accurately and context wise the resulting questions were more similar.
- "Which questions am I more likely to answer correctly?" task has been completed. This task required the same methodology in the "Are there any similar questions asked on Stack Exchange QA community websites?" task.
- Same Doc2Vec model is utilized for suggesting questions that can be answered by the user.

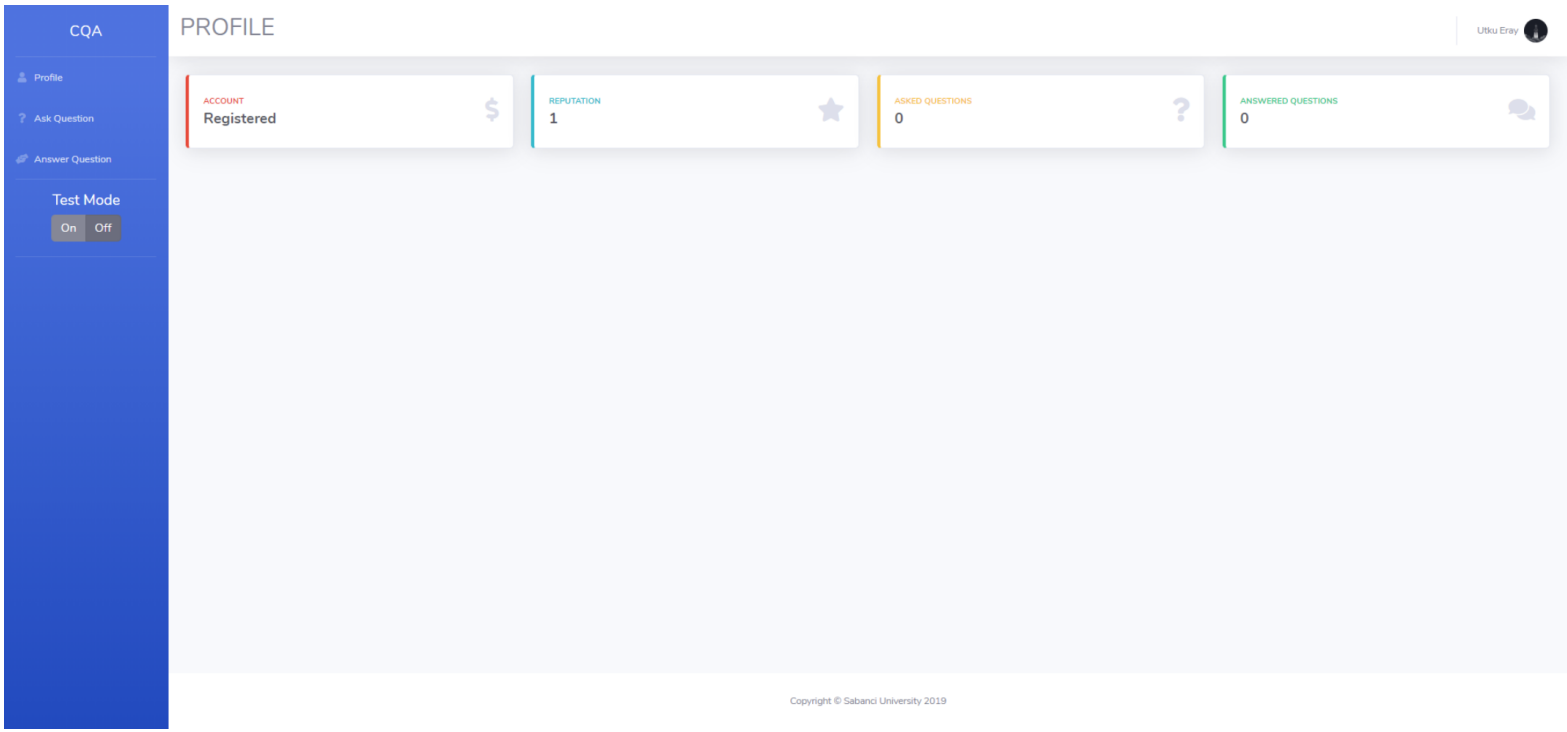
Front-End


CQA
Services
Contact
TEST YOUR QUESTIONS BEFOREHAND...
Our mission is to help you improve your questions with machine learning before and after you upload to any community question answering sites.
Start Using It Now !
Login with Stack Exchange

