**Name: \_\_\_Sri Vamshi Polela\_\_\_\_**

**CSDA 5430 Predictive Analytics**

**Week 6 Assignment**

1. **Discuss the necessity of feature selection on the case study QOL (quality of life and chronic disease) that we implemented in class. How does feature selection compare with dimension reduction such as PCA in dimensionality reduction?**

**Feature selection:**

* It is a process of selecting the features which are most impacting on a target variable.
* Faster Prediction: This helps us to predict in less time compared to prediction using all features.
* Data Efficiency: It can also decrease the collection of data for businesspeople.

**Case Study:**

* In the case of the Quality of Life (QOL) and chronic disease dataset, feature selection is like selecting the most important features of information from a large dataset like QOL. It helps researchers focus on specific factors that strongly influence the quality of life for patients with chronic illnesses. This makes the analysis more manageable and can reveal important insights.

**Dimension Reduction:**

* Dimension reduction is a process of simplifying complex datasets by reducing the number of dimensions or features. It involves creating a new set of dimensions that captures the most significant variability in the data, often using techniques like Principal Component Analysis (PCA). By doing so, dimension reduction aims to retain essential information while making the data more manageable, facilitating easier analysis and interpretation.

**Comparison of feature selection and dimension reduction such as PCA**

* Comparing feature selection to Principal Component Analysis (PCA) is like choosing between selecting a few important details and transforming all details into a new set of details. Feature selection keeps the original details but focuses on the most relevant ones, making it easier to understand their impact. PCA, on the other hand, transforms all details into a different form, which might be harder to interpret directly. In healthcare, where understanding specific factors is crucial, choosing the right details through feature selection is often more useful. **(Okay. Variable selection identifies representative subset from original set of variables to retain information in the data, whereas dimension reduction identifies a new set of variables that approximate, or as function of, original variables to reduce the complexity of the data. The former is a discrete process, and the latter is a continuous process.)**

1. **From the discussion in first week, we understand that variety is one of big data’s characteristics, which relates to unstructured data. There seems to be a direct relationship between preprocessing the unstructured data and the effectiveness of analytics on the unstructured data. Using text mining as example, discuss unstructured data and how they relate to data analytics process.**

* In the first week's discussion, we learned that variety is a key characteristic of big data, often associated with unstructured data such as text. unstructured data lacks a clear, tabular structure. Examples include text documents, emails, social media posts, images, audio, and video files. Unstructured data is often more varied and doesn't fit neatly into rows and columns, making it challenging to analyze using traditional methods.
* When dealing with unstructured data, such as in text mining, preprocessing becomes crucial. This involves cleaning, organizing, and transforming the data into a structured format that analytics tools can understand. The effectiveness of analytics on unstructured data relies heavily on how well the preprocessing is done. By converting unstructured text into meaningful features, like identifying key words or sentiments, data analytics can extract valuable insights, patterns, and knowledge from the diverse and varied information contained in unstructured data. **(Okay. Text is unstructured data with free words and sentences flow representing linguistic sentiments. Text mining is more difficult than mining structured data business collected in their operational database. As we’ve seen in this class, most traditional data analytics algorithms work well on structured data, like cross-sectional data, but may not directly on unstructured data. So, text data need be mining-friendly, meaning preprocessed well, before being fed into any mining algorithm.)**

1. **Consider the following text version of a post to an online learning forum in a statistics course:**

**Text, letter

Description automatically generated**

1. **Identify 10 non-word tokens in the passage.**

Non-word tokens are elements in the text that are not considered words. In the provided passage, the following are 10 non-word tokens:

1. !
2. <
3. >
4. / (forward slash)
5. =
6. “
7. ”
8. &
9. ;
10. \ (backward slash)
11. ? (Question mark)
12. ,
13. .
14. {
15. }
16. : **(Okay. Non-word tokens include html tags and punctuations, such as: !, “, comma.)**
17. **Suppose that this passage constitutes a document to be classified, but you are not certain of the business goal of the classification task. Identify material (at least 20% of the terms) that, in your judgment, could be discarded fairly safely without knowing that goal.**

|  |  |
| --- | --- |
| * Thanks * John * <br /> * Xfont * size="3" * &quot; * Illustrations * demos * for * need | * students * to * work * through * on * their * own * </font›> * Do * we |

* In the absence of knowledge about the specific business goal of the classification task, we can remove these terms.

