Exercise Sheet for the Lecture
Intelligent Learning and Information Systems

Summer Term 2017

10 - Clustering and Text Mining

programming exercise

electronic submission (strict): July 21, 2017 (12:00)

Joshua Chen (joshua.chen@uni-bonn.de)
Sergey Kuptsov (sergeykuptsovde@gmail.com)

This assignment is concerned with *clustering*. Submissions are due **electronically** and each group has to give a presentation (PowerPoint, PDF,...) in the exercise. The presentation cannot be longer than 5-6 minutes.

1 - Clustering handwritten digits [60 points]

- Download the *Semeion Handwritten Digit Data Set* from the UCI machine learning repository[*].
- Depending on the modulo-8 residue of your group number, cluster the digits in one of the following ways:

modulo-8 residue of your group number	your clustering algorithm and the number of clusters
0	k-means with k=10
1	k-means with k=30
2	Single Link, with k=10 clusters
3	Single Link, with k=30 clusters
4	Complete Link, with k=10 clusters
5	Complete Link, with k=30 clusters
6	Average Link, with k=10 clusters
7	Average Link, with k=30 clusters

Use the *Euclidean distance* as distance measure.

- Prepare a slide with the “confusion matrix” of your clustering: each row of your matrix should correspond to one of your clusters, while the columns should correspond to the digits $0, 1, \ldots, 9$. The entry $e_{i,j}$ (i.e., the value in the i-th row and j-th column of the matrix) should indicate how many images in the i-th cluster display the j-th digit.
- Select one of your clusters in which images of many different digits are present. Plot some representative images of this cluster on another slides.

[*] http://archive.ics.uci.edu/ml/datasets/semeion+handwritten+digit

- Discuss and compare the solutions in the class: What are the similarities and differences between your clustering and the clusterings of the other groups? Is the clustering with $k = 30$ a "refined" version of the clustering produced with $k = 10$? Which is the "best" clustering algorithm for this application?...

2 - Clustering of texts [optional, 60 bonus points]

- Determine the *language of your group* (*LG*) as follows:

  if all the members of your group speak a language L that is neither English nor German **then** $LG = L$;

  otherwise :

  > **if** all the members of the group speak German **then** $LG =$'German';
  >
  > **otherwise** : $LG =$'English'.

- Consider the following topics: *sports*, *politics*, *culture* and *science*. From each of these topics, download 10 documents from the internet that were written in the language of your group. (That is: as 4 topics are given, you are expected to download $4 \times 10 = 40$ documents in total.)

- Prepare the *binary* bag-of-words representation of your documents. Each document is represented as a point in a high-dimensional vector space. Each dimension of the vector space corresponds to one of the words in W, where W is the set of all the words that appear in *at least one* of the documents that you have downloaded. Consider the vector $v = (v_1, ...v_{|W|})$ corresponding to document d: $v_i = 1$ if the word corresponding to the i-th dimension of the vector space appears in document d. Otherwise: $v_i = 0$. For simplicity, you *can* treat syntactically different forms of the same word as different words, for example, you can treat "computer", "Computer" and "computers" as three different words.

- Prepare the *TFIDF* representation of your documents. For more details on TFIDF, see:
  `http://blog.christianperone.com/2011/09/machine-learning-text-feature-extraction-tf-idf-part-i/`
  and
  `http://blog.christianperone.com/2011/10/machine-learning-text-feature-extraction-tf-idf-part-ii/`.
  The above description assumes that you consider a supervised machine learning problem, i.e., your data is split into training and test sets. While preparing the TFIDF representations of your documents, you can assume that all of your documents belong to the training set.

- Find 4 clusters of documents with the AverageLink algorithm using the *Cosine distance* with the

  (a) binary bag-of-words representation, and

  (b) TFIDF representation of your documents.

- Does the clusters correspond to the topics? Similarly to Task 1, prepare a slide with the "confusion matrix" of your clustering.

- Discuss and compare the solutions in the class: in which cases are the texts clustered according to your expectations and when is the result of clustering counterintuitive? Up to what extent can the differences be attributed to working with different languages and how much do the solutions depend on the particular selection of texts?...