University of Ljubljana
Faculty of Computer and Information Science

June 13, 2017

Computational topology

Text classification using persistent homology

Matija Čufar, Domen Keglevič

1. Introduction

Persistent homology is a method from topological data analysis which enables us to calculate and compare features of topological spaces at different spatial resolutions. Some features persist at wide range of spatial resolutions and are more likely to represent the underlying space. This approach has been applied to many different domains ranging from image analysis [3] to genome studies [1].

In this seminar assignment we try to use persistence homology for text classification. We pick four different text domains and try to distinguish them based on their persistence diagrams.

2. Methods

We have chosen texts from the following domains:

- Excerpts from the Old and New Testaments of the Bible,
- articles from `phys.org`,
- recipes from `allrecipes.com`.

For each of the domains, we picked ten texts, each at least 100 words long. We used the Gudhi [2], a library for topological data analysis, to compute persistent homology on the texts.

To compute the persistent homology of a data set, we first need to represent it as a simplicial, or some other kind of complex. We used the following two approaches to build simplicial complexes for each of the domains:

2.1. Feature based approach.

In our first approach we associated each text with a point in $\mathbb{R}^6$. Each dimension was representing one of text features. We used the following features:

- the ratio of (average word length)/(longest word length),
- the ratio of (average sentence length)/(longest sentence length),
- the ratio of the total number of three words with the highest tf-idf value among all the words,
- the ratio of the number of words of length ≤ 8 among all words,
- the ratio of the number of words of length ≥ 9 among all words.
- the ratio of (number of different words)/(number of all words)

Note that all features are normalized in order to avoid giving too much weight to those that would have large size otherwise. After computing features and mapping texts to points in $\mathbb{R}^6$ we built Alpha and Vietoris-Rips complexes on each of the domains.

2.2. **Distribution based approach.** Our second approach involved computing distributions of word and sentence lengths. This gives us two distributions for each text. Two distributions of the same type (word or sentence) can then be compared using the following distance measures:

- The Hellinger distance:

$$H(P,Q) = \sqrt{\frac{1}{2}\sum_{i=1}^{k}\left(\sqrt{p_i}-\sqrt{q_i}\right)^2},$$

- the Chi-squared distance:

$$\chi^2(P,Q) = \frac{1}{2}\sum_{i=1}^{n}\frac{(p_i-q_i)^2}{(p_i+q_i+\varepsilon)},$$

- the Euclidean distance:

$$E(P,Q) = \sqrt{\sum_{i=1}^{k}\left(p_i-q_i\right)^2},$$

where P and Q are the discrete distributions and p_i and q_i are the i-th bins of corresponding distributions. The ε in the Chi-squared distance is a small positive constant used to avoid dividing by zero.

We used these distances to compute distance matrix for each of the domains and used distance matrices to build Vietoris-Rips complexes.

2.3. **Domain comparison.** When the simplicial complexes were built, we calculated persistence diagrams for each complex and computed the bottleneck distances between them. We expected the distances would help us distinguish between the texts. For example, the distance between the texts taken from the Bible should be smaller than the rest.

3. Results

3.1. **Feature based approach.** Bottleneck distance matrix for the Alpha complex is shown in Table 1. Barcode plots are shown in Figure 1.

As we can see from the distance matrix, the method did not produce meaningful results. The first thing we notice is that all the distances are very small, with the highest being only 0.008. Other examples that show the ineffectiveness of this method are the fact that the Old Testament differs the most from the New Testament and the fact that according to this method, physics article abstracts are very similar to recipes.

Something we notice from the barcode plots is that *TODO: napisat neki o 4d zadevah v receptih*

Results for the Vietoris-Rips complex are similar to the results for Alpha complex.

3.2. **Distribution based approach.** *TODO: tekst*

	Old Testament	New Testament	phys.org	recipes
Old Testament	0.000	0.008	0.003	0.003
New Testament	0.008	0.000	0.008	0.008
phys.org	0.003	0.008	0.000	0.002
recipes	0.003	0.008	0.002	0.000

TABLE 1. The distance matrix calculated from the Alpha complexes.

