Εξόρυξη Μεγάλων Δεδομένων

1η Εργαστηριακή Άσκηση

«Οπτικοποίηση Αξιολογήσεων Ταινιών»

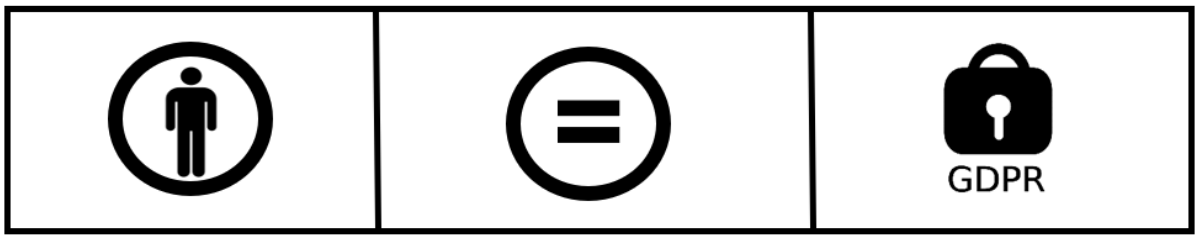
Βίννη Παναγιώτα

Α.Μ. : 170

Εξάμηνο: 2ο

penyvinni@gmail.com

Πίνακας Περιεχομένων



Στόχος

Στην παρούσα άσκηση, θα εξερευνήσουμε τη χρήση του εργαλείου της Google - Universal Sentence Encoder – για διανυσματοποίηση μικρών προτάσεων. Θα συνδυάσουμε μεθόδους μείωσης διάστασης όπως είναι η t-SNE για την ανάλυση κι οπτικοποίηση πολυδιάστατων του συνόλου δεδομένων SST-2. Το SST-2 είναι ένα σύνολο δεδομένων που περιλαμβάνει κριτικές ταινιών διαχωρισμένες σε θετικές και αρνητικές.

Περιγραφή Άσκησης

1. Φόρτωση και Προετοιμασία του Συνόλου Δεδομένων SST-2: Αρχικά, θα φορτώσετε το σύνολο δεδομένων SST-2 και θα προετοιμάσετε τα δεδομένα για την επεξεργασία.
2. Διανυσματοποίηση των Κειμένων με το Universal Sentence Encoder: Στη συνέχεια, θα χρησιμοποιήσετε το Universal Sentence Encoder για να μετατρέψετε τα κείμενα σε διανύσματα. Θα αξιολογήσετε την ποιότητα των διανυσμάτων και θα εξετάσετε τυχόν προβλήματα που μπορεί να προκύψουν.
3. Εφαρμογή της Μεθόδου t-SNE για Μείωση Διαστάσεων: Στο επόμενο βήμα, θα χρησιμοποιήσετε τη μέθοδο t-SNE για να μειώσετε τις διαστάσεις των διανυσμάτων που δημιουργήθηκαν από το Universal Sentence Encoder.
4. Οπτικοποίηση των Δεδομένων: Τέλος, θα οπτικοποιήσετε τα δεδομένα στον νέο χώρο διαστάσεων που δημιουργήθηκε από τη μέθοδο t-SNE. Αυτή η οπτικοποίηση θα σας βοηθήσει να κατανοήσετε τυχόν πρότυπα ή συσχετίσεις μεταξύ των κριτικών ταινιών.

Επεξήγηση του κώδικα

1. **Εισαγωγή βιβλιοθηκών:** Ο κώδικας ξεκινά με την εισαγωγή απαραίτητων βιβλιοθηκών, όπως η pandas για τον χειρισμό δεδομένων, η sklearn για τις λειτουργίες μηχανικής μάθησης, το TensorFlow Hub για τη φόρτωση του μοντέλου Universal Sentence Encoder (USE) και η matplotlib για την οπτικοποίηση.

2. **Φόρτωση και προετοιμασία του συνόλου δεδομένων:**
   - Το σύνολο δεδομένων SST-2 φορτώνεται από το αρχείο «sst2-train.csv» σε ένα πλαίσιο δεδομένων pandas DataFrame.
   - Η δομή του DataFrame (οι πρώτες γραμμές) εκτυπώνεται για τον έλεγχο των ονομάτων των στηλών.
   - Το σύνολο δεδομένων χωρίζεται σε χαρακτηριστικά (κείμενα - sentences) και ετικέτες (labels).
   - Το σύνολο δεδομένων χωρίζεται περαιτέρω σε σύνολα εκπαίδευσης και επικύρωσης χρησιμοποιώντας τη συνάρτηση train_test_split από το sklearn.

