A questing on pair's similarity detection with data mining applications using Natural language processing and machine learning: QUORA

[1]**Dr.Lalitha kumari pilli**, Assoc.Prof. dept of CSE, Malla Reddy Institute of Technology, Kompally, Maisammaguda, Dulapally, Secunderabad, 500100
[2]**P.Dhana Sri**, Asst. Prof., dept of CSE, Malla Reddy Institute of Technology, Kompally, Maisammaguda, Dulapally, Secunderabad, 500100
[3]**Dr.G.Nanda Kishor Kumar**, Professor, dept of CSE, Malla Reddy Institute of Technology, Kompally, Maisammaguda, Dulapally, Secunderabad, 500100

ABSTRACT

Processing and synthesising relevant information from a vast volume of unstructured data is known as "data mining." Here, we'll refer to the process of extracting meaningful information from massive data sets as "ontology" instead of "data mining," choosing an alternate name. Semantic web research has undergone an upswing in the number of creative advancements in recent years. The question-and-answer database of the semantic web may be used to find the most regularly recurring question-and-answer pairs on a specific subject. This work is growing increasingly harder for developers to prove, hence I've meant my current effort to be chock-full of crucial details. We'll employ QUORA, a platform where users may publish and amend questions and answers. We'll utilize this to assess the present plan. This tool enables users to work cooperatively on particular topics, give comments, and change previously published responses. This form of cooperation is done as a thread on a particular topic with a list of related or similar questions in order to discourage people from answering the same questions over and over again. Natural Language Processing (NLP) principles and multiple machine learning (ML) methodologies are used to find and eliminate the most varied queries and replies from a dataset for one particular subject matter.

Key Words: Ontology, Machine Learning, Natural Language Processioning, QUORA, Distinct Questions, Duplicate Questions, Semantic Web, Research Inventions.

1. INTRODUCTION

The amount of text on the internet has increased dramatically in recent years. That's why it's necessary to have certain top-notch algorithms that outperform other algorithms in the data repository if we want to manage and analyse meaningful information. Despite popular belief, mining isn't the sole method for locating and extracting information from structured or relational data.The amount of text on the internet has skyrocketed recently. Thus, in order to manage and process useful information, we need some excellent algorithms that can outperform other algorithms in order to obtain all the essential information from the data repositories. A common misconception is that only mining may be used to extract information from structured or relational data, although this is not always the case. Many different ways can be used to get text information from data that is structured or has questions and answers or long paragraphs with data that is linked.

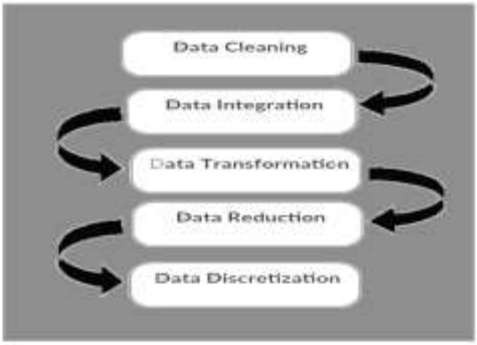


Figure 1. Representation the Data Pre - Processing Steps

Data pre-processing is considered a crucial step in every data mining application in order to translate high-level instructions into machine instructions. Data that has not been preprocessed will be exceedingly inconsistent, and the model won't be able to provide precise findings.

2. LITERATURE SURVEY

The most important phase in the software development process is the literature review. Prior to beginning the development of a new application or model, the time factor, economy and corporate strength must all be considered? We may begin constructing the application after all of these variables are reviewed and approved.

MOTIVATION

[1]. Text-to-Speech Dialogue Generation Using Discourse Structure Analysis.

H. Prendinger is the author.

For the most part, this article tries to focus on how to categorise input such as questions and replies collected from a text or passage. Generally speaking, there are two ways to categorise this text: Assessment of educational attainment using dialogue and an interactive Q&A system.

[2]. Interactional Question-Answering Experiments"

D. Moldovan is the author of this work.

An interactive question-and-answer system for educational purposes is the primary focus of this study. By generating questions and answers depending on the preferences of each unique user, the authors want to compare their work with that of another article that aims to first discover domain knowledge and then determine the context of that domain.

[3]. An Aid to Independent Study SIGCUE Outlook: Automatic Question Generation from Text.

