Πρόγραμμα Μεταπτυχιακών Σπουδών "Πληροφορικής και Δικτύων"
Τμήμα Πληροφορικής και Τηλεπικοινωνιών
Πανεπιστήμιο Ιωαννίνων
Εξόρυξη Μεγάλων Δεδομένων

Εργασία 1: Οπτικοποίηση Αξιολογήσεων Ταινιών

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.

Πρόγραμμα Μεταπτυχιακών Σπουδών "Πληροφορικής και Δικτύων"
Τμήμα Πληροφορικής και Τηλεπικοινωνιών
Πανεπιστήμιο Ιωαννίνων
Εξόρυξη Μεγάλων Δεδομένων

Μοντέλο Universal Sentence Encoder

Το Universal Sentence Encoder (USE) της Google είναι ένα προηγμένο μοντέλο κειμένου που έχει εκπαιδευτεί για τη δημιουργία διανυσμάτων κειμένου υψηλής ποιότητας. Η κύρια ιδέα πίσω από το USE είναι να μετατρέπει κείμενο σε διανύσματα χαρακτηριστικών που περιλαμβάνουν τόσο τη σημασιολογική όσο και τη συντακτική του πληροφορία. Το μοντέλο έχει εκπαιδευτεί σε έναν μεγάλο όγκο κειμένων από τον διαδίκτυο, καθιστώντας το ιδιαίτερα αποτελεσματικό σε διάφορες εφαρμογές επεξεργασίας φυσικής γλώσσας.

Η χρήση του Universal Sentença Encoder είναι αρκετά εύκολη και προσιτή. Μπορεί να χρησιμοποιηθεί είτε ως μοντέλο πρόβλεψης, όπου δίνεται ένα κείμενο ως είσοδος και παράγεται το αντίστοιχο διάνυσμα κειμένου ως έξοδος, είτε ως μοντέλο προεκπαίδευσης, όπου μπορεί να ενσωματωθεί σε πιο περίπλοκες αρχιτεκτονικές μοντέλων μηχανικής μάθησης.

Μια από τις κύριες εφαρμογές του USE είναι η αναζήτηση παρόμοιων κειμένων ή προτάσεων, η ταξινόμηση κειμένων, η σημασιολογική ανάλυση κειμένων και η ανίχνευση παραδειγμάτων. Επίσης, το USE χρησιμοποιείται ευρέως για την εκπαίδευση άλλων μοντέλων μηχανικής μάθησης, όπως ανάλυση συναισθημάτων, αυτόματη περίληψη κειμένου και μετάφραση.

Google Colab: https://shorturl.at/cuFZ0

```
import tensorflow as tf
import tensorflow_hub as hub

module_url = "https://tfhub.dev/google/universal-sentence-encoder/4" #@param
["https://tfhub.dev/google/universal-sentence-encoder/4",
"https://tfhub.dev/google/universal-sentence-encoder-large/5"]
model = hub.load(module_url)
print ("module %s loaded" % module_url)
def embed(input):
  return model(input)


#@title Compute a representation for each message, showing various lengths
supported.
word = "Elephant"
sentence = "I am a sentence for which I would like to get its embedding."
paragraph = (
    "Universal Sentence Encoder embeddings also support short paragraphs. "
```

```
    "There is no hard limit on how long the paragraph is. Roughly, the longer
"
    "the more 'diluted' the embedding will be.")
messages = [word, sentence, paragraph]

# Reduce logging output.
logging.set_verbosity(logging.ERROR)

message_embeddings = embed(messages)

for i, message_embedding in enumerate(np.array(message_embeddings).tolist()):
  print("Message: {}".format(messages[i]))
  print("Embedding size: {}".format(len(message_embedding)))
  message_embedding_snippet = ", ".join(
      (str(x) for x in message_embedding[:3]))
  print("Embedding: [{}, ...]\n".format(message_embedding_snippet))
```

Μέθοδος t-SNE

Google Colab: https://shorturl.at/cuFZ0

Αυτός ο κώδικας θα δημιουργήσει έναν τυχαίο πίνακα δεδομένων 100x10, θα εφαρμόσει τη μέθοδο t-SNE για να μειώσει τις διαστάσεις του πίνακα σε δύο διαστάσεις και τέλος θα οπτικοποιήσει τα δεδομένα σε ένα διάγραμμα στο Google Colab.

```
# Εγκατάσταση των απαραίτητων βιβλιοθηκών
#!pip install scikit-learn matplotlib

# Εισαγωγή των απαραίτητων βιβλιοθηκών
import numpy as np
import matplotlib.pyplot as plt
from sklearn.manifold import TSNE

# Δημιουργία τυχαίων δεδομένων
np.random.seed(42)
X = np.random.randn(100, 10)

# Εφαρμογή της μεθόδου t-SNE για μείωση των διαστάσεων
tsne = TSNE(n_components=2, random_state=42)
```

```
X_tsne = tsne.fit_transform(X)

# Οπτικοποίηση των δεδομένων
plt.figure(figsize=(10, 6))
plt.scatter(X_tsne[:, 0], X_tsne[:, 1], marker='o', c='b', edgecolor='k')
plt.title('t-SNE Οπτικοποίηση Δεδομένων')
plt.xlabel('Συνιστώσα 1')
plt.ylabel('Συνιστώσα 2')
plt.grid(True)
plt.show()
```

Σύνολο Δεδομένων

Το σύνολο δεδομένων Stanford Sentiment Treebank (SST-2) είναι ένα δημοφιλές σύνολο δεδομένων για εργασίες ανάλυσης συναισθήματος. Περιέχει κριτικές ταινιών ως προάσεις, με θετικές ή αρνητικές κριτικές. Κάθε κριτική ταινίας αποτελείται από μια πρόταση. Η ετικέτα συναισθήματος αντιπροσωπεύει το συνολικό συναίσθημα της πρότασης.

Παραδείγματα:

Θετική: "Αυτή η ταινία είναι ένα αριστούργημα!" (5 αστερία)

Αρνητική: "Η ταινία ήταν βαρετή και προβλέψιμη." (1 αστέρι)

Αναφορές

1. Universal Sentence Encoding, https://t.ly/457Rt
2. Σύνολο Δεδομένων SST2, https://t.ly/q5dnL
3. t-distributed Stochastic Neighbor Embedding, https://t.ly/m6BJ7
4. Scikit, https://t.ly/c1enp