NLP for NOTEEVENTS data

Naveen Kilaru

Introduction

Title: NLP Tutorial Using MIMIC Data

Objective: Learn to apply modern NLP techniques on EHR data, focusing on medical notes.

Key Takeaway: Extract valuable insights from unstructured clinical data to enhance machine learning models.

Learning Outcomes

Apply **SpaCy** & **SciSpaCy** for entity extraction.

Implement **Word2Vec** for word embedding.

Visualize data using **t-SNE** for dimensionality reduction.

Why medical notes matter?

Rich source of unstructured data in **EHRs**.

Often ignored due to complexity and lack of analysis.

NLP techniques reveal patterns and trends for better clinical decisions.

Tools and Libraries

Python 3.x

Google ColabSpaCy & SciSpaCy for entity extraction

Gensim for Word2Vec

scikit-learn for t-SNE

Matplotlib & Seaborn for visualization

Getting medical notes MIMIC

Data Source: MIMIC-III

- NOTEEVENTS
- DIAGNOSIS_ICD
- ADMISSION tables

Size: >1 GB

Used **Colab Pro** for faster processing

Entity Extraction

Load **SpaCy** or **SciSpaCy** models

Extract named entities from clinical notes

Visualize entities for clearer insights

Word Embedding

Preprocess notes using **Gensim**

Train Word2Vec to capture semantic relationships

Explore similar terms for deeper insights

Dimensionality reduction t-SNE

Apply **t-SNE** on word embeddings

Visualize complex relationships in 2D space

Spot clusters and anomalies in medical terminology

Observations

Structured data from unstructured notes improves ML models

NLP unlocks hidden insights in EHRs