J. H. Wolf is one of the authors.

The problem of producing questions and answers for a specific context was highlighted in this suggested study. According to the author's analogy, creating an exam paper is a highly difficult process that requires a lot of time and effort, therefore instructors have to put in quite a bit of effort in order to provide adequate MCQs for the educational system.

[4]. A Study of Text Classification Preprocessing Methods

BY Ammar Kadhim and colleagues

For text classification applications, preprocessing of text data is critical, as the author discusses in great detail in this draught paper. Pre-processing involves reducing a large number of words into one word and extracting the most important information from that word.

3. DATABASE SUGGESTED FOR USE

We want to leverage the Quora dataset, which is freely available online. This dataset contains more than 4 lakhs question pairs for training the model, and a testing dataset of 23 lac question pairs, which may be used to evaluate the model. The following are a few of the dataset's most important characteristics.

The dataset includes the following fields:

a. The training set question pair's id

b. Qid1 and Qid2 are the unique identifiers of each question, respectively (only available in train.csv)

c. full text of each of the following questions

d. Is duplicate - the target variable, set to true

Let's take a look at some instances of question pairings that are similar and different.

EXACTLY THE SAME COMBINATIONS OF QUESTIONS

A decent geologist is hard to come by.

What do I need to do to become an excellent geologist? '

QUESTION PAIRS WITH DIFFERENT ANSWERS

What is the step-by-step procedure for investing in the Indian stock market?

Investing in the stock market requires a step-by-step guidance.

4. DESCRIPTION OF THE PROPOSED METHOD

In this part, we'll talk about some of the methods we've considered for evaluating our present objective's efficiency. A two-phase technique is being used to test the efficacy of our existing application:

1. NLP-Based Feature Extraction

The characteristics may be extracted in two ways in this NLP:

a. The Minimum Requirements

b. The Feature Set Isn't Clear

2. Using a machine learning model for similarity prediction

At this stage, we compare the present application with a variety of machine learning models, then evaluate the performance of each model to determine which model provides the most accurate results. XGBOOST provides the most accuracy in our present application when it comes to predicting the optimal similarity query and pair.

1. Extraction of Feature Sets Using NLP

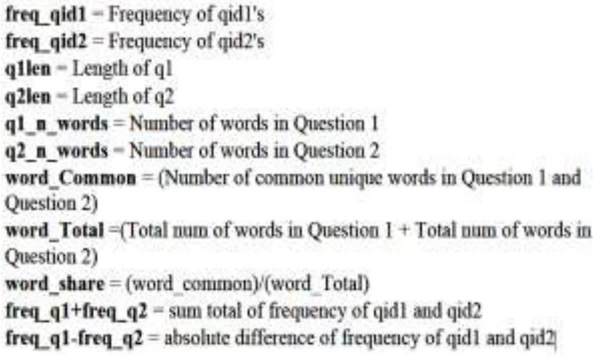
There are two techniques to extract features in this NLP:

A. The bare-bones features.

B. The Set of Fuzzy Feature

A) The bare-bones features.

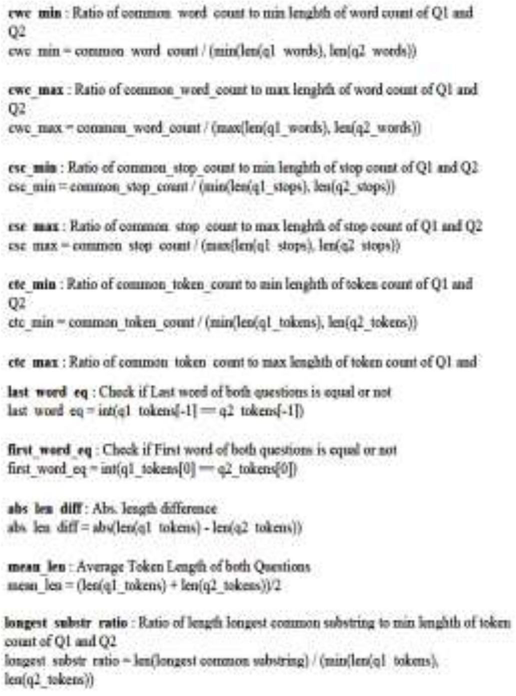
For starters, here are a few things to look for in a feature set:



freq_qid1 = Frequency of qid1's
freq_qid2 = Frequency of qid2's
q1len = Length of q1
q2len = Length of q2
q1_n_words = Number of words in Question 1
q2_n_words = Number of words in Question 2
word_Common = (Number of common unique words in Question 1 and Question 2)
word_Total =(Total num of words in Question 1 + Total num of words in Question 2)
word_share = (word_common)/(word_Total)
freq_q1+freq_q2 = sum total of frequency of qid1 and qid2
freq_q1-freq_q2 = absolute difference of frequency of qid1 and qid2

