Computational topology - group project

Text analysis 2

December 3, 2020

In this project your goal is to use persistence diagrams as a tool for classifying different types of texts. The idea is to build simplicial complexes based on several features for texts from different domains and see if persistence diagrams of these simplicial complexes detect the difference between the domains. This is a toy project but you might want to try it out also on a larger data base, using different text statistics.

- **Data:** Pick three different types of English language texts available on Project Gutenberg site, for example novels, song lyrics, and dramas.

  Split each text into sentences and compute the embedding for each sentence using LASER, which can be freely downloaded from the internet. For every document compute the document embedding by averaging the embeddings of sentences in the document. For evey type of texts produce a matrix D^k ($k \in \{1, 2, 3\}$) such that the i-th column of the matrix D^k is the embedding of the i-th document of that type. Since document embeds into a high-dimensional vector space, the documents (columns) are represented by vector of very large dimension.

- **Dimension reduction:** To reduce dimension, use singular value decompositions (SVD) of the document-term matrices: factor the matrices into $D^k = U^k \times \Sigma^k \times (V^k)^T$ where Σ is the diagonal matrix with singular values of D^k on the diagonal in decreasing order, U^k has the left singular vector and V^k has the right singular vectors in its rows. The j-th row of the matrix V^k corresponds to the embedding of the document representing the j-th column in the matrix D^k.

  So the first r entries in the row of the matrix V^k represent an embedding of the corresponding document into r-dimensional space. Pick $r \leq 5$.

- **The model:** For each of the three domains build persistence diagrams on the r-dimensional data points representing documents in the following way. Compute the biggest distance R between any two of them, divide the interval $[0, R]$ into 10 subintervals of equal lengths with partition points $r_1 < r_2 < \cdots < r_{10}$ and compute the corresponding filtration of α-shapes complexes (it suffices to consider simplices up to dimension 2).

- **Conclusion:** Compare the resulting persistence diagrams in dimensions 0, 1 and 2 obtained from the different domains. Do they differ between the domains?

Repeat the experiment after first removing stop words (i.e. words with a high frequency and insignificant meaning) from the documents. You can either use the following list of 25 stop words:

a, an, and, are, as, at, be, by., for, from, has, he, in, is, it, its,
of, on, that, the, to, was, were, will, with

or the longer list for English.

In your opinion, do the persistence diagrams obtained in this way classify the different domains? Does removing stop words affect the success of the classification model?