3. **Διανυσματοποίηση των κειμένων με τον καθολικό κωδικοποιητή προτάσεων:**
   - Το μοντέλο Universal Sentence Encoder (USE) φορτώνεται από το TensorFlow Hub («https://tfhub.dev/google/universal-sentence-encoder/4»).
   - Ορίζεται μια συνάρτηση embed_text για την ενσωμάτωση κειμένων με τη χρήση του φορτωμένου μοντέλου USE.
   - Τα κείμενα εκπαίδευσης και επικύρωσης ενσωματώνονται χρησιμοποιώντας τη συνάρτηση embed_text, παράγοντας embeddings για κάθε κείμενο.

4. **Εφαρμογή της μεθόδου t-SNE για τη μείωση των διαστάσεων:**
   - Εφαρμόζεται η μέθοδος t-SNE (t-distributed Stochastic Neighbor Embedding) για τη μείωση των πολυδιάστατων embeddings που παράγονται από το μοντέλο USE σε ένα δισδιάστατο χώρο.
   - Για το σκοπό αυτό χρησιμοποιείται η κλάση TSNE από την ενότητα manifold του sklearn.

5. **Οπτικοποίηση των δεδομένων:**
    - Τα μειωμένης διάστασης embeddings (train_embeddings_tsne) απεικονίζονται με τη χρήση διαγράμματος διασποράς.
    - Κάθε σημείο στο διάγραμμα αντιπροσωπεύει ένα ενσωματωμένο κείμενο, όπου οι άξονες x και y αντιπροσωπεύουν τις δύο συνιστώσες που λαμβάνονται από την t-SNE.
    - Τα σημεία χρωματίζονται με βάση τις αντίστοιχες ετικέτες τους (κλάσεις) χρησιμοποιώντας έναν χρωματικό χάρτη («RdYlGn»).
    - Το γράφημα απεικονίζεται με τη χρήση της matplotlib.

Συμπερασματικά, αυτό το πρόγραμμα φορτώνει ένα σύνολο δεδομένων, ενσωματώνει τα δεδομένα κειμένου χρησιμοποιώντας τον Universal Sentence Encoder, μειώνει τις διαστάσεις των embeddings χρησιμοποιώντας t-SNE και τα απεικονίζει σε ένα δισδιάστατο χώρο για να αποκτήσει πληροφορίες σχετικά με την κατανομή των δεδομένων.

Ο Κώδικας Υλοποίησης

```
# Importing necessary libraries
import pandas as pd  # For data manipulation
from sklearn.model_selection import train_test_split  # For splitting the
dataset
import tensorflow_hub as hub  # For loading Universal Sentence Encoder (USE)
model
from sklearn.manifold import TSNE  # For t-SNE dimensionality reduction
import matplotlib.pyplot as plt  # For data visualization

# Step 1: Loading and Preparing the SST-2 Dataset

# Load the dataset
dataset_path = "sst2-train.csv"
df = pd.read_csv(dataset_path)  # Reading data into a DataFrame

# Print DataFrame structure to check column names
print(df.head())  # Displaying the first few rows of the DataFrame

# Split into features (texts) and labels
texts = df['sentence'].tolist()  # Extracting text data
labels = df['label'].tolist()    # Extracting label data

# Split the dataset into training and validation sets
train_texts, val_texts, train_labels, val_labels = train_test_split(texts,
labels, test_size=0.2, random_state=42)

# Step 2: Vectorize the Texts with the Universal Sentence Encoder

# Load Universal Sentence Encoder module
module_url = "https://tfhub.dev/google/universal-sentence-encoder/4"
model = hub.load(module_url)  # Loading Universal Sentence Encoder model
print("Universal Sentence Encoder loaded")  # Printing confirmation message

# Define a function to embed texts
def embed_text(texts):
    embeddings = model(texts)  # Obtaining embeddings for texts using the USE
model
    return embeddings

# Vectorize training and validation texts
train_embeddings = embed_text(train_texts)  # Embedding training texts
val_embeddings = embed_text(val_texts)      # Embedding validation texts

# Example: Print first 3 embeddings and corresponding texts
```