As we can see from the above characteristics list, roughly 11 unique features are present. Word Total and Word Common are two of the most important characteristics in the standard feature set, however there are many more.

B) Fuzzy Feature Set



cwc_min : Ratio of common_word_count to min lenghth of word count of Q1 and Q2
cwc_min = common_word_count / (min(len(q1_words), len(q2_words))

cwc_max : Ratio of common_word_count to max lenghth of word count of Q1 and Q2
cwc_max = common_word_count / (max(len(q1_words), len(q2_words))

csc_min : Ratio of common_stop_count to min lenghth of stop count of Q1 and Q2
csc_min = common_stop_count / (min(len(q1_stops), len(q2_stops))

csc_max : Ratio of common_stop_count to max lenghth of stop count of Q1 and Q2
csc_max = common_stop_count / (max(len(q1_stops), len(q2_stops))

ctc_min : Ratio of common_token_count to min lenghth of token count of Q1 and Q2
ctc_min = common_token_count / (min(len(q1_tokens), len(q2_tokens))

ctc_max : Ratio of common_token_count to max lenghth of token count of Q1 and

last_word_eq : Check if Last word of both questions is equal or not
last_word_eq = int(q1_tokens[-1] == q2_tokens[-1])

first_word_eq : Check if First word of both questions is equal or not
first_word_eq = int(q1_tokens[0] == q2_tokens[0])

abs_len_diff : Abs. length difference
abs_len_diff = abs(len(q1_tokens) - len(q2_tokens))

mean_len : Average Token Length of both Questions
mean_len = (len(q1_tokens) + len(q2_tokens))/2

longest_substr_ratio : Ratio of length longest common substring to min lenghth of token count of Q1 and Q2
longest_substr_ratio = len(longest common substring) / (min(len(q1_tokens), len(q2_tokens))

From the above fuzzy features set we can see there are nearly 11 distinct features present in that basic feature set. Each and every feature is having distinct importance and they are used to calculate the most common question answer pair's similarity.

2) Using a Machine Learning Model to Predict Similarity.

Xgboost, a gradient boosting ensemble machine learning method based on a decision tree, is being used in this application. To surpass all other algorithms or frameworks, this Xgboost focuses on extracting unstructured data (pictures, text, etc.). When it comes to small-to-medium-sized structured/tabular data, this decision tree is deemed the best algorithm by the majority of users.

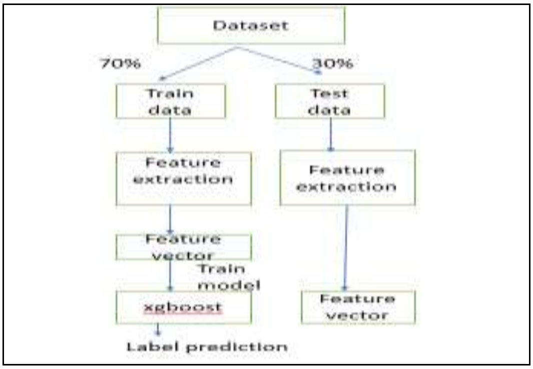


Figure 2 depicts our proposed model's architecture.

STEP WISE EXPLANTION

Step-by-step instructions for our current model may be found below This goes like this:

The first step is to import the input dataset, which includes a lot of relevant information about a certain subject. The dataset used for this analysis is the Quora dataset, which consists of an unstructured collection of question and answer pairings on a single subject.

Before dividing the dataset into training and testing, a pre-processing approach is needed. 70 percent of the dataset is used for training purposes, while 30 percent of the dataset is used for testing purposes.

A feature vector is subsequently constructed from the input data in Step 3, which includes features gleaned from both the training data and the test data.

