CSE 4990/6990 – BIG DATA AND DATA SCIENCE

PREDICTING GOOGLE APP RATING

Team: 5DBminds
Presenter: Lucas Ribeiro

PROJECT OVERVIEW

"PREDICTING PLAYSTORE RATING FOR APPS"

USING PLAYSTORES DATA OF A LARGE NUMBER OF EXISTING APPLICATIONS

TO PREDICT RATINGS FOR FUTURE APPS BASED ON THE ATTRIBUTES OF THAT APPLICATION

WHERE WE WERE

PREVIOUSLY, WE HAD OUR DATA FROM JSON FORMAT CONVERTED INTO CSV ALONG WITH REMOVAL OF UNNECESSARY DATA. WE WERE PLANNING TO LOOK AT ASSIGNING NUMERICAL VALUES TO THE FEATURES AND THEN APPLY VARIOUS SUPERVISED LEARNING ALGORITHMS.

CURRENT STATUS

THE DATA ITSELF IS IN A JSON FORMAT, TO USE IT IN OUR PROJECT WE HAVE

SPLIT IT INTO SMALLER DATA CHUNKS

LOADED INTO PYTHON DATA STRUCTURE DICTIONARY

ASSIGNED IT TO PANDAS DATAFRAME USING THE ATTRIBUTES AS COLUMNS

DATA PREPROCESSING

- CLEANING(NOT USEFUL DATA)
    - BINNING METHODS
    - **CLUSTERING METHODS**
    - **COMBINED COMPUTER AND HUMAN INSPECTION**

DATA PREPROCESSING

- INTEGRATION(DATA FROM MULTIPLE SOURCES)

DATA PREPROCESSING

- TRANSFORMATION(DATA THAT IS NOT APPROPRIATED YET FOR MINING)
    - NORMALIZATION(SCALING)
    - **SMOOTHING(REMOVE NOISE)**
    - AGGREGATION(SUMMARIZE DATAS)
    - **GENERALIZATION(REPLACE DATA WITH HIGH LEVEL CONCEPTS)**

DATA PREPROCESSING

- REDUCTION(DATA TOO BIG THAT NEED TO BE REDUCED WITH INTEGRITY)
    - DATA CUBE AGGREGATION
    - **DIMENSION REDUCTION**
    - **DATA COMPRENSION**
    - NUMEROSITY REDUCTION
    - **DISCRETIZATION AND CONCEPT HIERARCHY GENERATION**

DATA TRANSFORMATION

- FEATURES HAVING TEXTUAL VALUES
  - THAT NEED VECTORIZATION
  - THAT NEED NUMERICAL RATING
- FEATURES HAVING NUMERICAL VALUES

TEXTUAL FEATURES - VECTORIZATION

- CONVERSION OF TEXTUAL FEATURES TO CORRESPONDING NUMERICAL VALUES
- APPS' ATTRIBUTES NEEDS VECTORIZATION ARE:
- 1. DESCRIPTIONS
- 2. CATEGORY
- 3. REVIEW-TEXT
- 4. DEVELOPER
- 5. APP NAME

TEXTUAL FEATURES

- **CONTENT RATING**
- **ISTOPDEVELOPER**
- **HAVEINAPPPURCHASE**
- **ISFREE**

FEATURES - NUMERICAL VALUES

- **FEATURES THAT ARE DIRECTLY USABLE**
    - APP-SIZE, PRICE,
- **FEATURES THAT NEED SOME SORT OF TRANSFORMATION**
    - PUB DATE
    - LASTDATE
    - INSTALLATION
- **FEATURES HAVING RANGE OR DATE**
    - INSTALLATION EG. 50-100(WILL BE TRANSFORMED IN TWO COLUMNS)

HOW WILL IT BE DONE?

- CLASSIFICATION
- REGRESSION
- CLUSTERING
- DIMENSIONALITY REDUCTION
- MODEL SELECTION
- PREPROCESSING

ISSUES

- NUMBER OF COLUMNS IN THE VECTORIZED MATRIX IS HIGH WHICH CAN CAUSE MEMORY ERROR OR VERY HIGH PROCESSING TIME
- DESCRIPTION COLUMNS CONTAINS UNICODE THAT NEED TO BE TRANSFORMED
- CATEGORY NEED TO HAVE WEIGHTS WHICH COULD BIAS OUR PREDICTION

NEXT PHASE

1. **VECTORIZATION**
    - NARESH ADHIKARI
2. **CLUSTERING**
    - LUCAS RIBEIRO
3. **APPLY FOLLOWING SUPERVISED ALGORITHM**
4. **LINEAR REGRESSION**
    - NAILA BUSHRA
5. **NAIVE BAYES**
    - AYUSH RAJ ARYVAL
6. **K NEAREST NEIGHBOUR**
    - NARESH ADHIKARI

THANK YOU