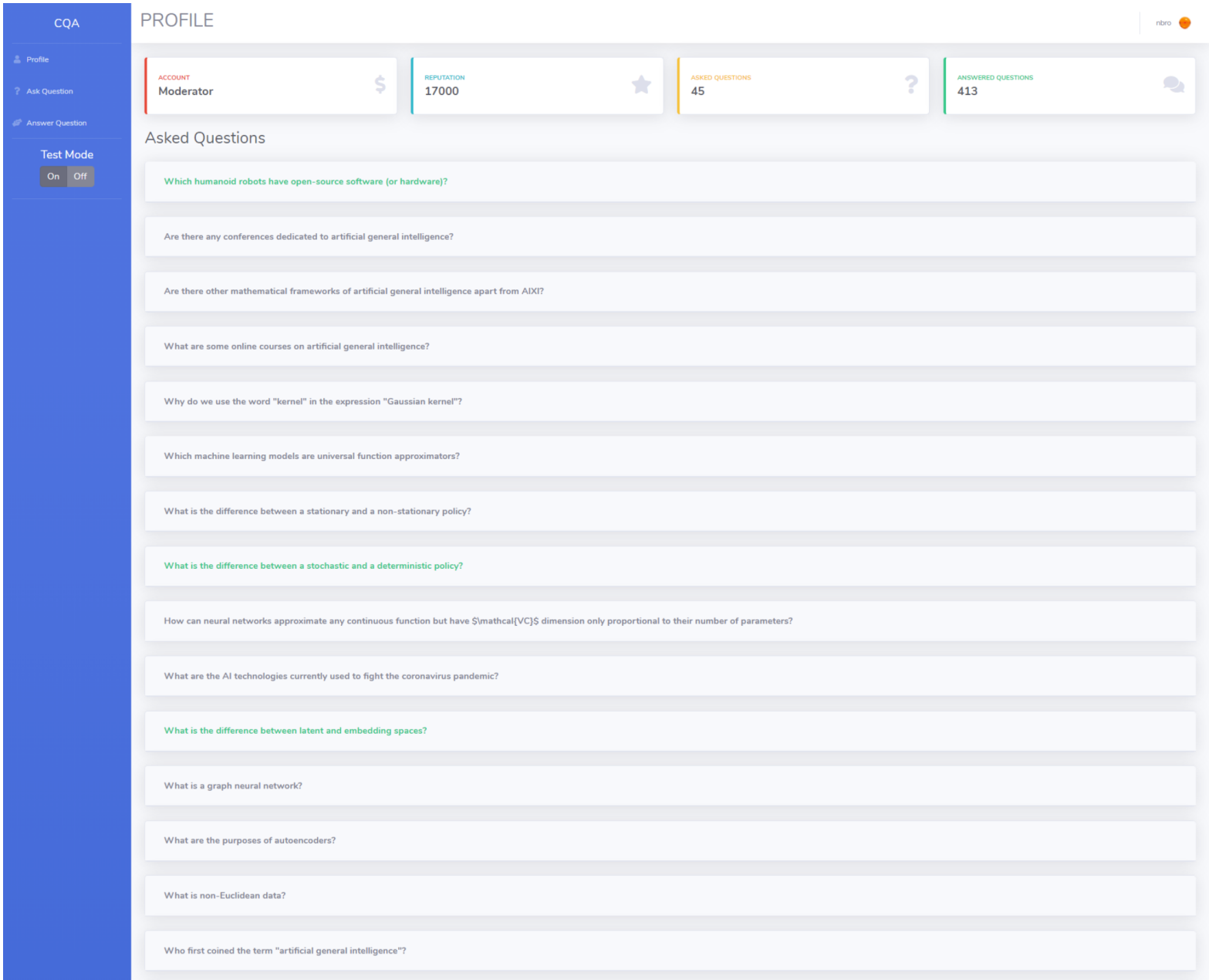

Services We Use
Python 3
JavaScript
NLTK
Google Colab



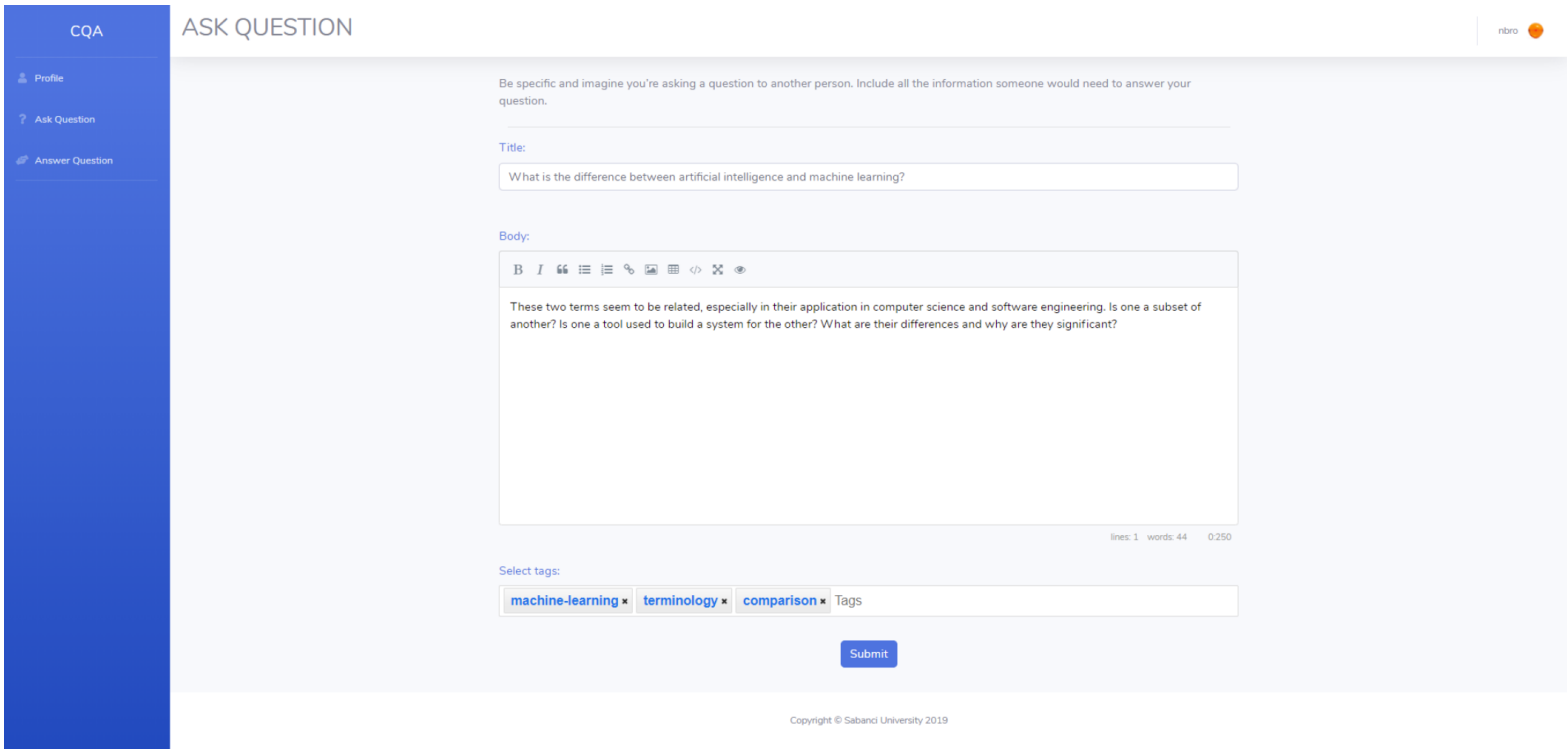
Front-End


CQA
Profile
Ask Question
Answer Question
Test Mode
On
Off
PROFILE
Utku Eray
ACCOUNT
Registered
REPUTATION
1
ASKED QUESTIONS
0
ANSWERED QUESTIONS
0
Copyright © Sabanci University 2019

