Big Data Analytics

Dr Sirintra Vaiwsri | Email: sirintra.v@itm.kmutnb.ac.th

Text Analytics

Text Analytics (Dietrich et al., 2015)

- Text analytics refers to the representation, processing, and modelling of textual data to create useful insight.
- It often suffers from the high dimensionality of data to be analysed.
- The text analysis steps usually consist of:
    - Parsing
    - Searching
    - Text mining

Text Analysis Steps (Dietrich et al., 2015)

- Parsing refers to the process of taking unstructured text and imposing a structure for further analysis.
- Searching refers to the identification of the documents that contain search items, also called “key terms”.
- Text mining uses the terms and indexes (created from the parsing and searching) to discover meaningful insights.
    - Clustering and Classification can be applied to text analysis.

Text Analysis Steps (Dietrich et al., 2015)

- All three steps do not have to be present in every text analysis project.
- For example, the project might focus on the parsing task that uses one or more text preprocessing techniques
  - Such as part-of-speech (POS) tagging, named entity recognition, lemmatisation, or stemming.

POS Tagging (Dietrich et al., 2015)

- The goal of POS tagging is to build a model that receives a sentence as an input.
- For example,

  Sentence: he saw a fox

  POS tagging: PRP VBD DT NN where, according to Penn Treebank tags (Taylor et al., 2003), PRP is personal pronoun, VBD is a verb (past tense), DT is a determiner, and a NN is noun

Lemmatisation (Dietrich et al., 2015)

- The goal of lemmatisation is to find the correct dictionary base form of a word.
- For example,

  Sentence: Pyspark causes many problems.

  Lemmatisation: Pyspark cause many problem

Stemming (Dietrich et al., 2015)

- The goal of stemming is to reduce variant forms.
- Stemming does not need a dictionary.
- The algorithm is such as Porter's Stemming Algorithm.
- For example,

  Sentence: Pyspark causes many problems.

  Stemming: Pyspark caus mani problem

Text Analysis Process (Dietrich et al., 2015)

1. Collect raw text
2. Represent text
3. Conduct analysis such as Term Frequency-Inverse Document Frequency (TFIDF), Topic Modeling, and Sentiment Analysis.
4. Gain insights

Text Analysis Implementation

- Import libraries
    - SparkSession
    - IntegerType from pyspark.sql.types
    - All from pyspark.sql.functions
    - HashingTF, Tokenizer, StopWordRemover from pyspark.ml.feature
    - Pipeline from pyspark.ml
    - LogisticRegression from pyspark.ml.classification
    - MuiclassClassificationEvaluator from pyspark.ml.evaluation

Text Analysis Implementation

- Create SparkSession
- Read data from file (reviews_rated.csv)
- Select Review Text and Rating columns where Review Text can be trimmed using trim() and Rating should be converted to IntergerType
- Show data
- Create Tokenizer using:
    - <your tokenizer> = Tokenizer(inputol=<“your review text column”>, outputCol=<“your review text words column”>

Text Analysis Implementation

- Create StopWordsRemover using:
  - <your stop word remover> = StopWordsRemover(inputol=<your tokenizer>.getOutputCol(), outputCol=<“your meaningful words column”>

- Create HashingTF using:
  - <your hashing TF> = HashingTF(inputol=<your stop word remover>.getOutputCol(), outputCol=“features”

Text Analysis Implementation

- Create Pipeline by using <your tokenizer>, <your stop word remover>, and <your hashing TF> as stages
- Create train and test datasets
- Show train dataset
- Fit train data to the pipeline
- Transform train data to a new train Dataframe
- Transform test data to a new test Dataframe

Text Analysis Implementation

- Create LogisticRegression
- Fit train Dataframe to the created LogisticRegression
- Transform the test Dataframe to the model
- Show <“your meaningful words column”>, < “your label column”>, and <“your prediction column”>
- Create MulticlassClassificationEvaluator
- Evaluate the accuracy using your created MulticlassClassificationEvaluator

Assignment (2 points)

- Please implement the text analysis following slides 10-4.
- Please show your code and your results to get 2 points.
- Results: 3 Dataframes and 1 Accuracy value

References

- Dietrich, D., Heller, B., & Yang, B. (2015). *Data science & big data analytics: discovering, analyzing, visualizing and presenting data*. Wiley.
- Taylor, A., Marcus, M., & Santorini, B. (2003). The Penn treebank: an overview. *Treebanks: Building and using parsed corpora*, 5-22.