These terms are primarily related to salutations, HTML formatting, and common words that might not carry significant semantic value for understanding the core content or intent of the passage.

**(Okay. Basically, everything in the top row and the bottom three rows could be discarded, since they are for formatting purpose for better communicating with human beings.)**

1. **Suppose that the classification task is to predict whether this post requires the attention of the instructor, or whether a teaching assistant might suffice. Identify the 20% of the terms that you think might be most helpful in that task.**

* **Keywords Indicating Need for Attention:**
* finish
* project
* illustration
* demos
* find
* students

**(More can be used. Including the phrases “do we need to finish project” and “where to find illustrations and demos”.)**

1. **What aspect of the passage is most problematic from the standpoint of simply using a bag-of-words approach, as opposed to an approach in which meaning is extracted?**

**Bag-of-Words Approach:**

* In a bag-of-words model, the text is represented as an unordered set of words, disregarding grammar and word order.
* Each word in the text is treated as a separate and independent feature, and the overall meaning of the text is derived from the frequency of individual words.
* It doesn't consider the structure, context, or relationships between words, and the order of words is not preserved.

**Problematic Aspect in Bag-of-Words Approach:**

* The passage contains HTML tags, entities, and incomplete structures (e.g., <br />, &quot;), which are non-word elements.
* In a bag-of-words model, these elements might be treated as separate tokens, potentially leading to noise and a less accurate representation of the true meaning of the text.
* The lack of consideration for the structure and context in a bag-of-words approach makes it challenging to handle these non-word elements appropriately. **(The bag-of-words approach considers only word tokens so url’s, special characters, and html tags are not top problems, and can be easily removed. Probably the more problematic aspect of the post is that, there are two voices – that of the inquirer and that of the responder. The top paragraph is a quoted sentence from the instructor and bottom is the question from student. A bag of words jumbles this all together, and any meaning associated with there being two voices would be lost.)**

**8/10 points**

1. **In text mining, TF-IDF is used to measure the relevance of terms (tokens) in a collection of documents. Consider the following four sentences, after text preprocessing (such as text clean up, tokenization, and text reduction), a TDM (term-document matrix) in TF can be obtained, as shown. Based on the TDM, calculate the numbers for normalized TF, IDF, and TF-IDF of each term.**

|  |  |
| --- | --- |
| **Sentences** | **TDM in TF** |
| 1. this is the first sentence!! 2. this is a second Sentence: 3. the third sentence, is here 4. fourth of all sentences from first to fourth. | Text  Description automatically generated |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Terms | TF (normalized)  1 2 3 4 | | | | IDF | TF-IDF  1 2 3 4 | | | |
| first | 1/2 | 0 | 0 | 1/4 | log2(4/2) = 1 | 0.5 | 0 | 0 | 0.25 |
| fourth | 0 | 0 | 0 | 2/4 | log2(4/1) = 2 | 0 | 0 | 0 | 1 |
| second | 0 | 1/2 | 0 | 0 | log2(4/1) = 2 | 0 | 1 | 0 | 0 |
| sentenc | 1/2 | 1/2 | 1/2 | 1/4 | log2(4/4) = 0 | 0 | 0 | 0 | 0 |
| third | 0 | 0 | 1/2 | 0 | log2(4/1) = 2 | 0 | 0 | 1 | 0 |

**Term Frequency:**

**Inverse Document Frequency:**

**Combining TF and IDF**

* TF-IDF (term frequency - inverse document frequency) measures the relative importance of a term.

𝑇𝐹−𝐼𝐷𝐹(𝑡,𝑑) = 𝑇𝐹(𝑡,𝑑)×𝐼𝐷𝐹(𝑡)

**(All good.)**

1. The zipped file *AutoAndElectronics1.zip* contains two folders: auto posts and electronics posts, each contains a set of 500 posts organized in small files. The posts are taken from Internet groups devoted to autos and electronics, so they are either auto-related or electronics-related. Two examples are provided at the end.

In R, perform the following steps as discussed in class on the data. We would like to build a classification model on the data to classify Internet discussion posts as either auto-related or electronics-related.

* Step 1: Collecting data (Discuss the business problem and how the data can support the analytics. You may use the following line of R code to import the zip file into a corpus object.)