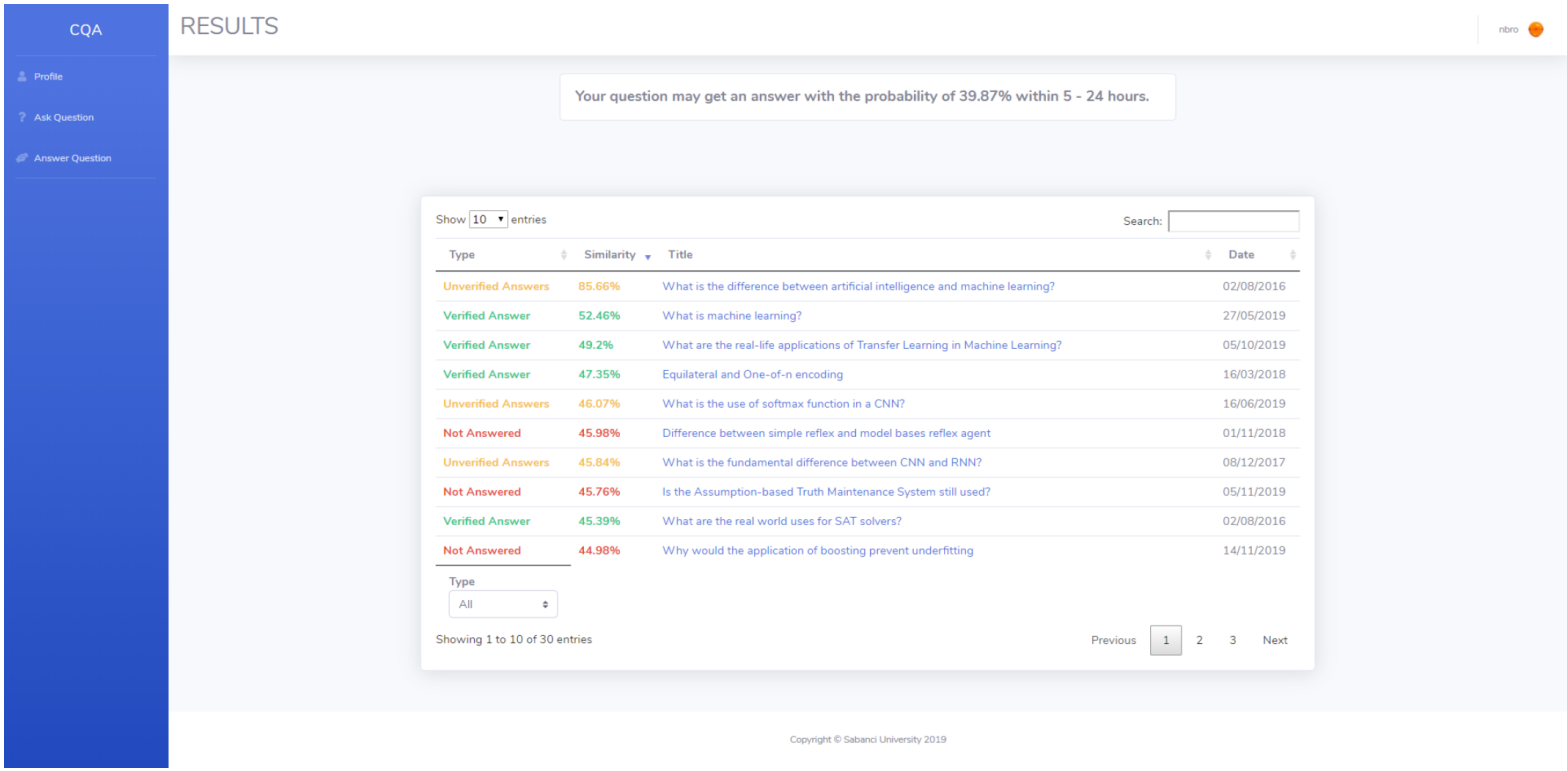
Front-End


CQA
Profile
Ask Question
Answer Question
Test Mode
On
Off
PROFILE
nbro
ACCOUNT
Moderator
REPUTATION
17000
ASKED QUESTIONS
45
ANSWERED QUESTIONS
413
Asked Questions
Which humanoid robots have open-source software (or hardware)?
Are there any conferences dedicated to artificial general intelligence?
Are there other mathematical frameworks of artificial general intelligence apart from AIXI?
What are some online courses on artificial general intelligence?
Why do we use the word "kernel" in the expression "Gaussian kernel"?
Which machine learning models are universal function approximators?
What is the difference between a stationary and a non-stationary policy?
What is the difference between a stochastic and a deterministic policy?
How can neural networks approximate any continuous function but have $\mathcal{VC}$ dimension only proportional to their number of parameters?
What are the AI technologies currently used to fight the coronavirus pandemic?
What is the difference between latent and embedding spaces?
What is a graph neural network?
What are the purposes of autoencoders?
What is non-Euclidean data?
Who first coined the term "artificial general intelligence"?