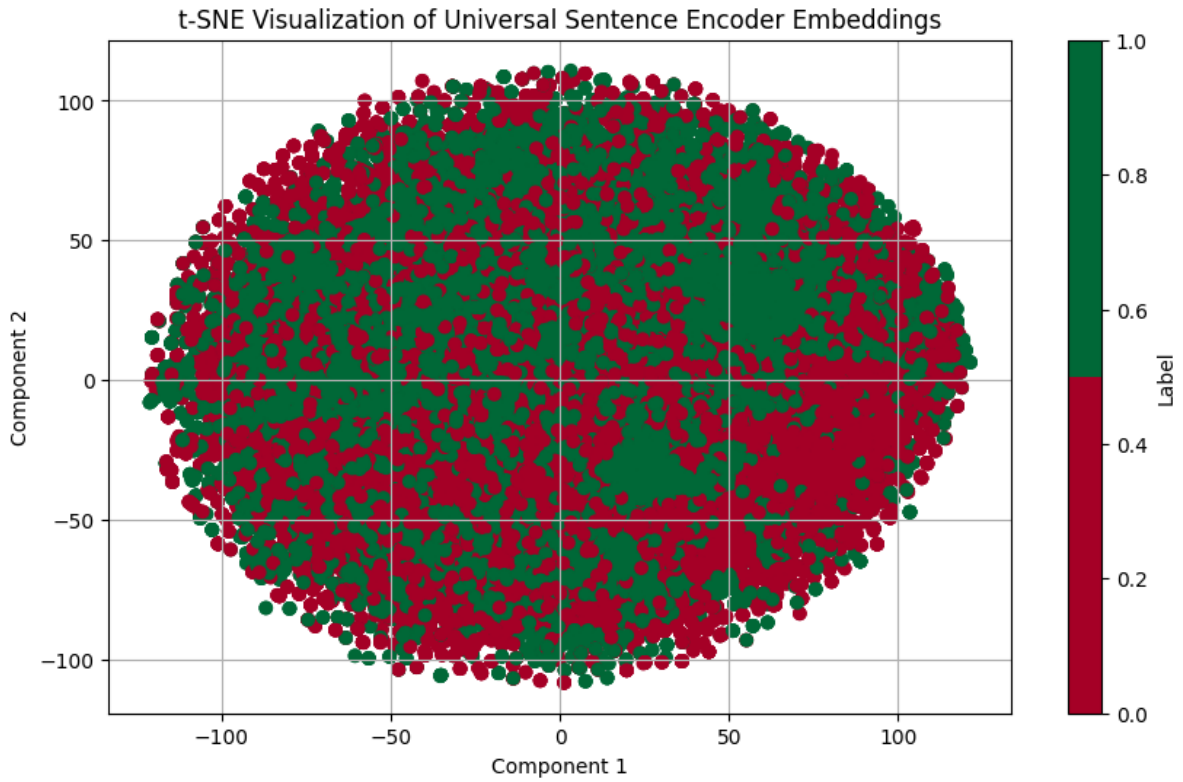
```
for i in range(3):
    print("Text:", train_texts[i])  # Displaying the text
    print("Embedding:", train_embeddings[i])  # Displaying the corresponding
embedding
    print("Label:", train_labels[i])  # Displaying the corresponding label
    print()

# Step 3: Application of the t-SNE Method for Dimensionality Reduction

# Apply t-SNE to reduce dimensions of the embeddings
tsne = TSNE(n_components=2, random_state=42)  # Initializing t-SNE object
train_embeddings_tsne = tsne.fit_transform(train_embeddings)  # Applying t-SNE
to training embeddings

# Step 4: Visualizing the Data

# Visualize the data in the new dimensional space created by t-SNE
plt.figure(figsize=(10, 6))  # Setting figure size
plt.scatter(train_embeddings_tsne[:, 0], train_embeddings_tsne[:, 1],
c=train_labels, cmap=plt.cm.get_cmap('RdYlGn', 2))  # Creating scatter plot
plt.title('t-SNE Visualization of Universal Sentence Encoder Embeddings')  #
Setting title
plt.xlabel('Component 1')  # Setting x-axis label
plt.ylabel('Component 2')  # Setting y-axis label
plt.colorbar(label='Label')  # Adding colorbar with label
plt.grid(True)  # Adding grid
plt.show()  # Displaying the plot
```

