TEST Engineering & Management

Information Retrieval and Document Classification

Alisha Gupta [2]

[1] Symbiosis Institute of Business Management Pune; Symbiosis International (Deemed University) Pune

Article Info
Volume 82
Page Number: 5886-5892
Publication Issue:
May - June 2020

Article History
Article Received: *19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *17 May 2020*

Abstract

There appears to be various information available online in the form of document. Finding these kinds of documents and retaining them, corresponding to their category has never been more automatic. This paper acknowledges the issue of classifying genre of different English novels with the help of different Natural Language Processing and Machine Learning methods. Different novels are collected and divided into training Dataset and test Dataset. Originally for the purpose of classification uses three dissimilar varieties of Fiction genre specifically Romantic, Fairy Tales and Thriller. The genres that have been taken are some of the most widely read genres of book among different age groups. Using different linguistic feature to obtain representative features for the genres. The training module uses the feature Datasets to provide the base for classification feature.

Keywords; *Classification, Dataset, Genre*

I. INTRODUCTION

Document Classification or Document Categorization is the act of classifying documents into various categories as an example news articles, researches, sports articles etc. The explosive increase in the resources of structured and unstructured information in the form of news articles, digital libraries, blog repositories etc. has made document classification an important area of research. It is practically not possible for the human beings to go through all of the available documents to find the document of interest. So, from this, the concept of document classification rose. Document classification assigns the documents to one or more predefined categories as specified by the text contained in them. Various text processing techniques like Data Mining, Machine Learning and Natural Language Processing to work together to implement text classification.

This paper intended towards digital documents and in particular, collection of digital novels and books. The technique of document classification has been used to categorize the English novels into three different form of Fiction genre namely, Romance, Fairy Tales and Thriller.The implementation is based on Python module and NLTK (Natural Language Toolkit) module as the major blocks. Python is an integrated software application of wide complexity, domain specific, object-focused.

Natural Language Processing Toolkit (NLTK) is a collection of conceptual and mathematical natural language processing (NLP) resources and applications developed in the Python programming language for the language English. The toolkit also provides a book that contains a detailed explanation of the notions responsible for the language processing tasks. The main purpose of NLTK is there to support research and teaching in Natural Language Processing (NLP). NLTK is also the prominent framework on which to develop Python programming.

A. Structure:

The remaining amount of the document is broken down into the relevant segment: Segment II defines the work performed in connection with this document. Section III tells about the approach and features used for classification. Section IV explains the classification algorithm and the results obtained.Section V contains conclusion and the future work that can be done.

B. Related Work:

Genre Classification is a broad concept and has its application in various fields, as an example, music, movies, news articles, blogs and text-based documents.

Mu Yong [4] used the idea of genre classification to organize the genre of song ci poem using its keyword features. The Song Ci poetry is basically segregated into two parts: Wan - Yue (somewhat delicate) and the other is Hao-Fang (heroic). The Song Ci poem exhibits various characteristics like it's concise, exquisite and use of small characters to describe emotion and mood, which makes it possible to classify the genre of the poetry. Three criteria for classifying the genre have been implemented. They were K-Nearest neighbor, Naive Bayes classification and Support Vector Machine. The results concluded that Support vector machine provided the most efficient results with 95% accuracy.

Named Entity Recognition is the most widely used concept of Natural Language Processing (NLP). The purpose of the identification of specified entities is to identify the proper nouns into previously defined classes.Darvinder Kaur, et al. [2] explored various ways in which the Named Entities can be recognized in various Indian languages. Text pre-processing includes Stop Words Removal as one of the most significant steps. Jaideepsinha K. Raulji, et al. [5], conducted a study to design an algorithm to identify and remove stop words. The algorithm designed was then implemented for Sanskrit Language. A dictionary-based approach was used for the algorithm and its implementation. As per their approach, a target text is taken and is compared with the list of predefined stop words.