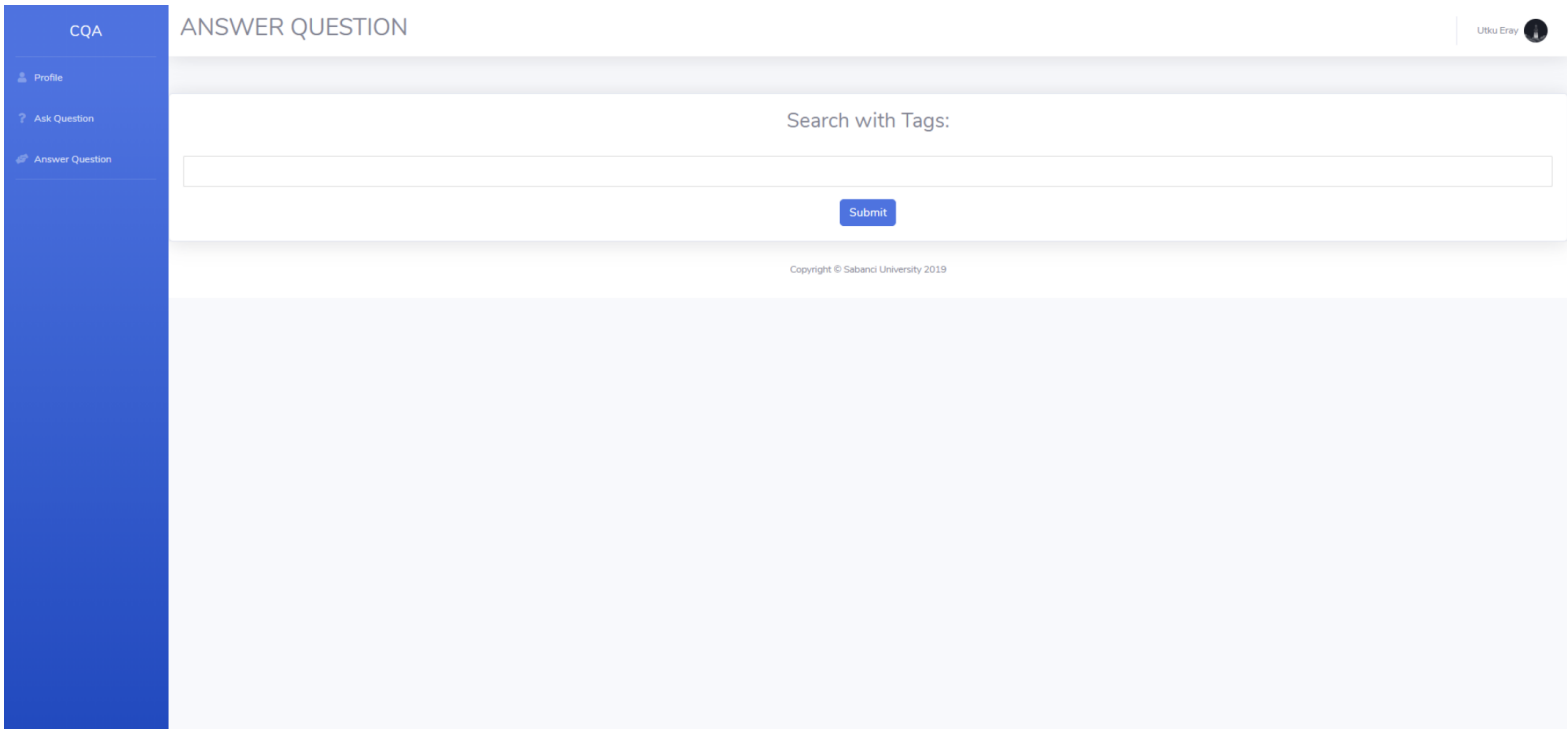
Front-End


CQA
ASK QUESTION
nbro
Profile
Ask Question
Answer Question
Be specific and imagine you're asking a question to another person. Include all the information someone would need to answer your question.
Title:
What is the difference between artificial intelligence and machine learning?
Body:
These two terms seem to be related, especially in their application in computer science and software engineering. Is one a subset of another? Is one a tool used to build a system for the other? What are their differences and why are they significant?
lines: 1 words: 44 0/250
Select tags:
machine-learning × terminology × comparison × Tags
Submit
Copyright © Sabanci University 2019