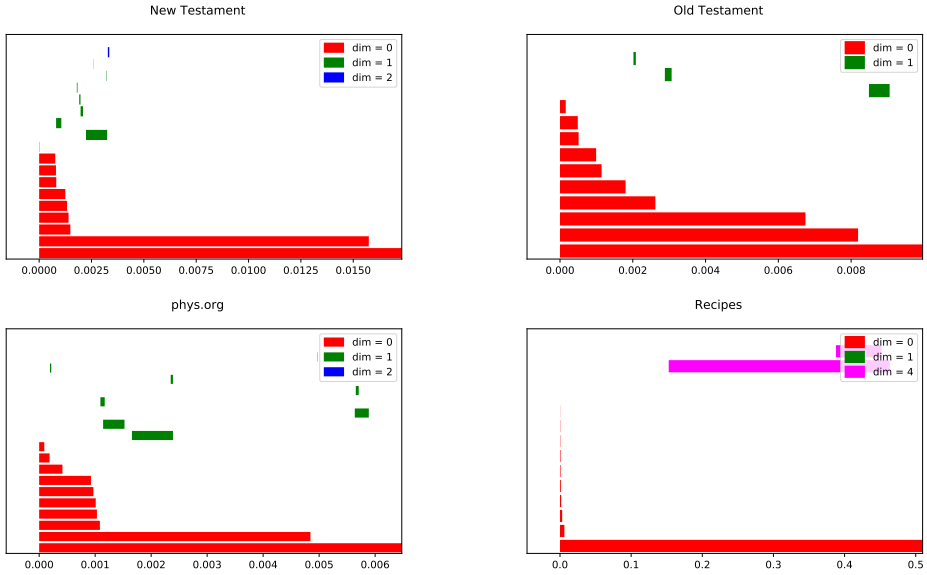

FIGURE 1. Persistence barcodes of the domains, calculated from the Alpha complexes.

	Old Testament	New Testament	phys.org	recipes
Old Testament	0.000	0.015	0.040	0.112
New Testament	0.015	0.000	0.033	0.103
phys.org	0.040	0.033	0.000	0.105
recipes	0.112	0.103	0.105	0.000

TABLE 2. The distance matrix calculated from the Vietoris-Rips complexes, built using the Hellinger distance between sentence length distribution.

4. Conclusion

As we saw in the previous section, the results are not very good.

5. Authors contributions

Both authors have contributed to most parts of the seminar work. The exception being that Domen prepared most of the presentation and Matija wrote most of the report. Detailed contributions can be seen in commit history of GitHub repository `https://github.com/dkegle/rt201617` which also contains all of the code and data.

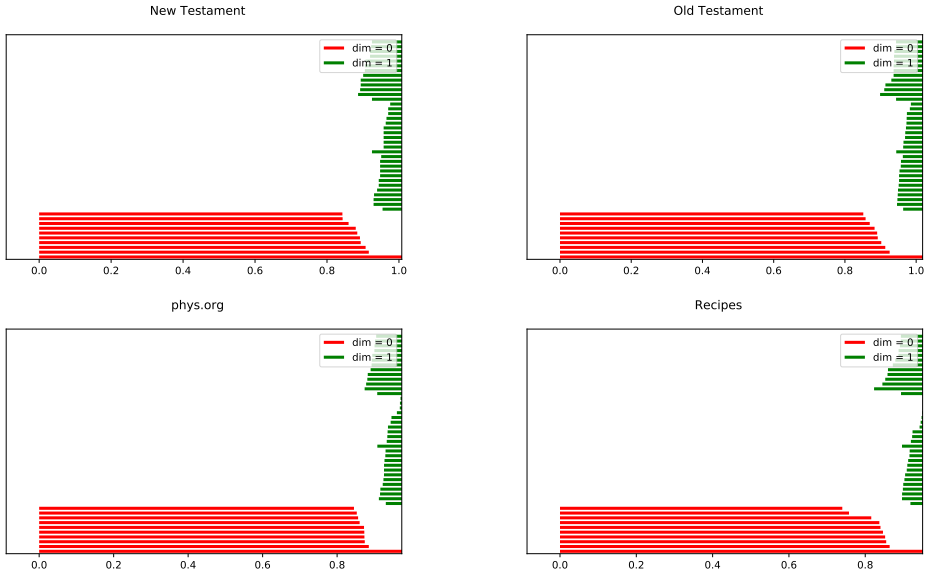

FIGURE 2. Persistence barcodes of the domains, calculated from the Vietoris-Rips complexes, built using the Hellinger distance between sentence length distributions.

REFERENCES

[1] Pablo G Camara, Daniel IS Rosenbloom, Kevin J Emmett, Arnold J Levine, and Raul Rabadan. Topological data analysis generates high-resolution, genome-wide maps of human recombination. *Cell Systems*, 3(1):83–94, 2016.

[2] Clément Maria, Jean-Daniel Boissonnat, Marc Glisse, and Mariette Yvinec. The gudhi library: Simplicial complexes and persistent homology. In *International Congress on Mathematical Software*, pages 167–174. Springer, 2014.

[3] Kazuaki Nakane, Akihiro Takiyama, Seiji Mori, and Nariaki Matsuura. Homology-based method for detecting regions of interest in colonic digital images. *Diagnostic pathology*, 10(1):36, 2015.