Holly Chiang, et al. [1], in his paper, discussed the technique to categorize the genre of the book purely on the basis of the cover and title of the book. He concluded that the text and image yielded same accuracy while classifying the genre. The basic purpose of this study was to determine how accurately a cover can define the genre of the book.

II. LITERATURE REVIEW

Jia Song et al., [6] carried out a thorough analysis of imbalanced category issues and proposed a two-directional data classification clustering-based sampling approach using research undertaken. This problem is common in technologies in the actual world. For instance, unethical telephone conversations, data extraction and sorting activities, word learning, etc., are identified. As reported by Stamatatos and others[7], there are several stylometric functions, along with quantities of punctuation. The importance is also placed in terms and intervals of punctuation to explain the stylometric characteristics along the same lines.Remember that Stamatatos et al. focuses on word different frequencies in the entire English language as a means of enhancing overall performance rather than using word quantities through their collection of ridiculously short newspaper posts.

Yong-Bae Lee, etc. [3], incorporated an immediate genre identification technique that use the characteristics derived from both the categorized genre and the subject documents. The findings indicated that the approach proposed was much superior than the practical application of quantitative students sometimes used to classify the topic.

A. Features

The following features have been investigated for genre classification. The content-based characteristics have also been collected and optimized on the basis of the entire collection.

B. Pronouns:

The ratio of pronouns like 'I', 'We', 'You' over total word count provide a basic classification feature. According to observation Romantic novels tend to have a higher pronoun ratio as compared to other.

C. Punctuation Marks:

The ratio of punctuation marks like ', : , ?' over total word count. The punctuation marks are important as it gives the knowledge of direct - indirect speech used or what type of conversation the characters are having in the novel like questions, assertions etc.

D. Named Entity Relation:

The ratio of numbers of names used or numbers of persons over total number of entities used in the novel. This finds entities marked as PERSON.

E. Length of Data:

This defines the length of the training novel. This is especially useful as it gives significant information about genre. According to observations Fairy Tales are short and crisp.

F. Set of Words:

This defines a particular set of words which are used most in a particular genre as compared to any other. These set of words are calculated by finding the classification of intensity of different words found in a particular genre and collection the data to form a set of words that can be used in classification. Defined a list of words present in 2 different genres namely Thriller and Romantic.

III. RESEARCH METHODOLOGY

In this paper the Dataset consisted of English novels categized into two categories test data and train data. The test data constitutes about 31 novels whereas the train data constitutes about 80 novels. The Dataset included varied novels from the genre Romantic, Fairy Tales and Thriller. They were so chosen to maintain uniqueness among different novels chosen.

The dataset was converted to text files and used for the process. Each novel's content goes through a series of modification that included Stop word removal, Stemming, Named Entity Recognition. The above algorithms remove the unnecessary words and markings, to give raw data and words in its root form. The root forms are required to bring out only the centralized meaning of the associated words without any grammar associated with it.

The processed data now goes through information retrieval phase try to acquire the novels characteristic features present in its language and grammar. For this purpose, there are features defined that works on the process of word frequencies, punctuation frequencies, common words associated, etc. Each feature has its own representative value for each attribute of the genre and the combined result for all features forms one part for classification of the genre for that particular novel.

The classification is carried out using K-means Analysis. It uses genre name and the corresponding values of the feature as the parameters for classification. This helps in mapping the features of trained data and the features obtained from the test file using CCA (Canonical Correlation Analysis) and then computing the average distance from the corresponding class. The Genre of the data being tested is defined as the minimum average distance from each class.

