CSE 515 Multimedia and Web Databases

Phase #2

(Due October 25th 2020, midnight)

Description: In this project, you will experiment with

- dimensionality reduction,
- unsupervised learning, and
- time series.

This project will build on the deliverables of the previous phase. Consider the data that was provided to you in the previous phase.

- **Task 0a:** Implement a program which creates a *gesture words* dictionary: Given a gestures folder that contains X, Y, Z, and W subfolders, a window length, w, a shift length, s, and a resolution, r,

  1. for each data file, f_i, the program creates a $f_i.wrd$ file containing the following:
     (a) for each component $c_l \in \{X, Y, Z, W\}$
        i. outputs the component ID, c_l into the output file
        ii. for each sensor s_j of f_i in the component c_l,
           A. outputs the sensor id s_j into the output file,
           B. computes and outputs, into the file, the average amplitude, $avg_{i,j}$, of the values,
           C. computes and outputs, into the file, the standard deviations, $std_{i,j}$, of the values,
           D. normalizes the entries to values between -1.0 and 1.0 (if the values are all the same, the values are normalized to 0.0)
           E. quantizes the entries into $2r$ levels as described in Phase 1,
           F. moves a w-length window on the corresponding time series (by shifting it s time units at a time), and at position h
           G. computes and outputs, into the file, the average quantized amplitude, $avgQ_{i,j,h}$, for the window h of sensor s_j
           H. computes and outputs, into the file, the symbolic quantized window descriptor, $winQ_{i,j,h}$ for the window h of sensor s_j

  IMPORTANT: *For the remaining tasks, the dictionary of the words consists of* $\langle$ *component Name, sensor ID* , $winQ$ $\rangle$ *triples.*

- **Task 0b:** Implement a program which, given a directory, dir, associates to each sensor s_j in each gesture file f_i two *gesture vectors*, based on

  – TF values (results for all sensors for the given gesture are written into a single output file, $tf\_vectors\_f_i.txt$)

  – TF-IDF values (results for all sensors for the given gesture are written into a single output file, $tfidf\_vectors\_f_i.txt$)

- **Task 1** Implement a program which, given a set of gesture files, a user selected vector model, and a k, identifies and reports the top-k latent semantics/topics, using
  - *user option* $\#1$*:* PCA,
  - *user option* $\#2$*:* SVD,
  - *user option* $\#3$*:* NMF,
  - *user option* $\#4$*:* LDA.

  You can use Matlab packages to compute these. The latent topics need to be presented in the form of $\underline{\langle word, score \rangle}$ pairs sorted in non-increasing order of scores.
- **Task 2:** Implement a program which, given a gesture file and a vector model, finds and ranks the 10 most similar gestures in the gesture database, relying on the following:
  - *user option* $\#1$*:* dot product of gesture vectors,
  - *user options* $\#2, 3, 4, 5$*:* top-k latent sensor semantics (PCA, SVD, NMF, LDA),
  - *user option* $\#6$*:* edit distance on symbolic quantized window descriptors (propose an edit cost function among symbols),
  - *user option* $\#7$*:* DTW distance on average quantized amplitudes

  Results are presented in the form of $\underline{\langle gesture, score \rangle}$ pairs sorted in non-increasing order of similarity scores.
- **Task 3: Latent Gesture Discovery Tasks**
  - **Task 3a:** Implement a program which, given a value p,
    1. creates a *gesture-gesture* similarity matrix using
       - *user option* $\#1$*:* dot product of gesture vectors,
       - *user options* $\#2, 3, 4, 5$*:* top-k latent sensor semantics (PCA, SVD, NMF, LDA),
       - *user option* $\#6$*:* edit distance on symbolic quantized window descriptors (propose an edit cost function among symbols),
       - *user option* $\#7$*:* DTW distance on average quantized amplitudes
    2. performs SVD on this *gesture-gesture* similarity matrix, and
    3. reports the top-p principle components (in terms of gesture membership) underlying this *gesture-gesture* similarity matrix

    Results are presented in the form of $\underline{\langle gesture, score \rangle}$ pairs sorted in non-increasing order of similarity scores.
  - **Task 3b:** Implement a program which, given a value p,
    1. creates a *gesture-gesture* similarity matrix using
       - *user option* $\#1$*:* dot product of gesture vectors,
       - *user options* $\#2, 3, 4, 5$*:* top-k latent sensor semantics (PCA, SVD, NMF, LDA),
       - *user option* $\#6$*:* edit distance on symbolic quantized window descriptors (propose an edit cost function among symbols),
       - *user option* $\#7$*:* DTW distance on average quantized amplitudes

2. performs NMF on this *gesture-gesture* similarity matrix, and
3. reports the top-p latent semantics (in terms of gesture membership) underlying this *gesture-gesture* similarity matrix

Results are presented in the form of $\underline{\langle gesture, score \rangle}$ pairs sorted in non-increasing order of similarity scores.

- **Task 4: Latent Gesture Clustering and Analysis Tasks**
  - **Task 4a:** Implement a program which considers top-p latent semantics of the gestures obtained using **Task 3a** and partitions the gestures into p groups based on their degree of membership to these p semantics.
  - **Task 4b:** Implement a program which considers top-p latent semantics of the gestures obtained using **Task 3b** and partitions the gestures into p groups based on their degree of membership to these p semantics.
  - **Task 4c:** Implement a program which considers a *gesture-gesture* similarity matrix obtained using
    - *user option* $\#1$*:* dot product of gesture vectors,
    - *user options* $\#2, 3, 4, 5$*:* top-k latent sensor semantics (PCA, SVD, NMF, LDA),
    - *user option* $\#6$*:* edit distance on symbolic quantized window descriptors (propose an edit cost function among symbols),
    - *user option* $\#7$*:* DTW distance on average quantized amplitudes

    and clusters the gestures into p groups using a k-means based technique (you will implement your own k-means algorithm).
  - **Task 4d:** Implement a program which considers a *gesture-gesture* similarity matrix obtained using
    - *user option* $\#1$*:* dot product of gesture vectors,
    - *user options* $\#2, 3, 4, 5$*:* top-k latent sensor semantics (PCA, SVD, NMF, LDA),
    - *user option* $\#6$*:* edit distance on symbolic quantized window descriptors (propose an edit cost function among symbols),
    - *user option* $\#7$*:* DTW distance on average quantized amplitudes

    and clusters the gestures into p groups using Laplacian-based spectral clustering technique (you will implement your own spectral clustering algorithm).

Deliverables:

- Your code (properly commented) and a README file.
- Your outputs for the provided sample inputs.
- A short report describing your work and the results.

Please place your code in a directory titled "Code", the outputs to a directory called "Outputs", and your report in a directory called "Report"; zip or tar all off them together and submit it through the digital dropbox.