COMS4721 HW5

Jeff Hudson (jdh2182)

Tuesday, April 28, 2015

Problem 1 (Markov Chains)

In this problem, you will rank 759 college football teams based only on the scores of every game in the 2014 season. Construct a 759×759 random walk matrix M on the college football teams. After processing all games, let M be the matrix formed by normalizing the rows of $\hat{M}$ so they sum to 1. Let w_t be the 759×1 state vector at step t. Set w_0 to the uniform distribution. Therefore, w_t is the distribution of the state after t steps given that the starting state at time 0 is uniformly distributed. Use w_t to rank the teams by sorting in decreasing value according to this vector.

List the top 20 teams and their corresponding values in w_t for $t = 10, 100, 200, 1000$.

```
## [1] "t = 10"
##       Team                Weight
##  [1,] "UW-Whitewater"     "0.015"
##  [2,] "MountUnion"        "0.013"
##  [3,] "ColoradoSt-Pueblo" "0.01"
##  [4,] "OhioState"         "0.009"
##  [5,] "Linfield"          "0.009"
##  [6,] "MinnSt-Mankato"    "0.008"
##  [7,] "Wartburg"          "0.008"
##  [8,] "Wesley"            "0.007"
##  [9,] "SouthernOregon"    "0.007"
## [10,] "Oregon"            "0.007"
## [11,] "Alabama"           "0.007"
## [12,] "NorthDakotaSt"     "0.007"
## [13,] "TCU"               "0.006"
## [14,] "MaryHardin-Baylor" "0.006"
## [15,] "FloridaSt"         "0.006"
## [16,] "Hobart"            "0.006"
## [17,] "JohnCarroll"       "0.006"
## [18,] "MarianIN"          "0.006"
## [19,] "Widener"           "0.006"
## [20,] "Concord"           "0.006"

## [1] "t = 100"
##       Team                Weight
##  [1,] "UW-Whitewater"     "0.029"
##  [2,] "OhioState"         "0.028"
##  [3,] "Oregon"            "0.023"
##  [4,] "Alabama"           "0.019"
##  [5,] "TCU"               "0.018"
##  [6,] "MountUnion"        "0.017"
##  [7,] "FloridaSt"         "0.015"
##  [8,] "MichiganSt"        "0.013"
##  [9,] "ColoradoSt-Pueblo" "0.013"
## [10,] "SouthernOregon"    "0.012"
## [11,] "Baylor"            "0.012"
```

