Text classification using persistent homology

Matija Čufar, Domen Keglevič

Faculty of Computer and Information Science

6 June 2017

Problem description

Problem

Given sets of texts from different domains, build a classifier which can distinguish between the domains.

Solution idea

- Use persistent homology.
- For each text calculate several features (get a vector in $\mathbb{R}^d$).
- Make a simplicial complex using distances between vectors of features.
- Calculate bottleneck distance between persistence diagrams.

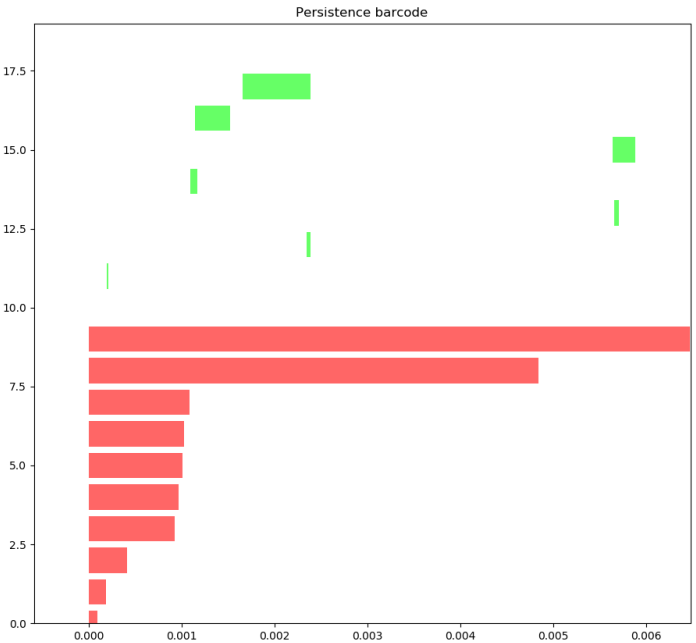
Which features?

We calculated several features:

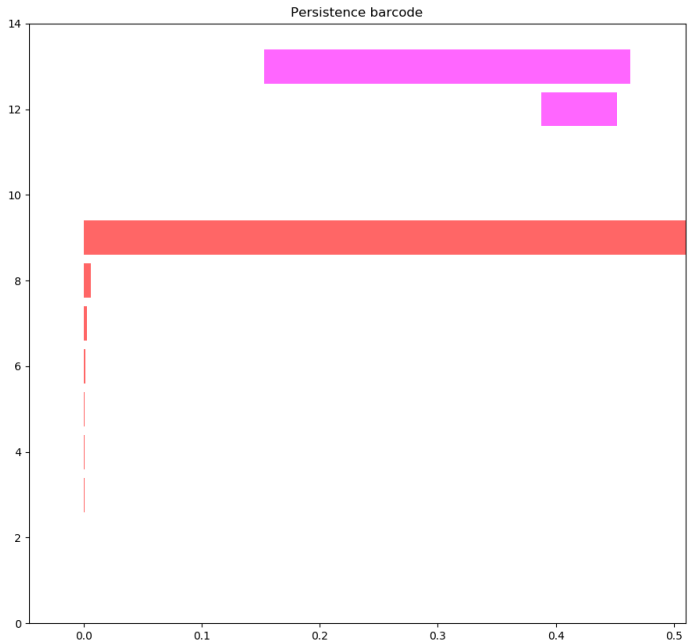
- Average word length / longest word length,
- Num of different words / num of all words,
- Three words with top tf-idf / num of all words,
- Etc.
- Separately: distribution of word and sentence length.

Results

Some barcodes:



