|  |  |  |
| --- | --- | --- |
| A logo on a black background  Description automatically generated | | |
| **Faculty of Science and Technology**  **CISC3025 – Natural Language Processing** | | |
| **Project Task 2: Implementation of a Naïve-Bayes Text Classification Model and its Performance Evaluation in Reuters Dataset** | | |
| **Group Members:** | | |
| Huang Yanzhen | | DC126732 |
|  |  |  |

Table of Contents

[Introduction 3](#_Toc160150013)

[Background 3](#_Toc160150014)

[Approaches, Designs and Methods 3](#_Toc160150015)

[3.1 Python Packages 3](#_Toc160150016)

[3.2 Preparations 3](#_Toc160150017)

[3.2.1 Retrieve Information from train.json 3](#_Toc160150018)

[3.2.2 String Tokenization 3](#_Toc160150019)

[3.2.3 Instance-wise Word-Frequency Dictionary Construction 5](#_Toc160150020)

[3.2.4 Vocabulary Construction 5](#_Toc160150021)

[3.2.5 Data Summarization using DataFrame 5](#_Toc160150022)

[3.2.6 Feature Selection 5](#_Toc160150023)

[3.3 Naïve-Bayes Probability Calculation & Model Implementation 5](#_Toc160150024)

[3.4 Evaluation using F-Score 5](#_Toc160150025)

[Results 5](#_Toc160150026)

[Conclusion 5](#_Toc160150027)

# Introduction

# Background

# Approaches, Designs and Methods

## 3.1 Python Packages

There are some python packages that are useful for fulfilling this project. Below shows a list of packages imported and the corresponding tasks they are used to perform.

|  |  |
| --- | --- |
| **Table 1:** Python packages used in this project | |
| **Package Name** | **Usage** |
| json | Compile the .json source file into structural objects. |
| re | Substitute varieties of delimiters or useless symbols in natural language writings by a single space. |
| nltk | Construct the tokenizer for tokenizing a string into an array of words with a given delimiter. |
| pandas | Construct a DataFrame object to contain the word frequency matrix, indexed by vocabulary. |
| numpy | Some basic math tasks, like initializing an array filled with 0. |
| collections | Turn a word array (which may contain duplicates) into a dictionary of word-frequency pairs. |

## 3.2 Word Count in Different Classes

This corresponds to the Requirement 1 and 2. In this part, this report demonstrates the ideas and methods for information retrieval and data processing for counting frequencies of all the individual words in different classes.

### 3.2.1 Retrieve Information from train.json

By applying the internal package of json in Python, it is easy to extract the training data into a two-dimensional list in python. The first attribute of each instance, i.e., the training id, is ignored by intuition.

Traversing all the instances, the program accumulates the document number of each types by reading the second attribute of the instance. The third attribute is the actual sentence, which would be further tokenized into word arrays, accompanied by its class.

### 3.2.2 String Tokenization

The key task of this phase is to determine whether a string is a valid word. By design, the following table shows some of the disambiguity rules that drives the idea of tokenization.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table2:** Disambiguity rules to define a word | | | | |
| **Pattern** | **Examples** | **Validity** | **Interpretation** | **Correction** |
| In-segment dots | U.S. | Yes | Country name abbreviation; |  |
| Co. | Yes | Abbreviation for “cooporation” |  |
| 13.10 | Yes | Floating point numbers |  |
| In- segment Slashes | autumn/winter | No | Selections, i.e. autumn or winter. | ‘autumn’, ‘winter’ |
| 13-3/2 | Yes | Hybrid number. |  |
| In-segment Hyphens | government-to-government | Yes | Connected subwords forming a long word. |  |
| In-segment Commas | 2,365,000 | Yes | A large number separated by a comma. |  |
| &lt; and > | &lt;Banca> | No | Escape sign for “<”. | ‘Banca’ |
| Ending with a punctuation mark | nations' | Yes | Possessive pronoun of the original noun plural. |  |
| policy. | No | Extra period. | ‘policy’ |
| Brackets surrounded | (Bracket) | No |  | ‘Bracket’ |
| Pure punctuation marks | ... | No | Abbreviation mark | ‘’ |
| -- | No | Extension mark | ‘’ |
| **…** | | | | |

Having the ideas, the tokenization is performed in sequence by two rules.

The first rule is made to avoid in-segment symbols that connects two disjoint words. For instance, the segment “said…Taiwan” should be separated into two disjoint words. Below shows the regular expression for replacing these symbols by a space.

|  |  |  |
| --- | --- | --- |
| **Table3:** Regular expression for the first rule | | |
| **Rule** | r'\"|\.\.+|\(|\)|\s--+\s|(?<=[A-Za-z])/|&[a-z]+;|>' | |
| **Patterns** | | **Interpretation** |
| \" | | Any double-quotation mark won’t be a part of a valid word. |
| \.\.+ | | Any continuous periods always separates two disjoint words. |
| \(|\) | | Any brackets won’t be a part of a valid word. |
| \s--+\s | | Any continuous hyphens always separates two disjoint words. |
| (?<=[A-Za-z])/ | | A slash only separates two disjoint words but doesn’t separate two numbers. |
| &[a-z]+;|> | | Remove the escape sign of “<”, as well as the actual sign of “>”. |

The second rule is made to normalize different delimiters. There are all kinds of delimiters in the natural language, like a comma followd by a space, or a period followed by a space. The second rule replaces these patterns by a single space.

|  |  |
| --- | --- |
| **Table4:** Regular expression for the seond rule | |
| **Rule** | r'(?<![A-Z])([.,?!"]\s+)' |
| **Interpretation** | This rule basically listed many possible situations for a delimiter. Any punctuation mark followed by any number of spaces is likely to be a delimiter. However, this rule doesn’t apply for abbreviations like U.S. or U.K. |