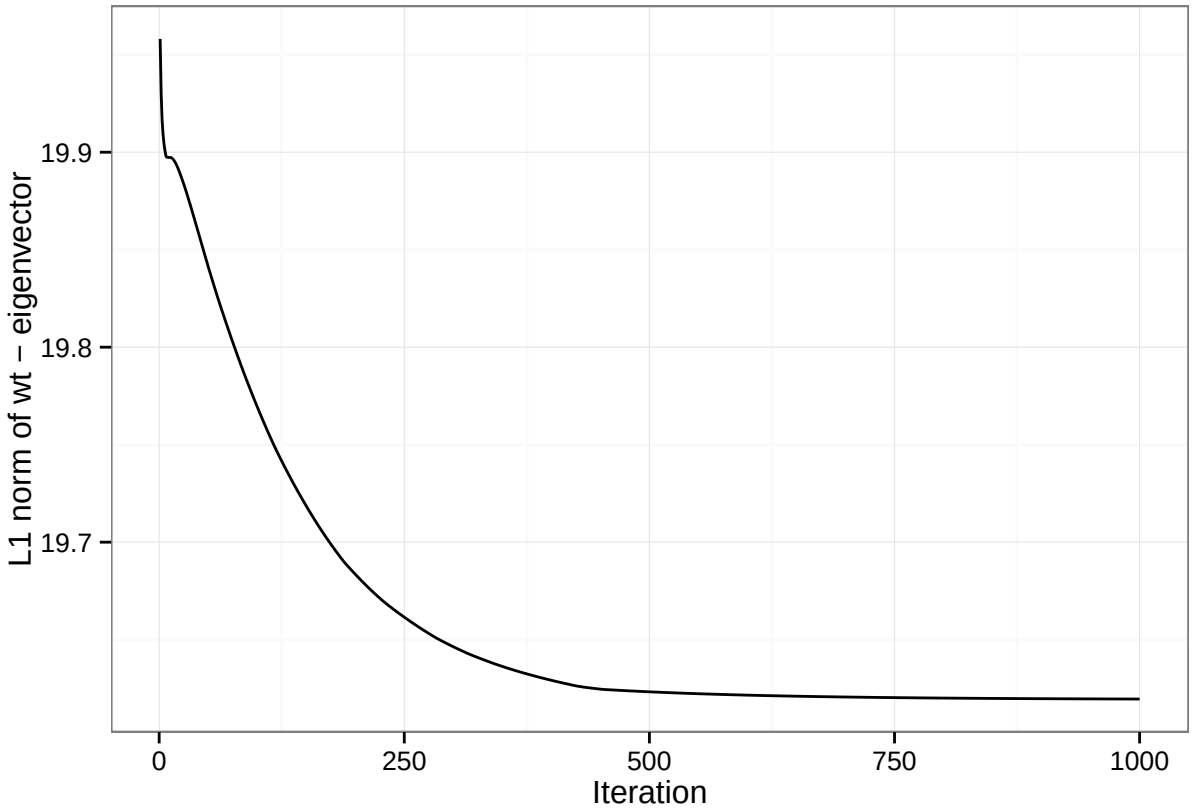
```
## [12,] "GeorgiaTech"       "0.012"
## [13,] "Wartburg"          "0.011"
## [14,] "UCLA"              "0.011"
## [15,] "Mississippi"       "0.01"
## [16,] "CarrollMT"         "0.01"
## [17,] "Georgia"           "0.01"
## [18,] "Arizona"           "0.01"
## [19,] "ArizonaSt"         "0.009"
## [20,] "MississippiSt"     "0.009"
```

```
## [1] "t = 200"
##       Team            Weight
##  [1,] "OhioState"     "0.036"
##  [2,] "Oregon"        "0.03"
##  [3,] "Alabama"       "0.024"
##  [4,] "TCU"           "0.024"
##  [5,] "FloridaSt"     "0.019"
##  [6,] "UW-Whitewater" "0.018"
##  [7,] "MichiganSt"    "0.017"
##  [8,] "Baylor"        "0.016"
##  [9,] "GeorgiaTech"   "0.015"
## [10,] "UCLA"          "0.014"
## [11,] "Mississippi"   "0.013"
## [12,] "Georgia"       "0.013"
## [13,] "Arizona"       "0.013"
## [14,] "ArizonaSt"     "0.012"
## [15,] "MississippiSt" "0.012"
## [16,] "Missouri"      "0.011"
## [17,] "Clemson"       "0.01"
## [18,] "SouthernCal"   "0.01"
## [19,] "Wisconsin"     "0.01"
## [20,] "Auburn"        "0.01"
```

```
## [1] "t = 1000"
##       Team            Weight
##  [1,] "OhioState"     "0.047"
##  [2,] "Oregon"        "0.039"
##  [3,] "Alabama"       "0.032"
##  [4,] "TCU"           "0.031"
##  [5,] "FloridaSt"     "0.025"
##  [6,] "MichiganSt"    "0.022"
##  [7,] "Baylor"        "0.021"
##  [8,] "GeorgiaTech"   "0.02"
##  [9,] "UCLA"          "0.018"
## [10,] "Mississippi"   "0.017"
## [11,] "Georgia"       "0.017"
## [12,] "Arizona"       "0.017"
## [13,] "ArizonaSt"     "0.015"
## [14,] "MississippiSt" "0.015"
## [15,] "Missouri"      "0.014"
## [16,] "Clemson"       "0.014"
## [17,] "SouthernCal"   "0.014"
## [18,] "Wisconsin"     "0.013"
## [19,] "Auburn"        "0.013"
```

```
## [20,] "Utah"          "0.013"
```

We saw that w_∞ corresponds to the first eigenvector of M^T: $M^T w_\infty = w_\infty \Rightarrow M^T u_1 = \lambda_1 u_1, \lambda_1 = 1$ and $w_\infty = u + 1/\Sigma_j u_1(j)$ (since by convention $u_1^T u_1 = 1$). Plot $||w_t - u_1/\Sigma_j u_1(j)||_1$ as a function of t for $t = 1, \ldots, 1000$.



