EPL448: Data Mining on the Web – Lab 2

University of Cyprus
Department of
Computer Science

Παύλος Αντωνίου

Γραφείο: Β109, ΘΕΕ01

Project topics

1. https://www.kaggle.com/competitions
   - Choose any competition from Kaggle competitions
2. https://www.kaggle.com/datasets
   - You can choose any dataset from Kaggle datasets
3. https://www.machinehack.com/hackathons
   - You choose any hackathon from Machine Hack hackathons

- We strongly recommend you choose a regression or classification project involving both numerical and categorical data (strings, dates, not images)
- **Project Delivery Day: April 14, 2024 @ 23:59**
- **Team Project: 2-3 persons / team**
- **Send 1 email per team by February 19th stating the members of the team and the selected topic**

Project

- Successful project submission consists of:
  - Source code of your implementation along with instructions of how to compile and run your program.
    - Source code should be always up-to-date with a github repository so as to measure each team member's contribution to implementation
  - Input dataset used for your experiments. You can provide a link in case dataset size is large.
  - A description of your project (6 pages, two columns), specifying in detail your goals, approach, milestones, evaluation methodology and experimental results. A document that describes how you did your experiments and what are your obtained results.
  - Project ppt presentation (15 min / team + 5 min Q/A)

Project implementation steps

- Problem definition
- Data cleaning (fill missing/correct erroneous values)
- Data encoding (string/dates to numerical)
- Data scaling (rescale columns if in different scales)
- Exploratory Data Analysis (EDA): visualize feature importance, feature correlation (with heatmaps & scatterplots), feature distributions (histograms), feature outliers (boxplots)
- Feature engineering (feature selection / extraction)
- Training and prediction phase (including cross-validation, multiple predictors (regressors / classifiers), parameter best values selection)

Task0: Αριθμός προϊόντων / χώρα

- Βρείτε τον αριθμό προϊόντων και το άθροισμα των πωλήσεων που έγιναν ανά χώρα, δοθέντος του αρχείου δεδομένων SalesJan2009.csv:

Transaction date	Product	Price	Payment Type	Name	City	State	Country	Account Created	Last Login	Latitude	Longitude
01-02-2009 6:17	Product1	1200	Mastercard	carolina	Basildon	England	United Kingdom	01-02-2009 6:00	01-02-2009 6:08	51.5	-1.1166667
01-02-2009 4:53	Product1	1200	Visa	Betina	Parkville	MO	United States	01-02-2009 4:42	01-02-2009 7:49	39.195	-94.68194
01-02-2009 13:08	Product1	1200	Mastercard	Federica e Andrea	Astoria	OR	United States	01-01-2009 16:21	01-03-2009 12:32	46.18806	-123.83
01-03-2009 14:44	Product1	1200	Visa	Gouya	Echuca	Victoria	Australia	9/25/05 21:13	01-03-2009 14:22	-36.13333	144.75
01-04-2009 12:56	Product2	3600	Visa	Gerd W	Cahaba Heights	AL	United States	11/15/08 15:47	01-04-2009 12:45	33.52056	-86.8025
01-04-2009 13:19	Product1	1200	Visa	LAURENCE	Mickleton	NJ	United States	9/24/08 15:19	01-04-2009 13:04	39.79	-75.23806

Task0: Αριθμός προϊόντων / χώρα

- Μπορείτε να χρησιμοποιήσετε τον κώδικα του WordCount και να τον τροποποιήσετε κατά το δοκούν:

  http://www.cs.ucy.ac.cy/courses/EPL448/labs/LAB02/WordCount.java

- Να λυθεί με ένα Map/Reduce πρόγραμμα.
- Παράδειγμα αρχείου εξόδου (part-r-00000):
  - `Argentina 1 1200`
  - `Australia 38 64800`
  - `Austria 7 10800`
  - …

Task1: N-Gram

- N-Gram είναι η συνεχόμενη ακολουθία Ν όρων από μια δεδομένη ακολουθία κειμένου ή ομιλίας
  - Βρίσκει εφαρμογή στη φωνητική αναγνώριση
- Οι όροι μπορεί να είναι συλλαβές, γράμματα λέξεις κτλ ανάλογα με την εφαρμογή
- Παράδειγμα:
  - Όροι: γράμματα, N = 3
    - Ερώτηση: Βρείτε τα 3-grams που προκύπτουν από την πρόταση "good morning"
    - Απάντηση: "goo", "ood", "od ", "d m", " mo", "mor", … κτλ.
  - Όροι: λέξεις, N = 2
    - Ερώτηση: Βρείτε τα 2-grams που προκύπτουν από την πρόταση "good morning my friend"
    - Απάντηση: "good morning", “morning my", “my friend"

Task1: N-Gram

- Αλλάξτε τον κώδικα του WordCount έτσι ώστε να μετρά πόσες φορές εμφανίζονται διαδοχικά πέντε συνεχόμενες λέξεις (5-Grams).

- Χρησιμοποιείστε τα πιο κάτω δεδομένα

  http://www.cs.ucy.ac.cy/courses/EPL448/labs/LAB02/dataset.zip

Task2: Anagram

- Ένας αναγραμματισμός είναι ο σχηματισμός λέξης με μετάθεση των γραμμάτων μιας άλλης λέξης
- Π.χ
  - Refills→fillers
  - Relayed→layered
  - Rentals→antlers
  - Rebuild→builder
- Βρείτε τους αναγραμματισμούς στο αρχείο: http://www.puzzlers.org/pub/wordlists/unixdict.txt
- Ανεβάστε το UNIX dictionary στο HDFS
  - hadoop fs -copyFromLocal /usr/share/dict/words input

Task2: Anagram

- Κάποια αποτελέσματα της διαδικασίας reduce:

```
2  hasn't,shan't
2  cascara,caracas
2  ramada,armada
2  drawback,backward
2  bacterial,calibrate
2  bandpass,passband
2  aboard,abroad
2  wabash,bashaw
3  banal,laban,nabla
```

Submission

- Implement each task as a separate eclipse project
  - In total you need to implement 3 eclipse projects
- Zip each eclipse project separately along with its result file (part-r-00000)
- Submit three zip files to Moodle by Thursday 8th of February @ 09:00 am (morning)