AmEx Ignite Proposal

Section A: Personal Details

Name: Akash Mondal **Institute:** Indian Institute of Technology Kharagpur

Roll No: 16CE33012 **Year:** 4th **Branch:** Civil Engineering

Degree: Dual Degree(B.Tech+M.Tech)

Prior background in machine learning/artificial intelligence:

Internships:

1. CodeFire Technologies | Summer 2018:

- Created an IR-based factoid question-answering system (chatbot), for answering queries related to school admission.
- Predicted transportation rate with accuracy nearly 90% in a small dataset and also Visualised feature importance with Tableau.
- Developed a tool to extract less structured web content from noisy web pages that appear only once like Headline, Author Name, etc.

2. National AI Resource Portal | Centre for AI IIT Kharagpur | Summer 2019:

- Implemented a screen scraper, capable of extracting and organizing data from various public ml repositories.
- Developed a tool in Python to automate the Indexing and the mapping process of the scraped data to the predefined metadata ontology.
- Used Apache Solr to create the search platform by indexing all the scraped data.

Projects:

1. News Classifier(Project):

- Used Tensorflow and NLTK to build a news classifier that classifies a given news headline to one of the many labels like politics, sports, etc.
- Augmented word vectors by applying Bag of Words model(BoW) on the collected text corpus to train the DNN Tensorflow model.
- Trained on a dataset containing various news headlines collected by scraping major Indian news websites using the newspaper library.

2. CNN For Car-Recognition(Project):

- Used Keras to implement a simple convolutional neural network car recognition system that achieves over 90% accuracy on test data
- Consists Conv layer, ReLU activation, Pooling layer two each and used augmentation techniques to increase the accuracy to 98%
- Used the Flask web framework to develop a keras and deep learning REST API and tested the server with both cURL and Python

Courses undertaken:

- Machine Learning, Programming and Data Structures,
- Probability and Statistics, Algorithm-l

Competitions:

- Flipkart GRiD Te[a]ch The Machines 2019(machine learning challenge)
- Microsoft code.fun.do (Hackathon 2018)

Section B: Proposal

(i) Describe the project you wish to work on. How did you come up with the idea?

Domain-based question answering system

The goal of this Information Retrieval based question answering system is to answer a user's question(In specific domain) by finding a text segment on some collection of labeled documents with different entities. The system will be flexible to different domains. It will be able to answer and be the Business Matter Expert, in the specific service it's providing, and potentially the specific domain it is servicing the business customer on. It can be used to build interfaces for websites, mobile applications, popular messaging platforms, and IoT devices that enable natural interactions between a business and it's users. Advantages over other chatbots:

1. It is robust – it relies only on a small number of syntax-analysis tools
2. It is easy to customize or extend with new features or knowledge.
3. It provides answers in real-time due to its simple architecture based on a small number of multi-class classifiers.

(ii) Provide a detailed project plan with project milestones and estimated timelines.

The proposed system uses a typical architecture consisting of three components linked sequentially. These are:

- **Question processing:** The question processing phase detects the type and extracts the number of pieces of information from the of the input questions(query). We can use NLTK to get all the entities (person, location, time, etc.) of the user query or we can implement a machine learning classifiers to serve this purpose.
- **Passage retrieval:** The second phase i.e. to get the relevant passage first we will convert the sentences to vector(bags of words model) and further it will be trained in a dataset(queries with labelled query type e.g. if it is asking about any person or is it asking about some places etc.) to recognize the user's intent and extracts prebuilt entities such as time, date, number etc.
- **Answer processing:** The Answer Extraction (AE) component identifies candidate answers from the relevant passage set and extracts the answer(s) most likely to respond to the user question. Here, we will convert the passage into a list of answer set than we assigned each answer a similarity score and for the final answer, we will return the sentence in the passage with the best score.

Timeline: each phase will take up to one or two weeks each(estimated), more couples of the week to increase the accuracy of the model, finding possible alternatives if performing poorly. As I have completed the first phase so I am expecting to complete it in 2-3 months.

(iii) Are you starting this project from scratch or is it an on-going project? If it is a work in progress, provide more details.

I have coded the phase-I i.e. question processing phase, using Python NLTK by first tokenizing the query and finally by removing the morphological affixes from words, leaving only the word stem by LancasterStemmer. I am yet to start the rest of the phases i.e. passage retrieval and ranking and answer processing phase.

(iv) Would you require financial funding for access to new data sources or computation? If yes, please elaborate on the requirement clearly stating the purpose.

Financial funding may require if we want to build a web interface of our system like for hosting the UI, backend and cloud credentials for training the model and to store the labeled documents for extracting answer, etc.

(v) Describe the techniques you intend to explore to accomplish your project. Also, specify details of how you plan to obtain data to train/test the algorithm.

I want to explore more about the way of representing text data when modeling text with machine learning algorithms. Though I have used Bags Of Words model in my previous internships I want to know if other existing models are better fit like SVD, word2vec, or GloVe, etc. Also, I want to explore how to priorities different sentences based on their word's POS tags and lexical context.

(vi) What kind of help do you expect from the Mentor who will be guiding you? (The mentor will be an experienced data science expert, working at AmEx)

I have a raw idea about the model. I am expecting help like in selecting the model and how to tune them for different domains.

(vii) What should be the success metrics on which your project should be evaluated?

By measuring task success rates. Task success is a major category for chatbot matrics. It will measure the percentage of query answered by the system vs, the query that had to be escalated to some person. It will measure the error rate within the process(answering queries) itself. In another word By tracking: Did it execute on the query to the satisfaction of the user or did a person have to answer it? Was there any error throughout the process that went through? Did it flow smoothly and was the intent understood?