What is $||w_{1000} - u_1/\Sigma_j u_1(j)||_1$? 19.6195414

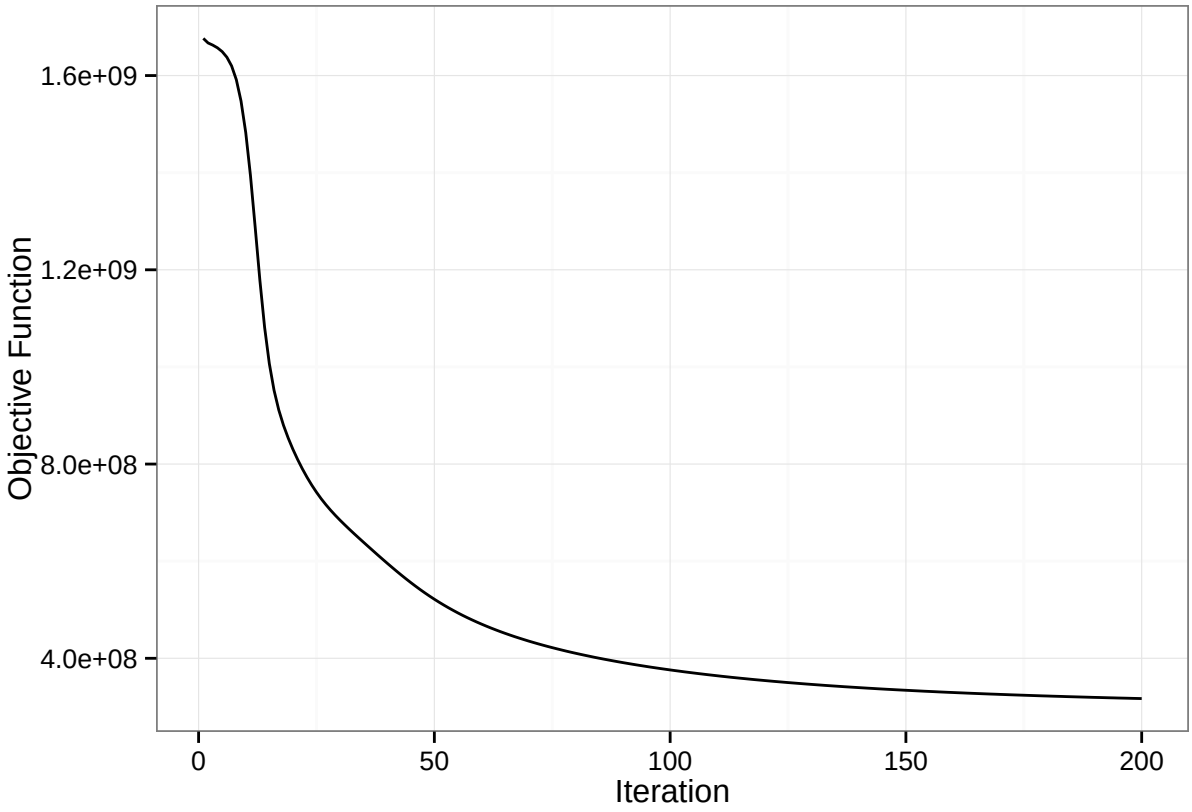
Problem 2 (Nonnegative matrix factorization)

In this problem you will factorize a $n \times m$ matrix X into a rank-K approximation WH, where W is $n \times K$, H is $K \times m$ and all values in the matrices are nonnegative. Each value in W and H can be initialized randomly, e.g., from a $Uniform(0, 1)$ distribution. (See a hint about the implementation below.)

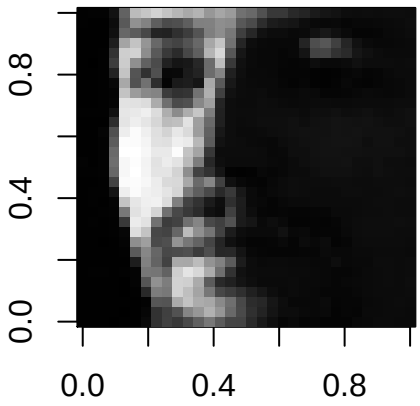
Part 1: The data to be used for Part 1 consists of 1000 images of faces, each originally 32×32, but vectorized to length 1024. The data matrix is therefore 1024×1000.

