Projectsamenvatting: Logistische regressie voor tekstclassificatie met meerdere labels

Dit project heeft als doel het ontwikkelen van een logistisch regressiemodel voor tekstclassificatie met meerdere labels. Het model wordt getraind om tekstuele gegevens te analyseren en deze te categoriseren in verschillende vooraf gedefinieerde labels.

De belangrijkste stappen van het project omvatten:

Voorbereiding en gegevensverwerking:

Importeren van de benodigde bibliotheken en modules voor gegevensmanipulatie, machine learning en natuurlijke taalverwerking.

Downloaden van de vereiste bronnen van NLTK, zoals tokenizermodellen, stopwoorden en POS-taggermodellen.

Definitie van hulpprogramma's en functies:

Implementatie van functies zoals "clean" voor het verwijderen van niet-alfabetische tekens, "token\_stop\_pos" voor tokenisatie, stopwoordverwijdering en POS-tagging, en "lemmatize" voor woordlemmatisering.

Verwerking van de dataset:

Laden van een CSV-bestand in een pandas DataFrame.

Toepassen van de "clean" functie op de tekstkolom om de gegevens schoon te maken.

Uitvoeren van tokenisatie, stopwoordverwijdering en POS-tagging met behulp van de "token\_stop\_pos" functie.

Lemmatisering van woorden op basis van de POS-tags met behulp van de "lemmatize" functie.

Identificeren en vervangen van grove woorden met behulp van reguliere expressies en een lijst met grove woorden.

Schoonmaken van speciale tekens in de tekst.

Dataset splitsen:

Opsplitsen van de dataset in training- en evaluatiesubsets met behulp van de train\_test\_split-functie van scikit-learn.

Modeltraining:

Creëren van een pipeline waarin teksttransformatie en het logistische regressiemodel worden toegepast.

Trainen van het model met een opgegeven aantal epochen.

Model evaluatie:

Evalueren van de getrainde modellen op de evaluatiesubset.

Berekenen van de nauwkeurigheid voor elk label om de prestaties van het model te beoordelen.

Modelopslag:

Opslaan van de getrainde modellen en de pipeline voor toekomstig gebruik met behulp van de joblib-bibliotheek.