COMPUTATIONAL LINGUISTICS I PROJECT
TOPOLOGICAL DATA ANALYSIS FOR WORD SENSE DISAMBIGUATION

TEAM MEMBERS:
• Michael Rawson • Sam Dooley • Mithun Bharadwaj •Rishabh Choudhary

INTRODUCTION:
Topological Data Analysis (TDA) provides a general framework to extract information from high dimensional, incomplete and noisy datasets. Persistent homology is an algebraic tool for measuring topological features of data. The aim of this project is to use persistent homology to find cycles in the local homology of embedding representations of words and develop it into an unsupervised clustering algorithm to disambiguate the different senses of a word.

MOTIVATION:
TDA has become a very active and broad area of research but its application in natural language processing (NLP) is in its infancy. Using persistent homology to analyze the distribution of word embeddings is a challenging and a relatively unexplored area. Further, the algorithm using the local homology of data-points is unsupervised and implementations of such algorithms run in polynomial time.

IMPLEMENTATION:
- Cluster the words into different connected groups using persistent homology.
- Persistent homology gives the birth and death time for the clusters. On a persistence diagram (Birth time vs Death time), any cluster close to the $x = y$ line i.e the difference between the birth and the death time is below some threshold is considered noise.
- The remaining clusters for the word can be considered as an estimate to the number of different meanings and the other words in each of the clusters are associated with different meanings.

BENCHMARK:
Benchmarking can be done in two ways against any word sense disambiguation (WSD) dataset:
1. Compare the number of clusters from the algorithm to the number of meanings for the word according to the dataset.
2. Evaluating the sense assigned to the word by the algorithm to be disambiguated against the dataset.

CHALLENGES:
- Clusters are susceptible to noise and most of the available data are usually noisy
- Hyper-parameter search for the clustering algorithm (locality ball around each point, noise threshold, embedding dimension size)

REFERENCE:
1. Tadas Temcinas, 2018 Local Homology of Word Embeddings
2. Devendra Singh Chaplot, et al. 2015, Unsupervised Word Sense Disambiguation Using Markov Random Field and Dependency Parser