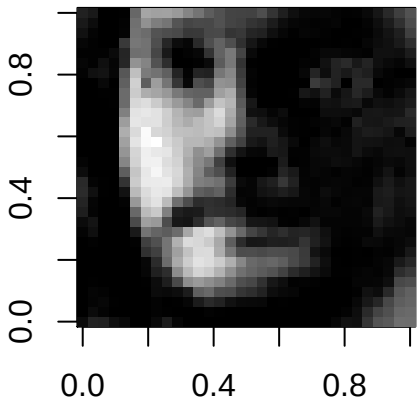
Implement and run the NMF algorithm on this data using the Euclidean penalty. Set the rank of the factorization to 25 and run for 200 iterations.

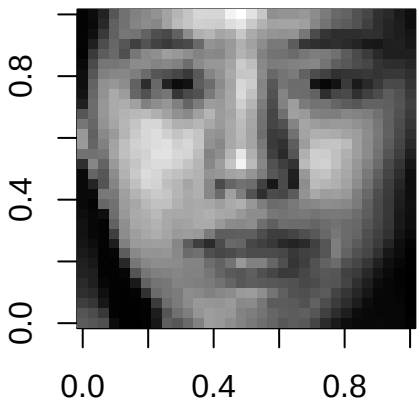
Plot the objective as a function of iteration.

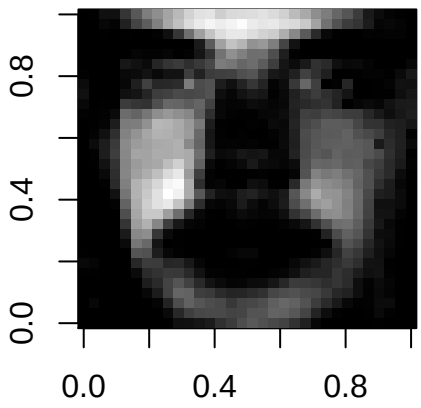


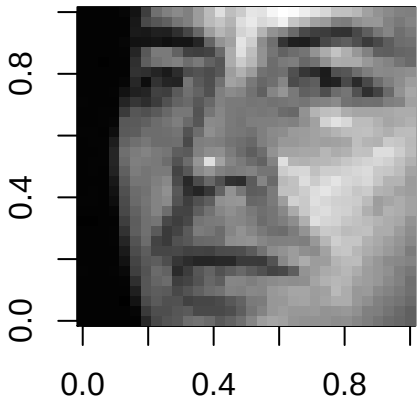
For 3 images in the data set, show the original image and the column of W for which the corresponding weight in H is the largest. This should be shown as two 32×32 images. Select the three images so that the columns displayed from W are different.

