Statistical data analysis
Dimensionality Reduction

WS

Introduction

The goal of this tutorial is to get familiar with some basic methods for dimensionality reduction, complete you own implementation of the Isomap algorithm, experiment with its parameters and compare with other techniques of dimensionality reduction.

1 Background

The data you will be working with are vector representations of words in a latent (unknown) high-dimensional space. This representation of words, also known as *word embedding*, differs from standard bag-of-words (BoW, TFIDF, etc.) representations in that the meaning of the words is *distributed* across all the dimensions. Generally speaking, the word embedding algorithms seek to learn a mapping projecting the original BoW representation (simple word index into a given vocabulary) into a lower-dimensional (but still too high for our cause) continuous vector-space, based on their distributional properties observed in some raw text corpus. This distributional semantics approach to word representations is based on the basic idea that linguistic items with similar distributions typically have similar meanings, i.e. words that often appear in a similar context (words that surround them) tend to have similar (vector) representations. The overall transformation process from BoW representation to the word embeddings representation and its 2-dimensional visualization is outlined in figure 1.

Specifically, the data you are presented with are vector representations coming from the most popular algorithm for word embedding known as word2vec [1] by Tomas Mikolov (VUT-Brno alumni). Word2vec is a (shallow) neural model learning the projection of BoW word representations into a latent space by the means of gradient descend. Your task is to further reduce the dimensionality of the word representations to get a visual insight into what has been learned.

2 Data

You are given 300-dimensional word2vec vector embeddings in the file *data.csv* with corresponding word labels in *labels.txt* for each line. Each of these words comes

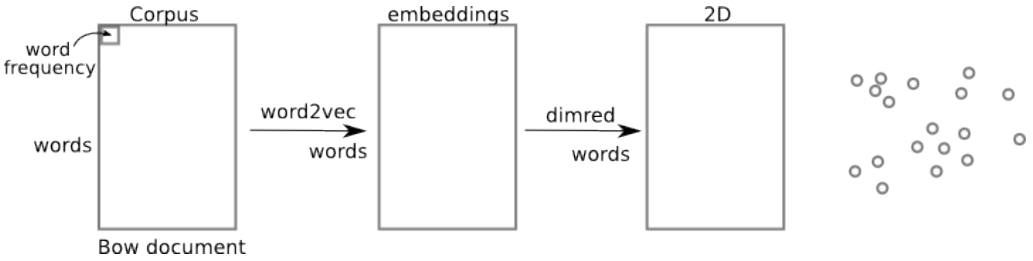

Figure 1: Process of transformation from BoW representation to the word embeddings representation and its reduction into 2-dimensional vector space.

from one of 10 selected classes of synonyms, which can be recognized (and depicted) w.r.t. labels denoted in the file *colors.csv*.

3 Tasks

1. **Load the dataset of 165 words**, each represented as a 300-dimensional vector. Each word is assigned to one of 10 clusters depending on its semantics.

   ```
   mydata <- read.csv('data.csv', header = FALSE)
   mylabels <- read.csv('labels.txt', header = FALSE)
   mycolors <- read.csv('colors.csv', header = FALSE)
   ```

   The data is in the matrix `mydata`, cluster assignment in `mycolors` and the actual words (useful for visualization) in `mylabels`. You can plot the data by using only the first 2 dimensions (see figure 2) by calling a following function:

   ```
   PlotPoints(mydata[,c(1,2)], mylabels, mycolors).
   ```

   The supporting code is provided in a file `reduction_task.R`.

2. **Implement ISO-MAP dimensionality reduction procedure. [2p]**

   - Determine the neighbors of each point [**0.5p**]
     - For simplicity, you can use `get.knn` method available in `FNN` package.
   - Construct the neighborhood graph (sparse matrix). [**1p**]
   - Compute shortest-paths (geodesic) matrix using your favourite algorithm. [**0.5p**]
     - Tip: Floyd-Warshall algorithm can be implemented easily here.
   - Project the geodesic distance matrix into 2D space with (Classical) Multidimensional Scaling (`cmdscale` functions in R).

   An expected result (for $k = 5$) should look similar (not necessarily exactly the same) to figure 3.

3. **Visually compare PCA, ISOMAP and t-SNE [1p]** by plotting the word2vec data, embedded into 2D using the `Plotpoints` function. Try finding the optimal k value for ISOMAP's nearest neighbour.

