Semantically Similar Sentences Search
Independent Study on ML related to Child Rescue
COMPSCI596E

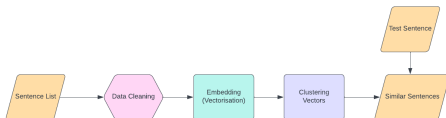Rahasya Barkur

University of Massachusetts Amherst

September 23, 2024

Outline

Motivation

- Short Term Goal: Clustering words with similar context
- Long Term Goal: Identifying possible child abuse from recorded calls by detecting threat words.

Methodology

- Input - A Sentence
- Output - List of sentences semantically similar to the input sentence.

Required Resources

- Open source Datasets
  - Dataset 1 - 20 Newsgroup Group Dataset [1]
  - Dataset 2 - News HeadLine Dataset [2]
- Computation - Pre Trained Model is utilised. Hence no GPU required
- Libraries Utilised : Numpy, Scikit Learn, NLTK, Sentence Transformers, Pandas, MatplotLib
- Pre Trained Vectorisation Models: BERT-CLS, Mini LM

References

[1] Tom Mitchell. *Twenty Newsgroups*. UCI Machine Learning Repository. 1997. DOI: 10.24432/C5C323.

[2] Rishabh Misra. *News Category Dataset*. June 2018. DOI: 10.13140/RG.2.2.20331.18729.

Thank you!
Questions? Suggestions?