2025

4Geeks Academy: data science cohort 12

DAY 29: NATURAL LANGUAGE PROCESSING

TODO

NLP	Overview & applications, text as features, some model types
IMAGE CLASSIFICATION PROJECT	Submit image-classifier-project-tutorial (Intro to Deep Learning module), if you haven't already
NLP PROJECT	Work on NLP Project Tutorial (Intro to NLP module), plan to finish by Friday

TOPICS

01 NATURAL LANGUAGE PROCESSING

02 TEXT AS FEATURES

03 COMMON NLP MODELS

NATURAL LANGUAGE PROCESSING

WHAT

Branch of data science and machine learning that deals with human language

- **1950s-1960s**: Rule-based - hand-crafted grammatical rules and dictionaries
- **1980s-1990s**: Statistics - probabilistic models, extraction of language patterns from large datasets
- **2010s**: Deep learning - recurrent neural networks learn representations of language from data
- **2017-Present**: Transformers - large language models (BERT, GPT, etc) achieve near human-level results on some language understanding and generation tasks.

WHY

Automate tasks that deal with written (or spoken) language:

- **Sentiment Analysis**: Product review analysis for e-commerce sites
- **Named Entity Recognition**: News article analysis to tag people and places mentioned
- **Machine Translation**: Real-time translation in messaging apps
- **Text Summarization**: Meeting notes summarization for business productivity

TEXT AS FEATURES

WHAT

Written text, like any other input needs to be encoded to numbers somehow

HOW

Common encoding methods:

- **Bag of Words** (BoW): Creates a vocabulary of all unique words in the corpus
  - Values indicate word frequency (count) in that document
  - Ignores word order and context
  - Simple but effective for many classification tasks

- **TF-IDF** (Term Frequency-Inverse Document Frequency): weights terms based on uniqueness in each document
  - TF (Term Frequency): How often a word appears in a document
  - IDF (Inverse Document Frequency): How rare a word is across all documents

- **Word Embeddings** (Word2Vec, GloVe): Maps words to vectors (typically 100-300 dimensions)
  - Captures semantic relationships between words
  - Words with similar meanings have similar vector representations
  - Pre-trained on large corpora or trained on your specific dataset

COMMON NLP MODELS

- **Supervised tasks** - Naive Bayes, Support Vector Machines (SVM), or Logistic Regression with bag-of-words or TF-IDF features

- **Language Understanding & Text Generation** - Transformer-based neural network models (BERT, GPT, etc) with vector embedding features

SVM MODELS

Support vector machine: classification or regression - finds a vector (or plane) that separates examples in feature space.

- ‘Support vector’: data points that lie closest to decision boundary
- Support vectors used to find optimal decision boundary
- More dimensions ~better classification

Scikit-learn implementations:

- **SVC():** support vector classifier, handles binary or multiclass classification
- **SVR():** support vector regressor, handles regression tasks