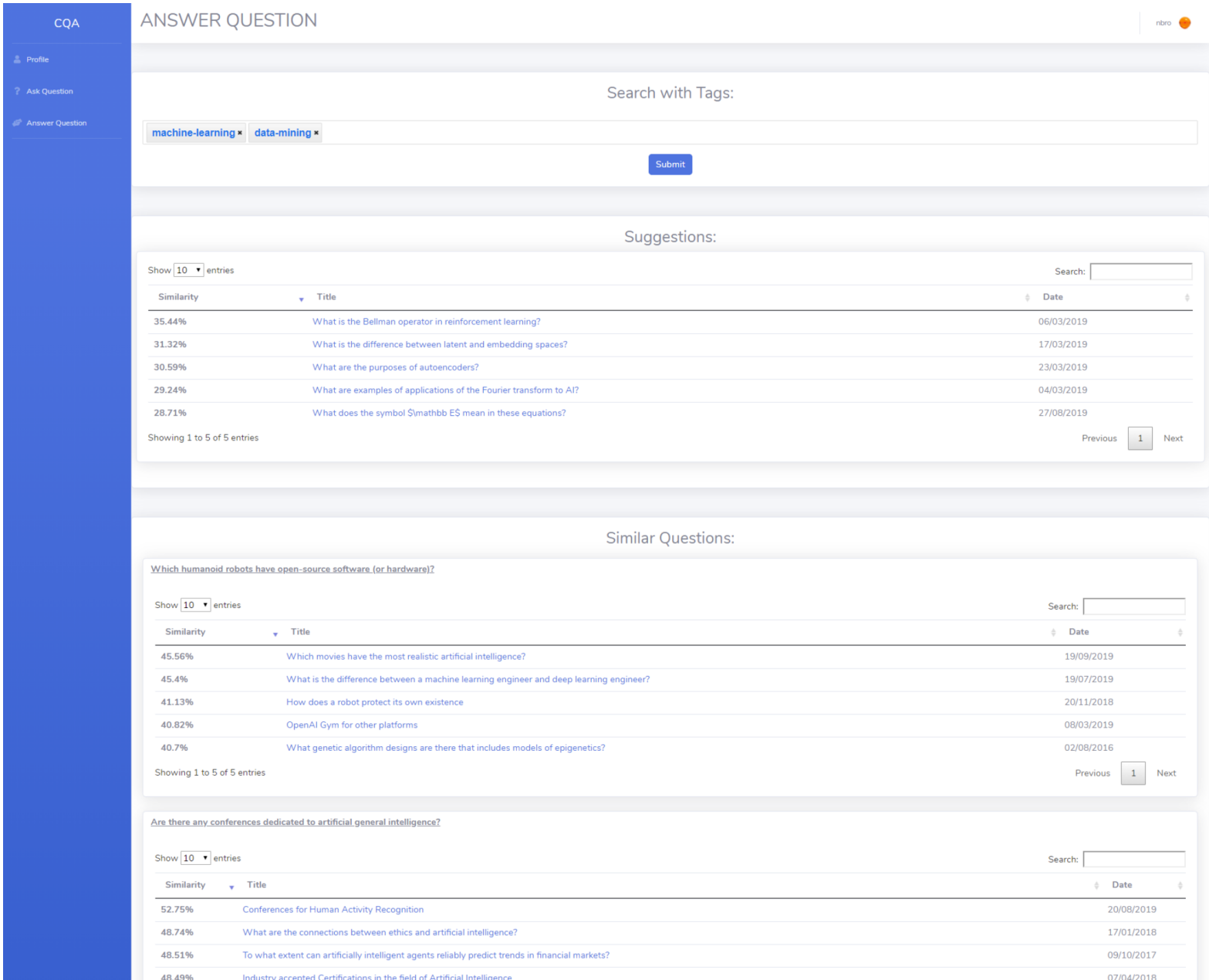
Front-End


CQA
Profile
Ask Question
Answer Question
RESULTS
nbro
Your question may get an answer with the probability of 39.87% within 5 - 24 hours.
Show 10 entries
Search:
Type
Similarity
Title
Date
Unverified Answers 85.66% What is the difference between artificial intelligence and machine learning? 02/08/2016
Verified Answer 52.46% What is machine learning? 27/05/2019
Verified Answer 49.2% What are the real-life applications of Transfer Learning in Machine Learning? 05/10/2019
Verified Answer 47.35% Equilateral and One-of-n encoding 16/03/2018
Unverified Answers 46.07% What is the use of softmax function in a CNN? 16/06/2019