Since all delimiters are unified to be arbitrary (mostly only one) number of spaces, it is very easy for a tokenizer to parse a sentence using the delimiter of “space”. However, there are still space to improve. There are still some words that contain some invalid words at the end. For instance, there is a word like “barry,”, having an extra comma. There were also invalid symbols like “--” presenting as a word string in the tokenization. Therefore, it is necessary to use the rstrip() function to filter these invalid tokens.

|  |
| --- |
| **Code 1:** Python code for filtering invalid tokens |
| for word in \_cur\_token\_array:  word = word.rstrip(',?!"-') # Remove extra punc marks at the end  cur\_token\_array.append(word) if word != "-" or "" else None |

Even though these rules and methods can’t guarantee a 100-percent accuracy of tokenization, it is completely useful for parsing the training data given.

After information retrieval and tokenizing, the data should be of the following form.

|  |
| --- |
| **Code2:** Data structure after processing |
| [  [‘acq’,[‘word1’,’word2’,’word3’,’word4’]],  [‘grain’,[‘word1’,’word2’,’word3’]],  [‘grain’,[‘word1’,’word2’,’word3’,’word4’,’word5’]],  [‘money-fx’,[‘word1’,’word2’]],  [‘acq’,[‘word1’,’word2’,’word3’,’word4’]],  …  ] |

### 3.2.3 Instance-wise Word-Frequency Dictionary Construction

This part only regards the scope of an instance. The goal of this part is to merge multiple identical words into a single key-value pair, where the key is the word string and the value is the frequency of the word.

An instance is a string-list pair, where the list represents the tokenized version of the original sentence and the string is the class of the sentence. The Counter() function extracted from package collections is used to turn the list into a dictionary of word-frequency pairs.

|  |  |
| --- | --- |
| **Table 5:** Construct word-frequency dictionary | |
| Instance before construction | [‘acq’,[‘the’,**’cake’**,’is’,’a’,’lying’,**’cake’**]],… |
| Instance after construction | [‘acq’,{‘the’:1,**‘cake’:2**,‘is’:1,‘a’:1,‘lying’:1}] |

The reason why converting original list into a dicitionary is necessary is that the order of the words doesn’t matter much for the present tasks, yet merging word counts saves space and time for frequency statistics.

### 3.2.4 Vocabulary Construction

Unlike dictionary construction, the vocabulary construction regards the scope of all instances of the data. The goal of constructing a vocabulary is to preserve a unique index for all words disregarding their frequencies, making it easier to store statistic values into the correct place.

By definition, the vocabulary embraces every word that had ever been present in any of the documents, and only store them once. Hence, it is intuitive to consider the vocabulary as a set. By traversing all the instances, the vocabulary only insert words it had never seen before into its storage. This is better when the instances are processed to merge identical words into one, saving the time for traversing. The built-in set() function is used to perform the construction.

|  |
| --- |
| **Code3:** Vocabulary Construction |
| vocab = **set()**  for instance in tokenized\_test\_data:  for word in instance[1]:  **vocab.add(word)**  vocab = list(vocab) |

### 3.2.5 Data Summarization using DataFrame

It is observed that the size of vocabulary is around 43,000. For each individual word, we need to count its frequencies in five different classes. Therefore, it is more efficient to construct a Token Frequency Matrix.

The row indexes is the vocabulary, while the column indexes are five different classes. By traversing every instances of the train data, the program first get the class of the instance. It then adds the frequency value of every unique word in the dictionary into the matrix. Notice that the matrix cell is located by the word-class pair. For each instance, the insertion is performed only on one column of the matrix since there’s only one class in an instance.

Below shows one possibiolity of the token frequency matrix. The matrix is realized using the DataFrame object from the pandas package, which by default generates index in a random order.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 6:** The token frequency matrix | | | | | |
| Index\Class | crude | grain | money-fx | acq | earn |
| prior | 4 | 7 | 1 | 27 | 501 |
| 30 | 33 | 47 | 34 | 133 | 300 |
| board | 6 | 24 | 12 | 343 | 192 |
| six | 32 | 22 | 106 | 81 | 286 |
| Year | 0 | 0 | 0 | 0 | 526 |
| LOSS | 0 | 0 | 2 | 0 | 470 |
| May | 11 | 47 | 11 | 54 | 333 |

This matrix is then sorted using the sum of each column. Since we want to then select 10,000 most frequent words as feature words, we need to sort this matrix. Due to lack of functionalities of DataFrame, it is required to calculate the sum during matrix construction, and then slice out the last column storing the sum of each row after sorting. The algorithm below shows the full construction process.

|  |  |
| --- | --- |
| **Algorithm 1:** Token Frequency Matrix construction | |
| 1 | **Procedure** ConstructTFMatrix(tokenizedDataFreq, vocab, classes); |
| 2 | Let TFM←new Matrix(row=vocab, col=classes) |
| 3 | **for** row **in** tokenizedDataFreq **do** |
| 4 | Let curData ← tokenizedDataFreq.class; |
| 5 | Let curWordDict ← tokenizedDataFreq.wordDictionary; |
| 6 | **for** word, frequency **in** curWordDict **do** |
| 7 | TFM [word, class] += frequency; |
| 8 | TFM ←sortByRowSum(TFM, order=descending); |
| 9 | **Output** TFMatrix. |

### 3.2.6 Feature Selection

## 3.3 Naïve-Bayes Probability Calculation & Model Implementation

This corresponds to the Requirement 3 and 4.

## 3.4 Evaluation using F-Score

# Results

# Conclusion