Scalable Knowledge Sharing Platform; An Artificial Intelligence Approach

by: Ibrahim O. Odufowora

Outlines

- Problem Description
- Understanding the data set
- Solution Methodology & Implementation
  - Data preprocessing
  - Model building & Parameter tuning
  - Model Validation
- Environment Setup
  - Development
  - Production
- References

Problem Description

- Identification of duplicate questions.
- A page for each logically distinct question.
- Identical Question:
    - “What are some of the best romantic movies in English?”
    - “What is the best romantic movie you have ever seen?”
- Non Identical Question:
    - “When do you use シ instead of し? “
    - “When do you use "&" instead of "and"?”
- A ‘scalable question platform’ should provide answers to each distinct question in a single resource location. This will enhance proper resources utilization.

Understanding The Data Set

id	qid1	qid2	question1	question2	is_duplicate
447	895	896	What are natural numbers?	What is a least natural number?	0
1518	3037	3038	Which pizzas are the most popularly ordered pizzas on Domino's menu?	How many calories does a Dominos pizza have?	0
3272	6542	6543	How do you start a bakery?	How can one start a bakery business?	1
3362	6722	6723	Should I learn python or Java first?	If I had to choose between learning Java and Python, what should I choose to learn first?	1

- No. Variables: 6
- Relevant Input Features: 2 (question1, question2)
- Output Feature: 1 (is_duplicate)
- No. Train samples: 404,290
- No. Test samples: 2,345,796

Solution Methodology & Implementation

- Methodology:
  - Automate the process of detecting semantically equivalent questions.
  - Direct a user to an existing resource/page if the question already exist.
- Implementation:
  - Build effective and scalable machine learning and natural learning processing models.
  - Data preprocessing.
  - Models building & Parameters tuning.
  - Models Validation.
  - Interface a scalable, online real-time RESTful API with the best model.
  - Load balance the API in order to minimize response time and to avoid overload.

Environment Setup

- Development
  - Build the models on the train set.
  - Use test set for models validation.
  - Interface a Flask RESTful API with the best model.
- Production
  - Move the best model and its optimal parameters to production.
  - Interface a RESTful API with WSGI server with the model parameters.
  - Load balance the API.
  - As more data suffaces, rebuild the model occasionally to improve predictability.

References

- Quora. 2017. *First Quora Dataset Release: Question Pairs*. [ONLINE] Available at: https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs. [Accessed 30 September 2018].