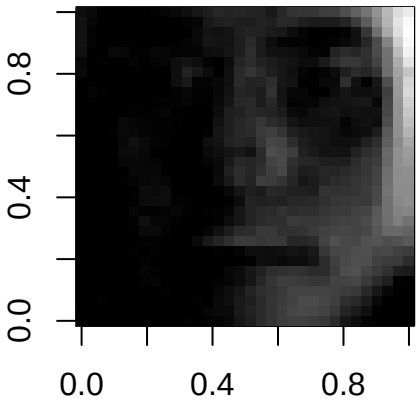






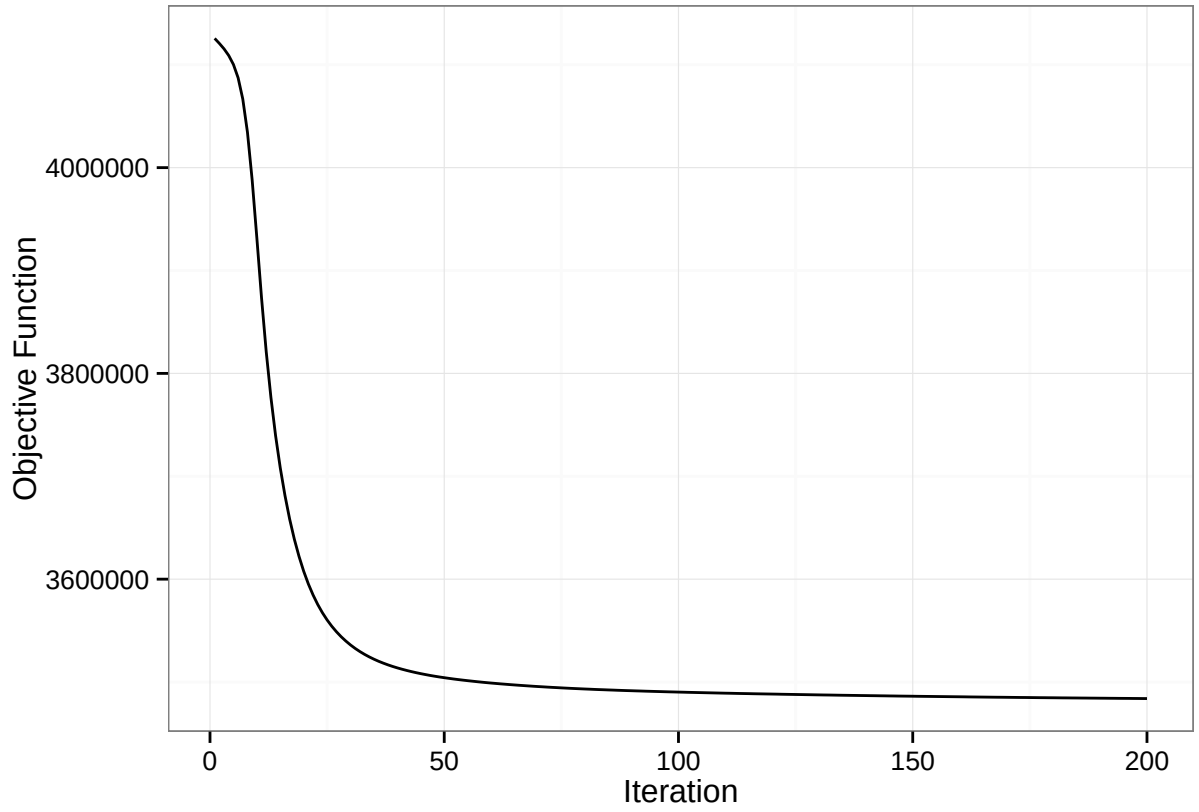






Part 2: The data to be used for Part 2 consists of 8447 documents from The New York Times. (See below for how to process the data.) The vocabulary size is 3012 words. You will need to use this data to constitute the matrix X, where X_{ij} is the number of times word i appears in document j. Therefore, X is 3012×8447 and most values in X will equal zero.
Implement and run the NMF algorithm on this data using the divergence penalty. Set the rank (i.e., number of topics) to 25 and run for 200 iterations.

Plot the objective as a function of iteration.



After running the algorithm, normalize the columns of W so they sum to one. Pick 5 columns of W. For each selected column show the 10 words having the largest probability according to the values in W and give the probabilities. The i th row of W corresponds to the i th word in the "dictionary" provided with the data.

```
## [1] "Topic 17"
##  Word       Probability
##  "law"      "0.0198"
##  "court"    "0.0138"
##  "case"     "0.0134"
##  "rule"     "0.0118"
##  "issue"    "0.0115"
##  "state"    "0.0109"
##  "official" "0.0108"
##  "legal"    "0.0102"
##  "lawyer"   "0.0101"
##  "decision" "0.0086"

## [1] "Topic 8"
##  Word         Probability
##  "food"       "0.017"
##  "serve"      "0.0091"
##  "restaurant" "0.0088"
##  "water"      "0.0088"
##  "eat"        "0.0079"
##  "dry"        "0.0078"
##  "taste"      "0.0077"
```

```
##  "fresh"      "0.0075"
##  "pound"      "0.0075"
##  "add"        "0.007"
```

```
## [1] "Topic 1"
##  Word      Probability
##  "hour"    "0.0262"
##  "travel"  "0.0164"
##  "trip"    "0.014"
##  "visit"   "0.0107"
##  "morning" "0.0104"
##  "hotel"   "0.0103"
##  "train"   "0.009"
##  "service" "0.0088"
##  "car"     "0.0088"
##  "phone"   "0.0087"
```

```
## [1] "Topic 12"
##  Word       Probability
##  "war"      "0.0246"
##  "military" "0.0204"
##  "american" "0.0174"
##  "force"    "0.0154"
##  "states"   "0.0131"
##  "attack"   "0.0105"
##  "troop"    "0.0103"
##  "soldier"  "0.0091"
##  "country"  "0.0086"
##  "official" "0.0086"
```

```
## [1] "Topic 19"
##  Word       Probability
##  "father"   "0.059"
##  "son"      "0.0492"
##  "mother"   "0.048"
##  "mrs"      "0.0463"
##  "daughter" "0.0383"
##  "graduate" "0.0304"
##  "marry"    "0.0253"
##  "family"   "0.0239"
##  "wife"     "0.0203"
##  "receive"  "0.0198"
```