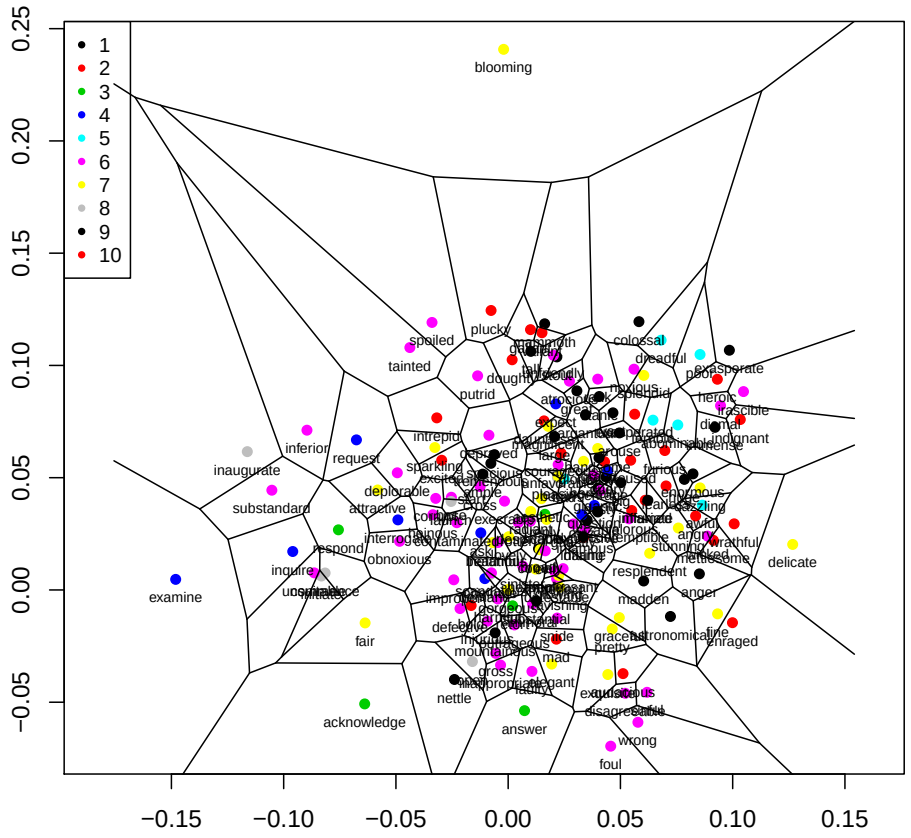

Figure 2: Random dimensionality reduction. Word placement does not match its semantics.

PCA is already provided to you as a part of the `stats` package. t-SNE method is a part of the `tsne` package
(see `https://cran.r-project.org/web/packages/tsne/tsne.pdf`).

4. **Classify the individual words with 10 classes defined by clusters/colors. [1p]**

   The supporting code in a function `Classify` performs training and testing of classification trees and gives the classification accuracy (percentage of correctly classified samples) as its result. Compare the accuracy of prediction on plain data, PCA, ISOMAP and t-SNE.

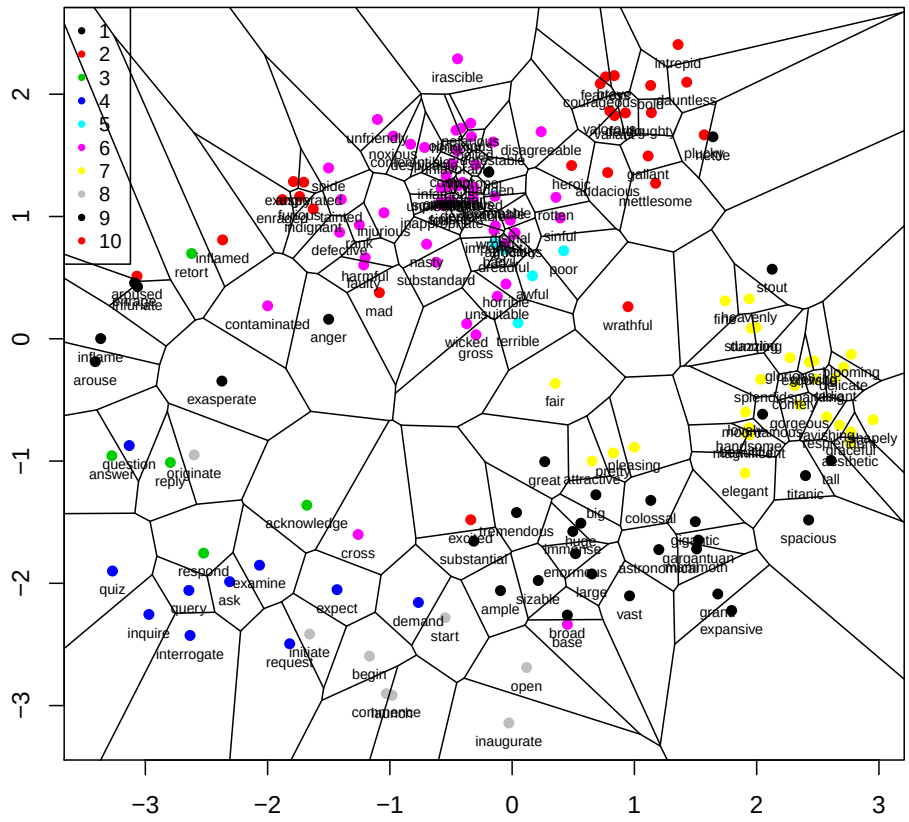

Figure 3: ISOMAP outcome. The first two dimensions much better capture word semantics.

4 Submission Form

Submit task as ZIP file containing all files necessary to run your code (try to test via Rscript reduction_task.R) and write a short report (2-3 pages). The report should contains a definition of the task, interesting implementation details, and your answers for the task number 3 and 4.

References

[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.