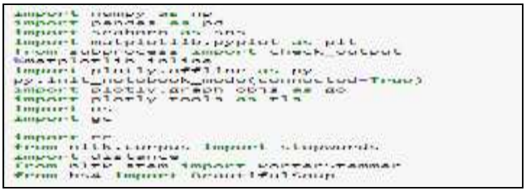
Our next step is to determine the most important attributes and then train our system on them.

XgBoost ML Model is used in the fourth step and then queries are categorised as distinct and duplicates. For each question, we attempt to label it and then categorise the best desired result based on that result.

4. EXPERIMENTAL RESULTS

A Python-based implementation is being used to demonstrate the planned application's performance. The first step is to import all the required libraries and then load the input dataset to discover the most accurate similarity between question and answer pairs.

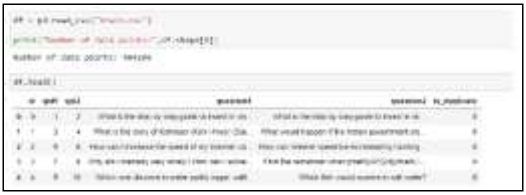
IMPORT LIBRARIES



```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from subprocess import check_output
%matplotlib inline
import plotly.offline as py
py.init_notebook_mode(connected=True)
import plotly.graph_objs as go
import plotly.tools as tls
import os
import gc

import re
from nltk.corpus import stopwords
import distance
from nltk.stem import PorterStemmer
from bs4 import BeautifulSoup
```

We can see from the window above that a number of libraries and packages are being utilised to demonstrate the present goal. We thus attempt to load and import all of the relevant libraries into our application.

LOAD INPUT DATASET



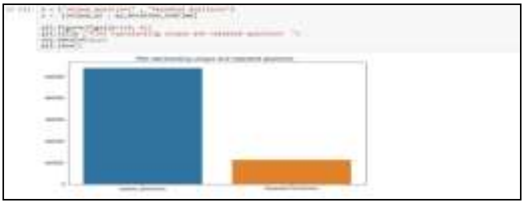
```
df = pd.read_csv("train.csv")

print("Number of data points:",df.shape[0])

Number of data points: 404290

df.head()
```

```
   id  qid1  qid2                                          question1                                          question2  is_duplicate
0   0     1     2  What is the step by step guide to invest in sh...  What is the step by step guide to invest in sh...             0
1   1     3     4  What is the story of Kohinoor (Koh-i-Noor) Dia...  What would happen if the Indian government sto...             0
2   2     5     6  How can I increase the speed of my internet co...  How can Internet speed be increased by hacking...             0
3   3     7     8  Why am I mentally very lonely? How can I solve...  Find the remainder when [math]23^{24}[/math] i...             0
4   4     9    10  Which one dissolve in water quikly sugar, salt...            Which fish would survive in salt water?             0
```

We can clearly see crucial variables like data points, the amount of data points, and required question pairs and their accompanying results whether they are distinct or repeated in the aforementioned window. For this duplicate property, we may use either

0 or 1 as the value. A '1' is recorded if the questions and responses are identical. In this case, '0' is used to indicate that they have not been copied.

DATA VISUALIZATION

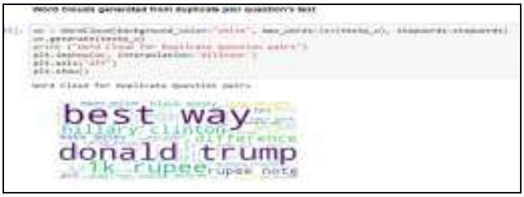


Two key criteria, such as unique questions and duplicated questions, may be seen clearly in the above window.

DATA PRE - PROCESSING

Several pre-processing techniques may be easily identified from the above window when features are extracted from the input data.

Word Cloud Formed by Duplicate Pair Question



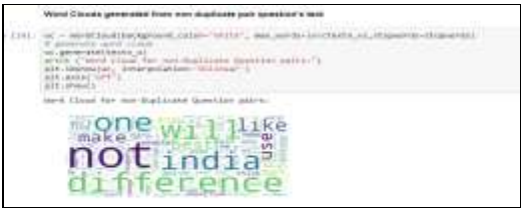
Word Clouds generated from duplicate pair question's text

```
wc = WordCloud(background_color="white", max_words=len(texts_d), stopwords=stopwords)
wc.generate(texts_d)
print ("Word Cloud for Duplicate Question pairs")
plt.imshow(wc, interpolation='bilinear')
plt.axis("off")
plt.show()
```

Word Cloud for Duplicate Question pairs

We can plainly see numerous terms that are retrieved from the majority of duplicated question pairings and so they create a single word cloud from the questions that are most often repeated.

