CS4242: PROGRAMMING ASSIGNMENT 2

Inference of User Interests from Multiple Social Networks

Due date: 21 Mar 2016 (Mon, 1800 Hrs)

Presentation Date: 22 Mar (Tue), 20 mins per group

Total Marks: 14 (instead of 11)

Objectives:

Modeling user interests will benefit many services and applications, such as adaptive E-learning, target advertisement and personalized services. On the other hand, user' online behaviors on the prominent social platforms, such as Facebook, Twitter and LinkedIn, are mostly interest-driven, which essentially reflect users' personal preferences and interests. Moreover, users are connected and organized by multiple social platforms, which characterize users from different angles. For example, Facebook, Twitter and LinkedIn emphasize social connection, information exchange and professional archive, respectively. Therefore, aggregating and exploring a user's footprint from multiple social platforms is a promising way to harvest a more comprehensive view of users.

This assignment aims to design an effective multi-label multi-source classifier and then infer interests of users based on a given dataset collected from multiple social networks. There are several challenges, such as pre-processing of raw data, extracting and aggregating suitable informative features from multiple sources, and designing effective classifier.

What You Need to Do

You need to implement a multi-label multi-source classifier based on the user-generated-content and then apply it on the user-generated-content (UGC) from multiple social networks for interest inference. It should incorporate the following functions:

1. **Basic Requirements:**
    - You need to implement the data laundry and feature extraction parts for obtaining suitable contents from different formatted raw data (e.g., pure text, web page sources and json files) and extracting suitable informative features of each source via any machine learning tools.
    - You need to implement any suitable effective machine learning techniques or modify the given basic multi-label single-source linear regression program via late fusion to implement your multi-label multi-source classifier.
    - You need to implement the required evaluation metrics (P@K and S@K) and tabulate the overall results to demonstrate the effectiveness of each source and the resulting classifier.
    - It is noted that your classifier should be able to classify a user's interests into 20 given interest categories based on his/her contents from Facebook, Twitter and LinkedIn. Notably one user may have multiple interests.
2. **Enhanced Requirements:**
    - You can appropriately combine features from different sources via early fusion.

- You can implement advanced multi-source multi-label models, such as multi-label SVM and even multi-source classification.
- One key aspect of testing is to systematically demonstrate the effectiveness of different combinations of multiple sources in the classifier compared to single sources.

What You are Given:

1. **Dataset**

- The list of 20 predefined interests includes music, marketing, entrepreneurship, blogging and new technology, which can be indexed by the number of lines in the interest.txt file.
- The training dataset consists of 420 users indexed from ‘U1’ to ‘U420’, and corresponding ground truth (420 rows and 20 columns) with the respect to the 20 interests, where the row and column respectively represent user and interest, and 1 denotes the user has the interest, and 0 otherwise. You can utilize the dataset to design and tune your classifier.
- The test dataset consists of 150 users indexed from ‘U421’ to ‘U570’, and the corresponding ground truth (150 rows and 20 columns). You can utilize the dataset to evaluate your systems.
- All the aforementioned dataset and codes are available via google drive: https://drive.google.com/folderview?id=0BzCduZQhBlNyWFlhc3BMODIzMmM&usp=sharing.
- During online demo and testing, we will give you data for additional 20 users from which you need to extract the interests of these users online.

2. **Basic Codes**

- The basic framework implements the basic multi-label single-source linear regression without the bias term, and the evaluation metric S@K.

3. **Tools**

- Some tools for machine learning and feature extraction (but you can use whatever you like to extract features and train classifiers):
    1. Multi-label SVM https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/multilabel/
       Multi-view Learning http://kdd16.droppages.com/
    2. LDA http://nlp.stanford.edu/software/tmt/tmt-0.4/
       Word2Vec https://code.google.com/archive/p/word2vec/
       Sentence2Vec https://github.com/klb3713/sentence2vec

Presentation and Online Testing

- You will need to present your work within a 20-min session, including question answering, during which you will present your work using PPT and demonstrate the effectiveness of your software on your Notebook.
- Testing dataset will be used for testing during online evaluation.
- The roles of all members must be well defined and each must understand all aspects of project.

Report

You need to submit before the deadline:

- A report of not more than 8-pages. It should include program structure, details of pre-processing, feature extraction, classifier, training and validation procedures. You also need to include tabulated results of testing, demonstrating effectiveness of your classifier.
- A short PPT file (for about 8 mins of presentation) that includes sufficient details for the instructors to understand the details of your program and testing.
- Source codes of your implementation.

Remarks:

(a) Techniques, flexibility and effectiveness of system is the most important; UI should be functional and hence do not spend too much time on refining UI.

(b) All members are required to present some aspects of the system.

(c) Extra marks will be given for excellent assignments.

Consultation:

For questions regarding this assignment, please consult:

- Mr. Wang Xiang (xiangwang1223@gmail.com)

****Late Submission Policy**

We impose the following penalties for late submissions. (a) Late but within 24 hours: 25% reduction in grades. (b) Late but within 3 days: 50% reduction in grades. (c) After 3 days: zero marks.