INSOFE

Inspire...Educate...Transform.

Inspire…Educate…Transform.

Text mining : Full day lab
20160910 - Batch 17 –CSE7306c

Group Activity -Author Identification

1. Make groups with 3 benches into 1
2. This is a group activity
3. 10 minutes will be given for this task.
4. Task: Author identification
    a. Identifying the author of a given text.
    b. Classification problem
    c. Documents with known authors are used for training
    d. The aim is to automatically determine the corresponding author of an anonymous text using the training data
5. You are given total 300 paragraphs of 3 different authors (100 paragraphs for each other)
6. Write the steps that you follow to solve this classification problem in detail

Steps for the GroupActivity : Author Identification

1. Read the text data
2. build a corpus using the text mining (tm) package
3. Data processing
   - Text cleansing
   - Key word extraction: candidate keyword list, TF-IDF
   - Term extraction, dimensionality reduction, feature selection, etc
4. clean up the corpus using tm_map()
   a. removeNumbers
   b. removePunctuation
   c. stripWhitespace
   d. content_transformer(tolower)
   e. removeWords using stopwords("english")
   f. stemDocument
   g. Covert to PlainTextDocument
5. Vectorization : Create DocumentTermMatrix
6. Convert DTM into dataframe
7. Do the Visualizations to understand the important attributes of each author
8. Decide up on the removal of sparse terms from a document-term matrix – decide the sparse percentage
9. Add the class label to the dataframe
10. Split the dataset into train and test
11. Build the Model (you can experiment with different models also)
12. Predict for the test dataset
13. Compute the evaluation metrics

Text mining

Text categorization (a.k.a. **text classification**) is the task of assigning predefined categories to free**-text** documents.

- Pre-given categories and labelled document examples
- Classify a new document
- A standard classification modelling (Supervised learning)

Steps involved in text classification

- Labeled text documents
- Text data pre-processing
  - Text cleansing
  - Key word extraction: candidate keyword list, TF-IDF
  - Term extraction, dimensionality reduction, feature selection, etc.
- Define Train and test sets
- Create classification model on train data
    - Vector space models:
      - SVM, KNN, Decision trees, Neural nets, etc.
    - Probabilistic models
      - Naïve Bayes classifier.
- Classification model evaluation
- Classification of unknown text documents.

Preprocessing steps

- library(tm)
- Create the corpus : Corpus(DataframeSource(data.frame(text)))
- Preprocessing using tm_map
- removeNumbers – removes numbers,
- removePunctuation – removes punctuation symbols,
- stripWhitespace – removes extra spaces,
- removeWords - stopwords(language=‘english’) – removes stopwords for the language specified -
- content_transformer(tolower) – transforms all upper case letters to lower case,

DocumentTermMatrix

- Once the data is cleansed, create the document term matrix.
  - Use as is DTM matrix
  - Use TFIDF matrix
  - Use Binary form
- Convert the document term matrix as a data.frame to find features and perform modeling

Keyword extraction (ngram)

- Manually inspect a sample of documents in each category and list down the keywords that are representative of the class. Example of some 1-grams are:
- “auto”
  - Airbags, acceleration, gear, speeding, automotives, etc.
- “med”
  - Diagnosis, treatment, illness, disease, death, symptoms, etc.
- “comp’
  - Graphics, visuals, image, processor, resolution, compiler, etc.

Define scores based on the number of keywords found for each class.

Adding class attribute

- Add the class attribute to the document term dataframe

```
class <- c(rep("auto",1000), rep("med",1000), rep("comp",1000))
FullData.p.DtM.C = cbind(dt_matrix,class)
```

At this stage, build models to check for accuracy of classification.

To improve further, reduce features, build dictionary of terms and score the document for classification.

Feature reduction: remove sparseness

1. Remove sparse terms

dt_matrix <- removeSparseTerms(FullData.p.DtM, 0.99)

terms with frequency atleast >N(1-0.99) will be retained. N is total number of docs.

Feature reduction: Random Forest

- Use random forest on the data and find terms that are important.
- From the random forest model summary, you obtain the importance of terms.
- varImp(rfmodel) or
- importance(rfmodel,type=2)
- Use reduced number of terms only in the model and ignore the rest.

Model building

- Split the data into train and test data sets
- Build the model
- Obtain train and test accuracy

Several aspects to consider

- There is no single way of approaching a text analysis.
- Text preprocessing
  - Dates/Numbers/Abbreviations/Alphanumerics/stop words/stemming/spell-check, etc. might be important based on the domain and quality of data. Need careful analysis and use of regular expressions to remove or retain.
- Keywords extractions
  - Manual inspection, frequency of words/wordcloud, POS, Named-entities, wordnet, etc.
  - Ngrams
  - Synonyms dictionary lookup, Abbreviations-Full forms dictionary, etc.
  - Sentiment, confounding characteristics, opinions, etc.
- Feature engineering
  - Use DTM matrix/TF-IDF/Binary data
  - Defining scores based on keywords
  - Defining semantic based rules, etc.

Several aspects to consider

- Feature reduction
  - Sparsity, C5.0 importance, RF importance, PCA, SVD, etc.
- Modeling and evaluation
  - Train-Test split, Stratified split in case of class imbalance, Merging categories, etc.
  - Semantic rules
  - LSA, text summarization – Ranking
  - Clustering – stability check, etc.
  - Classifier – confusion matrix, ROC, etc.

R References

- http://finzi.psych.upenn.edu/library/LSAfun/html/genericSummary.html
- https://rstudio-pubs-static.s3.amazonaws.com/31867_8236987cf0a8444e962ccd2aec46d9c3.html
- https://eight2late.wordpress.com/2015/07/22/a-gentle-introduction-to-cluster-analysis-using-r/
- http://www.academypublisher.com/jetwi/vol01/no1/jetwi01016076.pdf

Text mining infrastructure in R

- http://www3.ntu.edu.sg/sce/pakdd2006/tutorial/pakdd06-Tutorial%20Text%20Clustering.pdf

Mahout implementation of Text classification

- https://mahout.apache.org/users/classification/twenty-newsgroups.html

Wordnet dictionary in R

- https://cran.r-project.org/web/packages/wordnet/vignettes/wordnet.pdf

International School of Engineering

Plot 63/A, 1st Floor, Road # 13, Film Nagar, Jubilee Hills, Hyderabad - 500 033

For Individuals:	+91-9502334561/63 or 040-65743991
For Corporates:	+91-9618483483
Web:	http://www.insofe.edu.in
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	http://www.youtube.com/InsofeVideos
SlideShare:	http://www.slideshare.net/INSOFE
LinkedIn:	http://www.linkedin.com/company/international-school-of-engineering

This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.