# read zip file into a corpus

corp <- Corpus(ZipSource("AutoAndElectronics1.zip", recursive = T))

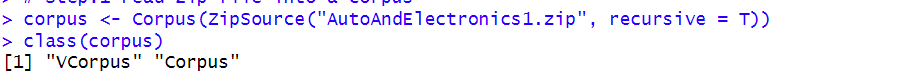
**Business problem:**

* + To classify, the posts taken from the internet discussion are either auto related or electronics related.
  + To achieve this, we must build a classification model which automatically categorizes these posts into either auto-related or electronics-related groups.

**Data Understanding:**

* Here the given data is unstructured data. To make this data an analytical usage we have convert it to structured data or analytical tool understandable data.
* To convert this as structured data, data preparation is needed that includes data cleanup, reduction, and tokenization, such as turning to lower case, removing white space, punctuation, numbers, stop-words, and stemming.
* The data from internet posts provides the basis for training a classification model, allowing it to learn patterns associated with auto and electronics topics. This allows automatic sorting of new posts, making it easier to analyze and organize online discussions. Ultimately, the data supports the creation of an efficient algorithm for classifying posts into auto-related or electronics-related categories.

**Loading the given zipfile into corpus object:**



**(Okay. The goal is to predict whether a new discussion post as either auto-related or electronics-related. Outcome variable is the label with values (auto-related or electronics-related) we attach on each document. The important class may depend on business goal.)**

* Step 2: Exploring data and preparing data (Perform necessary text preprocessing. Create training partition (60%) and testing partition (40%) with randomized partitioning. Set a seed so the results can be reproduced.)
* Text preprocessing include data cleanup, reduction, and tokenization, such as turning to lower case, removing white space, punctuation, numbers, stop-words, and stemming.

**Clean the Corpus Object:**

A computer screen shot of a program

Description automatically generated

* Create DTM or TDM objects for each group (auto-related (top 500) and electronics-related (bottom 500). Find or plot frequent words on each group. Plot a couple of word clouds. Identify some keywords for each group.

**Created DTM:**

**A screenshot of a computer code

Description automatically generated**

**Create DTM objects for each group (auto-related (top 500) and electronics-related (bottom 500).**

**A screenshot of a computer code

Description automatically generated**

**Frequent words for auto-related:**

A graph with blue lines

Description automatically generated with medium confidence

**Frequent words for electronic-related:**

**A graph of blue lines

Description automatically generated with medium confidence**

**Word cloud for Auto-related:**

**A close up of words

Description automatically generated**

**Word cloud for electronic-related:**

**A close up of words

Description automatically generated**

**Keywords which are common in both Auto and Electronic related posts:**

**A screenshot of a computer code

Description automatically generated**

* Create a TF-IDF weighted TDM (term-document matrix). Perform LSA (with dim=5) to receive a LSA object from the TF-IDF matrix.

A white background with black text

Description automatically generated

**Using LSA (Latent semantic indexing [or latent semantic analysis] (LSA) which involves dimensionality reduction.**

A number of numbers on a white background

Description automatically generated Remember use label for each group of posts (1 for auto-related and 0 for electronics-related)

**(Okay)**

* Step 3: Training a model on the data (Use the function *glm()* for the classification algorithm to run a logistic regression model. Remember to ignore label as part of predictors.)

A close-up of a math equation

Description automatically generated

**(The model looks good)**

* Step 4: Evaluating model performance (Evaluate the model performance against the test data with confusion matrix. Identify accuracy, sensitivity, and specificity. Use the default threshold value of 0.5 for classification.)

**Fitting logistic regression using train data and evaluating using holdout data:**

A screenshot of a computer

Description automatically generated

**(Identify accuracy, sensitivity and specificity? Which class is the positive class?)**

* Step 5: Improving and comparing model performance (Use any other classification algorithm(s) for modeling. Compare and comment on results.)

**I am using classification tree for comparing results of logistic regression:**

**A screenshot of a computer

Description automatically generated**

**Comparison: When positive class is ‘0’ means I took important class as Electronic-related.**

* The logistic regression model slightly outperforms the decision tree model in terms of overall accuracy (92.75% vs. 90.75%).
* Both models have high sensitivity (correctly identifying the positive class), with a value of 95.00%.
* The logistic regression model shows higher specificity (correctly identifying the negative class) compared to the decision tree (90.50% vs. 86.50%).

**From the above two models results we can choose logistic regression for classification task. (Normally we use 1 for important or positive class, instead of 0. If electronic related is important class, when labeling, it may be labeled as 1. Overall good.)**

**18/20 points**

**Total 56/60 points**