Ο κώδικας βρίσκεται στο παρακάτω αρχείο στο Google Collab:
https://colab.research.google.com/drive/1fWZgVJ281anGmQt3WE-3OgS70xhEk6VK?usp=sharing

Οπτικοποίηση των Αποτελεσμάτων

Η οπτικοποίηση που παρουσιάζεται παρακάτω απεικονίζει την κατανομή των δεδομένων κειμένου σε έναν μειωμένο δισδιάστατο χώρο που προκύπτει μέσω της εφαρμογής της τεχνικής t-SNE. Αυτή η οπτικοποίηση βασίζεται στα embeddings που παράγονται από το USE, ένα μοντέλο βαθιάς μάθησης ικανό να μετατρέπει τις εισόδους κειμένου σε πυκνές αριθμητικές αναπαραστάσεις που αποτυπώνουν το σημασιολογικό νόημα.

Στο διάγραμμα διασποράς, κάθε σημείο αντιπροσωπεύει ένα embedding κειμένου, με παρόμοια κείμενα που αναμένεται να συγκεντρώνονται μαζί. Το χρώμα κάθε σημείου αντιστοιχεί στην ετικέτα ή την κλάση του κειμένου, παρέχοντας πληροφορίες για την κατανομή των διαφορετικών κλάσεων στο χώρο των embeddings. Μέσω της μείωσης των διαστάσεων, τα υψηλής διάστασης embeddings που παράγονται από το μοντέλο USE μετατρέπονται σε χώρο χαμηλότερης διάστασης, διατηρώντας παράλληλα την τοπική δομή των δεδομένων. Αυτό επιτρέπει την οπτική εξέταση μοτίβων, συστάδων και διαχωρισμών μεταξύ των embeddings κειμένου.



Εικόνα 1: Οπτικοποίηση των δεδομένων στον νέο χώρο διαστάσεων (t-SNE)

Συμπεράσματα

- Το USE μετατρέπει αποτελεσματικά κάθε κείμενο σε μια υψηλής διάστασης αριθμητική αναπαράσταση (embeddings), τα οποία συλλαμβάνουν σημασιολογικές πληροφορίες σχετικά με τα κείμενα, επιτρέποντας τη σύγκριση και την ανάλυση δεδομένων κειμένου σε έναν αριθμητικό χώρο.
- Η t-SNE μειώνει επιτυχώς τα υψηλών διαστάσεων embeddings σε ένα δισδιάστατο χώρο και στοχεύει στη διατήρηση της τοπικής δομής των δεδομένων, επιτρέποντας την οπτικοποίηση συστάδων και μοτίβων δεδομένων.
- Το διάγραμμα διασποράς οπτικοποιεί τα embeddings σε έναν δισδιάστατο χώρο. Κάθε σημείο του διαγράμματος αντιπροσωπεύει ένα embedding κειμένου, όπου παρόμοια σημεία (κείμενα) αναμένεται να συσσωρεύονται μαζί.
- Το χρώμα κάθε σημείου αντιπροσωπεύει την αντίστοιχη ετικέτα ή κλάση του κειμένου, παρέχοντας πληροφορίες για την κατανομή των διαφορετικών κλάσεων στο χώρο των embeddings.
- Εάν παρατηρούνται διακριτές συστάδες στο διάγραμμα, αυτό υποδηλώνει ότι κείμενα με παρόμοιες σημασιολογικές έννοιες ομαδοποιούνται μαζί. Αυτό υποδηλώνει ότι τα USE embeddings αποτυπώνουν αποτελεσματικά τη σημασιολογική ομοιότητα.
- Εάν τα σημεία διαφορετικών κλάσεων είναι καλά διαχωρισμένα στο διάγραμμα, αυτό υποδηλώνει ότι τα embeddings περιέχουν διακριτική πληροφορία για εργασίες ταξινόμησης.
- Εάν υπάρχει επικάλυψη μεταξύ συστάδων ή κλάσεων, αυτό μπορεί να υποδηλώνει ασάφεια ή ομοιότητα μεταξύ ορισμένων κλάσεων ή κειμένων.