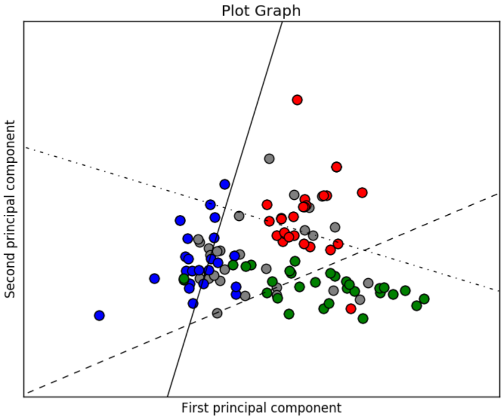

Fig. 1. Test Graph

In the above figure 1 the colored dots represent different genre being trained on namely "Fairy Tales" in Blue, "Romantic" in Green and "Thriller" in Red. The Grey dot defines the Data being tested as per the trained data.

The lines divide the whole area into 3 different hyper planes that defines the maximum probability of finding a book of a particular genre, which help us find the genre of the data being tested.

IV. RESULTS AND DISCUSSIONS

As per the observations on Test Data the result obtained is shown in Fig. 1 and Fig. 2 respectively.

Right =	28
Total =	31

Accuracy = 90.32%.

Fig. 2. Result

CONCLUSION

The main idea of this paper is to retrieve the information from the text and classify the document according to the genre. We use the technique to categorize the English novels into three different types of Fiction genre namely Romantic, Fairy Tales and Thriller. The genres taken are few of the most widely read genres of book among different age group. The various concepts of Natural Language Processing and Machine Learning, including stop word removal, named entity recognition, stemming, etc., have been implemented for processing the data to determine their respective genre. Around 31 books from different genres were collected as test data set and out of which 28 books were classified correctly, thereby giving accuracy of 90.32%.

FUTURE SCOPE

In the future, we wish to extend our study by using larger data sets and various types of genres. Furthermore, our classification may extend to web classification to explore new feature sets.

REFERENCES

Ge Yifan, Holly Chiang, and Connie Wu. "Classification of Book Genres By Cover and Title."

Kaur Darvinder, and Vishal Gupta. "A survey of named entity recognition in English and other Indian languages." IJCSI International Journal of Computer Science Issues 7.6 (2010): 1694-0814.

Lee, Yong-Bae, and Sung Hyon Myaeng. "Text genre classification with genre-revealing and subject-revealing features." Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2002.

Mu, Yong. "Using Keyword Features to Automatically Classify Genre of Song Ci Poem." Workshop on Chinese Lexical Semantics. Springer International `Publishing, 2015.

Raulji, Jaideepsinh K., and Jatinder Kumar R. Saini. "Stop-Word Removal Algorithm and its Implementation for Sanskrit Language."

Song, Jia, et al. A bi-directional sampling based on K-means method for imbalance text classification; Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on. IEEE, 2016.

Stamatatos, Efstathios, Nikos Fakotakis, and George Kokkinakis. "Text genre detection using common word frequencies." Proceedings of the 18th conference on Computational linguistics-Volume 2. Association for Computational Linguistics, 2000.

REFERENCES

1. Mohammed Faizan Zafar and Dr. Danish Ahmed Siddiqui (2019), “IMPACTS OF BRAND EXTENSIONS ON PARENT BRAND IMAGE”, Article in SSRN Electronic Journal, extracted from research gate.
2. Keller, K.L. (1993), “Conceptualizing, measuring, and managing customer-based brand equity”, Journal of Marketing, Vol. 57 No. 1, pp. 1-22.
3. Eva Martínez José M. Pina, (2003),”The negative impact of brand extensions on parent brand image”, Journal of Product & Brand Management, Vol. 12 Iss 7 pp. 432
4. Müge Arslan, F., & Korkut Altuna, O. (2010). The effect of brand extensions on product brand image. Journal of Product & Brand Management, 19(3), 170–180.
5. Smith, D.C. and Park, C.W. (1992), ‘‘The effects of brand extensions on market share and advertising efficiency’’, Journal of Marketing Rerearch, Vol. 29, August, pp. 296-313.
6. Eva Martínez, José M. Pina. 2005. Influence of Corporate Image on Brand Extensions: A Model Applied to the Service Sector. Journal of Marketing COMMUNICATIONS 11, 263-281.