Determining Category

Hossein Zandinejad

Abstract

Document Classification or Document Categorization is a problem in Library Science, Information Science, and Computer Science. The task is to assign a document to one or more classes or categories. This may be done "***manually***" (or "***intellectually***") or ***algorithmically***. The intellectual classification of documents has mostly been the province of Library Science, while the algorithmic classification of documents is mainly in Information Science and Computer Science. The problems are overlapping, however, and there is therefore interdisciplinary research on document classification. The documents to be classified may be ***texts***, ***images***, ***music***, etc. Here, in this project, we are going to do text classification.

1 Introduction

- The goal of this project is to determine the category of each application using Persian description. Text mining technology is now broadly applied to a wide variety of government, research, and business needs. All these groups may use text mining for records management and searching documents relevant to their daily activities.
- Here, I first describe the data I use and how I process it. Then, I explore whether these questions can be answered using techniques for classification and text mining. Finally, I run multiple Machine Learning and Deep Learning models and choose the most accurate one.
- In this project the difficulty was to find appropriate parameters for the Neural Network model because of the complexity of the dataset.
  In this paper, I examine several models and attempt to find the most efficient one.

2 The dataset and features

For this project I use the SummerCamp 1400 CafeBazaar dataset. This dataset consists of three features named "***app_id***", "***description_fa***" and "***label***". The "***app_id***" feature shows the ID of each 37899 application which is not of obvious importance in this project, With that being said, I deleted it and used the other features in constructing the models. The "***description_fa***" feature is a Persian description of the apps and games that is a string. The "label" feature is a label for each app that shows the category of it.

I use ***Fasttex*** in Python which is a library for efficient learning of word representation and sentence classification. The pre-trained word vector for Persian Language, trained on Common Crawl and Wikipedia has been used for this project. Then with the help of some functions (sentence_to_avg, read_fastext_vecs) and a dictionary that maps words to their fasttext vector representation, I prepared the dataset for NN-model. I also prepared the labels for training in Keras Library that I used.

	app_id	description_fa	label
0	0	<p>بازی مین روب یک برنامه فکری است که باید مین...	1
1	1	<p>در این بازی تعدادی عکس برای شما نشان داده می...	1
2	2	<p>مرگ پایا «تلاش نافرجام برای درک «بوفالو ...	7
3	3	<p>فیلم نما ، برنامه ای برای دانلود و پخش آنلا...	7
4	4	<p>*این برنامه حاوی بیش از 500 عکس و ژست برای ...	7

3 Methods

I. Neural Network: [1] [3] For the purposes of my project, I used Python's ***Sklearn***, ***Keras***, ***TensorFlow***, ***Pandas***, and ***NumPy*** packages to construct a neural network model on the test dataset. I achieved 708 points twice (%70.8), using NN-model with the following parameters:
$1.Batch\_size = 64$,
$epochs = 200$,
$loss\_function =' mean\_squered\_error'$
$, optimizer = adagrad(lr = 0.001)$
$2.Batch\_size = 370$,
$epochs = 40$,
$loss\_function =' mean\_squered\_error'$,
$optimizer =' adam'$

II. Linear SVC: [2] A support vector machine builds a model that seeks to maximize the margin between the separating hyperplane and data points. More specifically, the model parameters are found by solving the following optimization problem:

$$\min \frac{1}{2}|||\omega|||^2 + C\sum_{i=1}^{m}\xi_i$$
$$\text{s.t.} \quad y^{(i)}(\omega^T \mathbf{x}^{(i)} + b) > 1 - \xi_i, \quad i = 1, \ldots, m$$
$$\xi_i \geq 0$$

To implement SVM we used LinearSVC in the Sklearn package in Python. I achieved 754 points (%75.4), using a ***linear SVC*** model with the following parameters:
$C = 0.5$,
$multi_class = "crammer\_singer$,
$max\_iter = 5000"$

4 Future Works

Given more time I would work to develop or improve on several ideas. The first idea I had was to Bayesian GaussianMixture model and the second one was to run Logistic Regression Models.
The third method I wanted to try was to add persian_stop_words to our models but unfortunately, I did not have the time and a suitable stop word list.

5 Conclusion

I did a pre-process first like vectoring the words by the use of ***fasttext*** library. The next step was to try different models to check which one had the best performance. The ***LinearSVC*** model was the best of all in categorizing the apps and games with an accuracy of ***%75.4***.

References

[1] Charu C. Aggarwal. *Neural Network and Deep Learning*. Springer, 2018.

[2] Sklearn. Linearsvc.

[3] Sklearn. Neural network models.