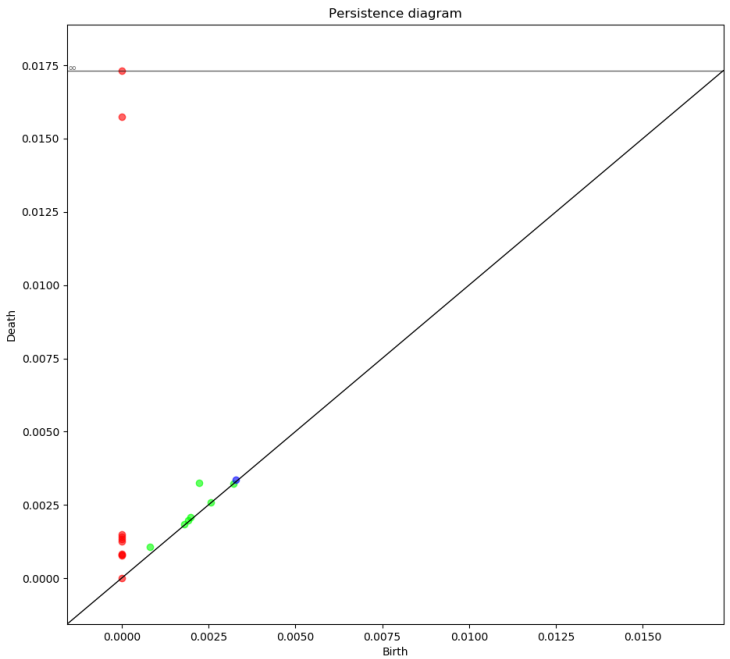
Physics news



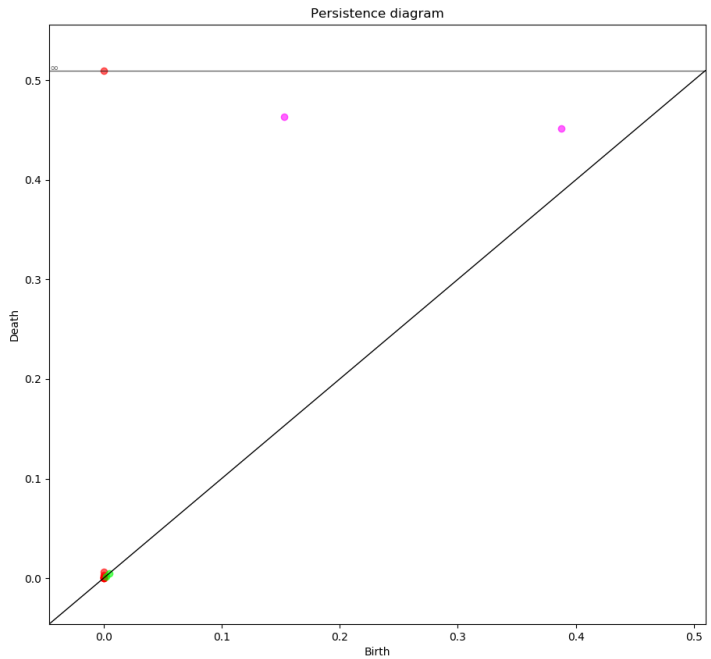
Recipes

Results

Some persistence diagrams:



Bible, New Testament



Recipes

Results (vectors of features)

	bible-new	recipes	phys.org	bible-old
bible-new	0.000	0.008	0.008	0.008
recipes	0.008	0.000	0.002	0.003
phys.org	0.008	0.002	0.000	0.003
bible-old	0.008	0.003	0.003	0.000

Bottleneck distance (alpha complex, points in $\mathbb{R}^d$)

	bible-new	recipes	phys.org	bible-old
bible-new	0.000	0.094	0.112	0.082
recipes	0.094	0.000	0.045	0.054
phys.org	0.112	0.045	0.000	0.082
bible-old	0.082	0.054	0.082	0.000

Bottleneck distance (Rips complex, points in $\mathbb{R}^d$)

Results (distribution of sentence lengths)

	bible-new	recipes	phys.org	bible-old
bible-new	0.000	0.109	0.250	0.185
recipes	0.109	0.000	0.165	0.132
phys.org	0.250	0.165	0.000	0.088
bible-old	0.185	0.132	0.088	0.000

Bottleneck distance (Rips with abstract points, $d(P, Q)$ is Euclidean distance).

	bible-new	recipes	phys.org	bible-old
bible-new	0.000	0.103	0.031	0.018
recipes	0.103	0.000	0.105	0.112
phys.org	0.031	0.105	0.000	0.040
bible-old	0.018	0.112	0.040	0.000

Bottleneck distance (Rips with abstract points, $d(P, Q) = \sqrt{\frac{1}{2}\sum\left(\sqrt{p_i} - \sqrt{q_i}\right)^2}$, Helliger distance).

The End