University of Ljubljana
Faculty of Computer and Information Science

June 12, 2017

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 library [1] to compute persistent homology on the texts.

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 five-dimensional space 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	0	0	0
New Testament	0	0	0	0
phys.org	0	0	0	0
recipes	0	0	0	0

TABLE 1. The distance matrix

FIGURE 1. TODO: grafi

- the Chi-squared distance:

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

- 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.

3. Results

4. Conclusion

TODO: rezultati so slabi itd
TODO: kaj bi blo za probat

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.