Not Answered 45.98% Difference between simple reflex and model bases reflex agent 01/11/2018
Unverified Answers 45.84% What is the fundamental difference between CNN and RNN? 08/12/2017
Not Answered 45.76% Is the Assumption-based Truth Maintenance System still used? 05/11/2019
Verified Answer 45.39% What are the real world uses for SAT solvers? 02/08/2016
Not Answered 44.98% Why would the application of boosting prevent underfitting 14/11/2019
Type
All
Showing 1 to 10 of 30 entries
Previous 1 2 3 Next
Copyright © Sabanci University 2019

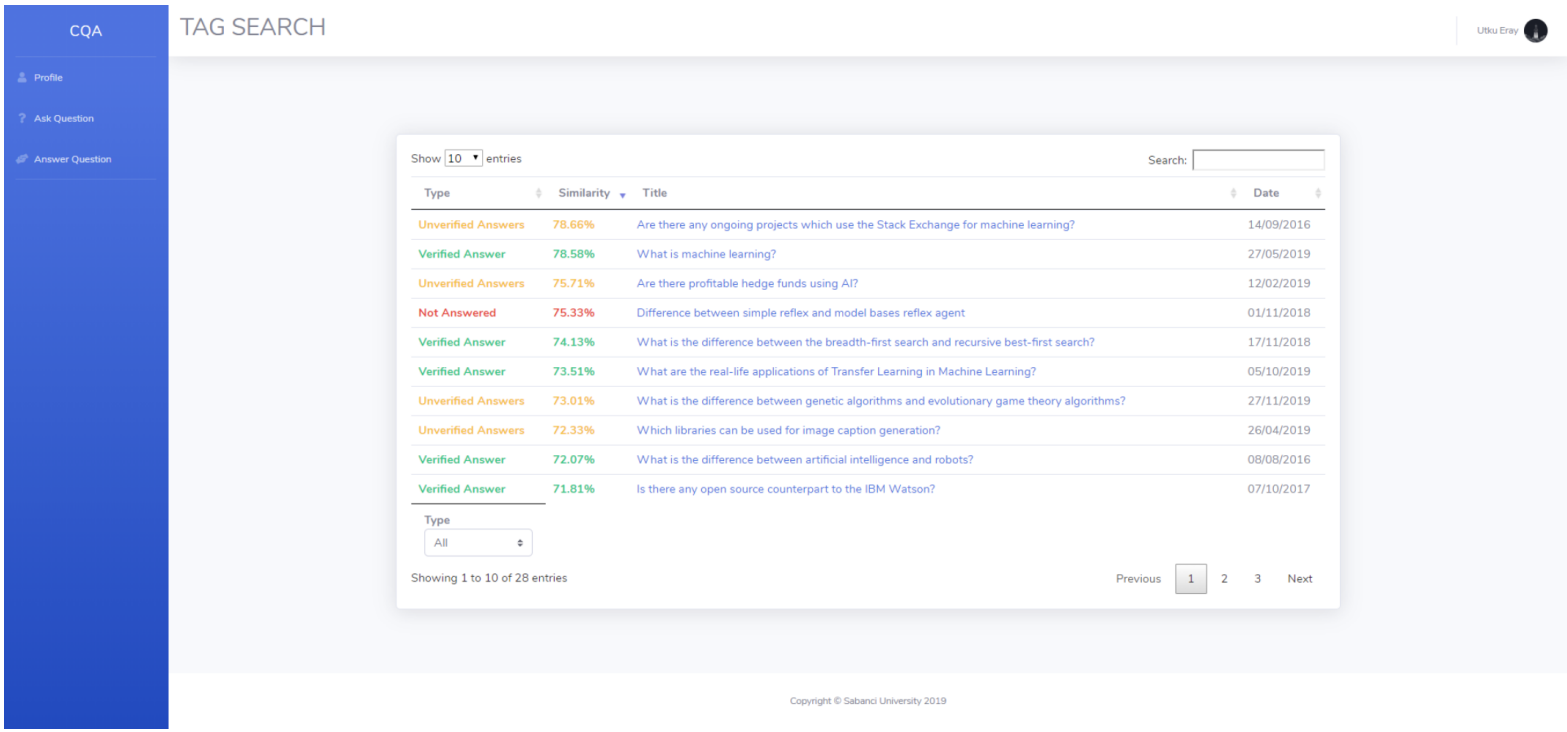
Front-End


CQA
ANSWER QUESTION
Utku Eray
Profile
Ask Question
Answer Question
Search with Tags:
Submit
Copyright © Sabanci University 2019

