CSE 515 Multimedia and Web Databases

Phase #1

(Due October 20th 2021, midnight)

Description: In this project, you will experiment with

- image features,
- vector models,
- dimensionality curse,
- graph analysis.

NOTES:

- In this phase, images to be inserted into the database will be labeled as `image-X-Y-Z.png`, where
  - $X \in \{cc, con, detail, emboss, jitter, neg, noise1, noise2, original, poster, rot, smooth, stipple\}$ denotes the type of the image,
  - $1 \leq Y \leq 40$ denotes the subject ID, and
  - $1 \leq Z \leq 10$ denotes the image sample ID.
- The tasks in this phase involve the three feature models and similarity/distance functions developed in the previous phase.
- You can use existing libraries LDA decomposition.
- You can use existing libraries for eigenvector and eigenvalue extraction.

PROJECT TASKS:

- **Task 1:** Implement a program which (a) given one of the three feature models, (b) a user specified value of X, a user specified value of k, (d) one of the four dimensionality reduction techniques (PCA, SVD, LDA, k-means) chosen by the user, reports the top-k latent semantics extracted using images of this type. Each latent semantic should be presented in the form of a list of subject-weight pairs, ordered in decreasing order of weights. Store the latent semantics in a properly labeled output file.
- **Task 2:** Implement a program which (a) given one of the three feature models, (b) a user specified value of Y, a user specified value of k, (d) one of the three dimensionality reduction techniques (PCA, SVD, LDA,k-means) chosen by the user, reports the top-k latent semantics extracted using images of this subject. Each latent semantic should be presented in the form of a list of type-weight pairs, ordered in decreasing order of weights. Store the latent semantics in a properly labeled output file.
- **Task 3** Implement a program which, (a) given one of the three feature models and (b) a value k,
  - creates (and saves) a type-type similarity matrix,

  - performs a user selected dimensionality reduction technique (PCA, SVD, LDA,k-means) on this type-type similarity matrix, and
  - reports the top-k latent semantics.

  Each latent semantic should be presented in the form of a list of type-weight pairs, ordered in decreasing order of weights. Store the latent semantics in a properly labeled output file.

- **Task 4** Implement a program which, (a) given one of the three feature models and (b) a value k,
  - creates (and saves) a subject-subject similarity matrix,
  - performs a user selected dimensionality reduction technique (PCA, SVD, LDA,k-means) on this subject-subject similarity matrix, and
  - reports the top-k latent semantics.

  Each latent semantic should be presented in the form of a list of subject-weight pairs, ordered in decreasing order of weights. Store the latent semantics in a properly labeled output file.

- **Task 5** Implement a program which, given the filename of a query image which may not be in the database and a latent semantics file, identifies and visualizes the most similar n images under the selected latent semantics.

- **Task 6** Implement a program which, given the filename of a query image which may not be in the database and a latent semantics file, associates a type label (X) to the image under the selected latent semantics.

- **Task 7** Implement a program which, given the filename of a query image which may not be in the database and a latent semantics file, associates a subject ID (Y) to the image under the selected latent semantics.

- **Task 8** Implement a program which (a) given a subject-subject similarity matrix, (b) a value n, and (c) a value m,
  - creates a similarity graph, $G(V, E)$, where V corresponds to the subjects in the database and E contains node pairs v_i, v_j such that, for each subject v_i, v_j is one of the n most similar subjects in the database
  - identifies the most significant m subjects in the collection using ASCOS++ measure. See

    Hung-Hsuan Chen and C. Lee Giles. 2015. ASCOS++: An Asymmetric Similarity Measure for Weighted Networks to Address the Problem of SimRank. ACM Trans. Knowl. Discov. Data 10, 2, Article 15 (October 2015).

- **Task 9** Implement a program which (a) given a subject-subject similarity matrix, (b) a value n, (c) a value m, and three subject IDs
  - creates a similarity graph, $G(V, E)$, where V corresponds to the subjects in the database and E contains node pairs v_i, v_j such that, for each subject v_i, v_j is one of the n most similar subjects in the database
  - identifies the most significant m subjects (relative to the input subjects) using personalized PageRank measure. See

    Huang, S., Li, X., Candan, K. S., Sapino, M. L. (2016). Reducing seed noise in personalized PageRank. Social Network Analysis and Mining, 6(1), 1-25.

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.