Computational topology

Text classification using persistent homology

Matija Čufar, Domen Keglevič

1. Introduction

TODO: nekaj o vztrajni homologiji in mogoče analizi teksta

In this report, we attempt to classify texts from four different domains by comparing their persistence diagrams.

2. Methods

We have chosen to attempt to classify texts from the following domains:

- Excrepts from the Old and New Testaments of the Bible,
- abstracts of 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 [1], 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 Alpha and Vietoris-Rips complexes.
Our first approach involved computing the following features for each of the texts:

- 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.

This gives us a point in $\mathbb{R}^5$ for each of the texts. We used these points to build Alpha and Vietoris-Rips complexes on each of the domains.

2.2. Distribtuion distance-based Vietoris-Rips complexes.
Our second approach involved computing the distributions of word and sentence lengths and calculating the distances between the texts using the following distance measures:

- The Hellinger distance:

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

	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.

- the Chi-squared distance:

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

- the Euclidean distance:

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

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

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

2.3. **Domain comparsion.** 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

In the next secitons, we present the results given by both our approaches.

3.1. **Feature-based Alpha and Vietoris-Rips complexes.** The bottleneck distance matrix for the Alpha complex is shown in Table 1. The 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*

We will not present the results for the Vietoris-Rips complex in this report, because the results are very similar.

3.2. **Distribtuion distance-based Vietoris-Rips complexes.** *TODO: tekst*

4. CONCLUSION

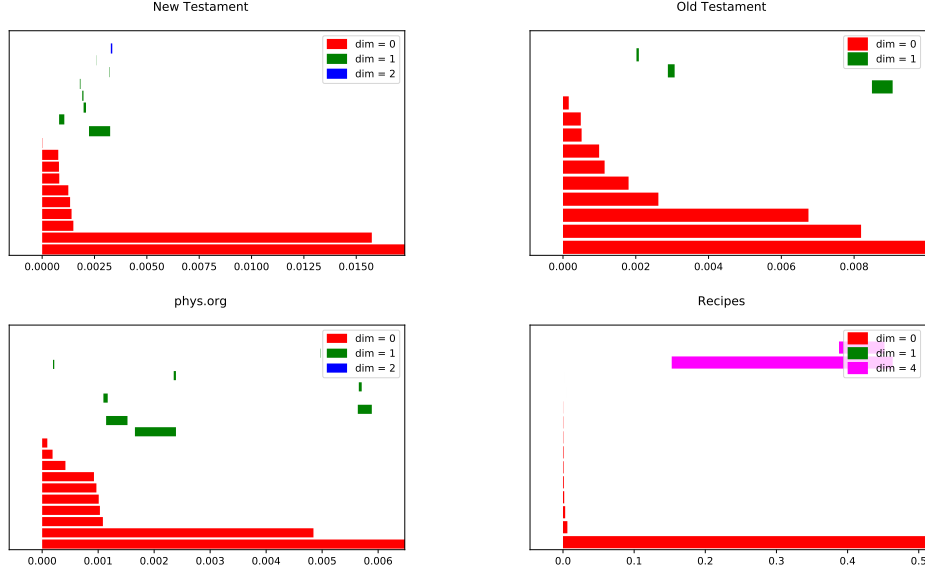As we saw in the previous section, the results are not very good.

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.
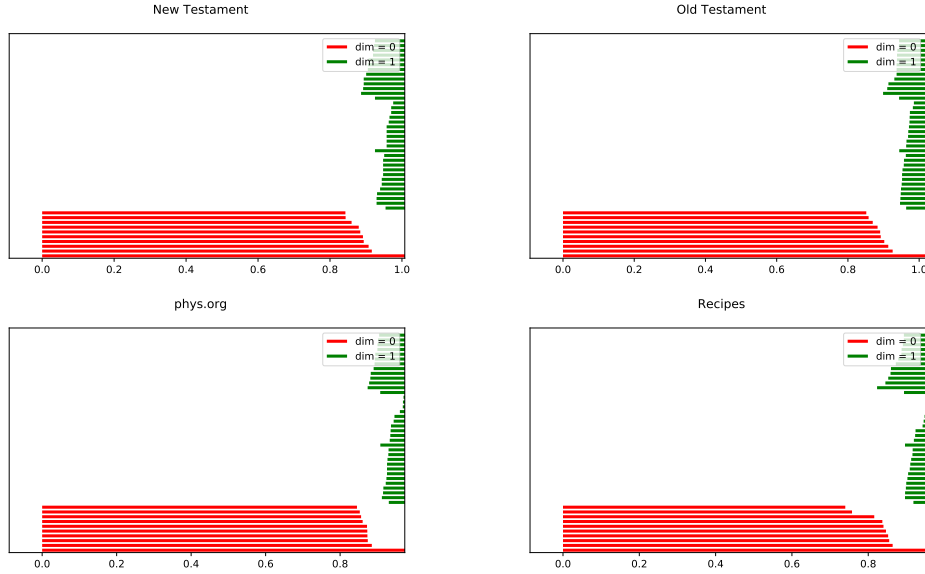

FIGURE 2. Persistence barcodes of the domains, calculated from the Vietoris-Rips complexes, built using the Hellinger distance between sentence length distributions.

5. Authors contributions

References

[1] 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.