Word Cloud Formed by Non - Duplicate Pair Question



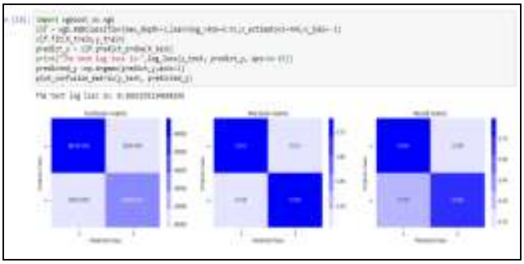
Word Clouds generated from non duplicate pair question's text

```
wc = WordCloud(background_color="white", max_words=len(textn_w), stopwords=stopwords)
# generate word cloud
wc.generate(textn_w)
print ("Word Cloud for non-Duplicate Question pairs:")
plt.imshow(wc, interpolation='bilinear')
plt.axis("off")
plt.show()
```

```
Word Cloud for non-Duplicate Question pairs:
```

There are various terms from non-duplicated pairings that appear often in the word cloud shown in the top window and hence they are all grouped together.

XgBoost Performance



```
import xgboost as xgb
clf = xgb.XGBClassifier(max_depth=[illegible], learning_rate=[illegible], n_estimators=[illegible], n_jobs=-1)
clf.fit(X_train,y_train)
predict_y = clf.predict_proba(X_test)
print("The test log loss is:",log_loss(y_test, predict_y, eps=1e-15))
predicted_y =np.argmax(predict_y,axis=1)
plot_confusion_matrix(y_test, predicted_y)
```

```
The test log loss is: 0.[illegible]
```

The XgBoost ML method generates a confusion matrix for the provided dataset, as seen in the upper window. Confusion Matrix, Precision Matrix, and Recall Matrix are all examples of performance charts.

5. CONCLUSSION

An entirely new approach for determining the most frequently asked questions and answers on Quora has been devised in this proposed work. I've tried a wide array of machine learning models to handle the issue of duplicate questions on Quora. My finest results were from XGBoost Model after multiple theoretical and practical studies. Aside from the Quora dataset itself, I think that machine learning methods may be used to further investigate the problem of Natural Language Understanding. As such, my study on machine learning techniques will undoubtedly include more models and efforts to improve existing models.

6. REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Proceedings of the International Conference on Learning Representations (ICLR).

[2] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in neural information processing systems, 2013.

[3] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment analysis is a big suitcase," IEEE Intelligent Systems, vol. 32, no. 6, pp. 74–80, 2017T.

[4] Eric Jones, Travis Oliphant, Pearu Peterson, et al. 2001–. SciPy: Open source scientific tools for Python. [Online; accessed ¡today¿]. http://www.scipy.org/

[5] Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Proc. ICLR

[6] Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016a. How transferable are neural networks in nlp applications? In Proc. EMNLP.

[7] Antonie, M. L., & Zaiane, O. R. (2002). Text document categorization by term association. In Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on (pp. 19-26). IEEE.

[8] Srividhya, V., & Anitha, R. (2010). Evaluating preprocessing techniques in text categorization. International journal of computer science and application, 47(11), 49-51.

[9] Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. Journal of machine learning research, 3(Mar), 1289-1305.

[10] Soucy, P., & Mineau, G. W. (2005, July). Beyond TFIDF weighting for text categorization in the vector space model.

[11] Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. WSEAS transactions on computers, 4(8), 966-974.

[12] Kamruzzaman, S. M., Haider, F., & Hasan, A. R. (2010). Text classification using data mining. arXiv preprint arXiv:1009.4987.

[13] 13) Shi, L., Mihalcea, R., & Tian, M. (2010, October). Cross language text learning. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (pp. 1057-1067). Association for Computational Linguistics.

[14] Kadhim, A. I., Cheah, Y. N., & Ahamed, N. H. (2014, December). Text Document Preprocessing and Dimension Reduction Techniques for text.