Front-End


CQA
Profile
Ask Question
Answer Question
ANSWER QUESTION
nbro
Search with Tags:
machine-learning ×
data-mining ×
Submit
Suggestions:
Show 10 entries
Search:
Similarity
Title
Date
35.44% What is the Bellman operator in reinforcement learning? 06/03/2019
31.32% What is the difference between latent and embedding spaces? 17/03/2019
30.59% What are the purposes of autoencoders? 23/03/2019
29.24% What are examples of applications of the Fourier transform to AI? 04/03/2019
28.71% What does the symbol $\mathbb E$ mean in these equations? 27/08/2019
Showing 1 to 5 of 5 entries
Previous 1 Next
Similar Questions:
Which humanoid robots have open-source software (or hardware)?
Show 10 entries
Search:
Similarity
Title
Date
45.56% Which movies have the most realistic artificial intelligence? 19/09/2019
45.4% What is the difference between a machine learning engineer and deep learning engineer? 19/07/2019
41.13% How does a robot protect its own existence 20/11/2018
40.82% OpenAI Gym for other platforms 08/03/2019
40.7% What genetic algorithm designs are there that includes models of epigenetics? 02/08/2016
Showing 1 to 5 of 5 entries
Previous 1 Next
Are there any conferences dedicated to artificial general intelligence?
Show 10 entries
Search:
Similarity
Title
Date
52.75% Conferences for Human Activity Recognition 20/08/2019
48.74% What are the connections between ethics and artificial intelligence? 17/01/2018
48.51% To what extent can artificially intelligent agents reliably predict trends in financial markets? 09/10/2017
48.49% Industry accepted Certifications in the field of Artificial Intelligence 07/04/2018

Front-End


CQA
Profile
Ask Question
Answer Question
TAG SEARCH
Utku Eray
Show 10 entries
Search:
Type
Similarity
Title
Date
Unverified Answers 78.66% Are there any ongoing projects which use the Stack Exchange for machine learning? 14/09/2016
Verified Answer 78.58% What is machine learning? 27/05/2019
Unverified Answers 75.71% Are there profitable hedge funds using AI? 12/02/2019
Not Answered 75.33% Difference between simple reflex and model bases reflex agent 01/11/2018
Verified Answer 74.13% What is the difference between the breadth-first search and recursive best-first search? 17/11/2018
Verified Answer 73.51% What are the real-life applications of Transfer Learning in Machine Learning? 05/10/2019
Unverified Answers 73.01% What is the difference between genetic algorithms and evolutionary game theory algorithms? 27/11/2019
Unverified Answers 72.33% Which libraries can be used for image caption generation? 26/04/2019
Verified Answer 72.07% What is the difference between artificial intelligence and robots? 08/08/2016
Verified Answer 71.81% Is there any open source counterpart to the IBM Watson? 07/10/2017
Type
All
Showing 1 to 10 of 28 entries
Previous 1 2 3 Next
Copyright © Sabanci University 2019

Conclusion & Future Work

This project aimed to improve Stack Exchange sites in terms of user utilization. The improvements have been implemented with a user-friendly website for ease of use.

For the future, other machine learning tasks can be defined such as prediction on the answer quality, prediction on whether a question will be closed...

References

- Le, Q., & Mikolov, T. (2013). Distributed Representations of Sentences and Documents. Retrieved from https://arxiv.org/pdf/1405.4053.pdf
- Gensim: Topic modelling for humans. Radim Řehůřek: Machine learning consulting. https://radimrehurek.com/gensim/
- Scikit-learn. scikit-learn: machine learning in Python — scikit-learn 0.16.1 documentation. Retrieved from https://scikit-learn.org/stable/
- Stack Exchange. 2019. Retrieved from https://stackexchange.com/sites

Thank You for Listening!