Topological Data Analysis in Natural Language Processing – A Tutorial

Wlodek Zadrozny
College of Computing, University of North Carolina at Charlotte
School of Data Science, University of North Carolina at Charlotte
wzadrozn@uncc.edu

Abstract

Topological Data Analysis (TDA) introduces methods that capture the underlying structure of shapes in data. Within the last two decades, TDA has been mostly examined in unsupervised machine learning tasks. TDA has been often considered an alternative to the conventional algorithms due to its capability to deal with high-dimensional data. in different tasks including but not limited to clustering, This tutorial will focus on applications of topological data analysis to text data.

Description of the Proposed Tutorial

The tutorial will cover the introductory concepts and definitions in Topological Data Analysis (TDA), focusing on text analysis. A discussion of available software will all be included in the tutorial. The goals of the tutorial are as follows:

1. Make the audience familiar with methods and tools of topological data analysis.
2. Teach methods combining topological data analysis, time series, and word embeddings, which has particular relevance for text understanding since. This approach can handle large texts and the encoding is sensitive to the word order.
3. Show how topological data analysis could be used to improve the accuracy of classification (even with small amounts of data), and augment deep learning by providing new feature types.

Content Level and Audience Prerequisites

The content level of the proposed tutorial roughly matches 40% beginner level, 30% intermediate level, and 30% advanced level. A general knowledge of NLP is preferred but not required. All required introductory TDA material will be covered during the tutorial.

General description of tutorial content

The focus of the tutorial is on the applications of Topological Data Analysis (TDA) in Natural Language Processing (NLP). Beginning with the the basic concepts and definitions, we will explain the uses of TDA in NLP. We will explain the step-by-step of the processes, as well as, the required/available software and packages for each step. We will provide easily understandable programs along with instructions.

Despite its growing popularity in data analysis, there has been only a handful of successful applications of TDA to NLP, perhaps because very few researchers has experience in both areas. On the other hand, the availability of easy-to-use libraries in both domains, points at the opportunity to bridge this gap and and to increase our understanding of the geometry of language beyond the the standard vector spaces, as well as the geometry embedded in deep neural networks.

Outline

This tutorial is designed in a way where, if an attendee is getting lost, they can proceed to the next step with prepackaged data and execute the code later at home. The main TDA package used in the tutorial will be Ripser. The code will be presented as Python notebooks running on Google Colab.

Table 1: Time-line of the tutorial.

No.	Segment	Presentation	Time-line
A	Introduction	Slides	10 minutes
B	TDA background and available software	Slides/Code	15 minutes
C	Positioning TDA in NLP	Slides	20 minutes
D	TDA methods for NLP	Code	35 minutes
E	Limitations, opportunities and conclusion	Slides	10 minutes

The following techniques will be discussed in the tutorial:

1. Extracting topological features from word embeddings representation of text.
2. Extracting topological features from TF/IDF vector space
3. Extracting topological features without utilizing conventional features.
4. Discovering topological features in deep neural networks.