ΘΕΩΡΙΑ

**Στόχος**: feature extraction με tf-idf και μετά evaluation με multi layer perception classifier MLP

Με βάση την βιβλιογραφία: 4476, **<https://arxiv.org/pdf/1802.05322.pdf>**

**1ο Βήμα:** pre-processing the data

Source: <https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089>

* Removing the genres that is not needed (Family, history, music, tv movie, animation, documentary)
* Lower case
* Stop words
* Punctuation
* Apostrophe
* Stop words
* Single characters
* Convert numbers
* Punctuation and stop words
* Stemming

**2ο Βήμα:** tf-idf

* Split into training (70%) and test set (30%) the description and the corresponding genres (source: <https://arxiv.org/pdf/1802.05322.pdf> )
* Transform and fit into the training set

Transform into the test set

(source: <https://arxiv.org/pdf/1802.05322.pdf> )

code:

1.[Creating a Movie Reviews Classifier Using TF-IDF in Python (analyticsvidhya.com)](https://www.analyticsvidhya.com/blog/2021/09/creating-a-movie-reviews-classifier-using-tf-idf-in-python/?) here he use the transform and the TfidfVectorizer like the source

2. <https://medium.com/geekculture/creating-content-based-movie-recommender-with-python-7f7d1b739c63> (not the same recommendation systems for the genres)

3. <https://github.com/AlexanderNixon/Machine-learning-reads/blob/master/Movie-content-based-recommender-using-tf-idf.ipynb> (not the same recommendation systems for the genres)

(εχει και περιγραφική στατιστική αλλά το πίνακα που ήθελα)

4. <http://www.sefidian.com/2022/02/28/understanding-tf-idf-with-python-example/>

(**general code**, with transform, two methods??)

<https://github.com/topics/movie-genre-classification>

(notebooks με το θεμα ισως υπαρχει κατι σημαντικο)

<https://medium.com/@kunalgupta4595/predicting-movie-genres-based-on-plot-summaries-bae646e70e04> (σημαντικό έχει και τα ngrams)

<https://github.com/LinggarM/Movie-Genre-Classification-based-on-Synopsis-using-Deep-Neural-Network-and-TF-IDF-Vectorizer/blob/main/Movie_Genre_Classification_based_on_synopsis_using_Deep_Neural_Network_and_TF_IDF_Vectorizer.ipynb> (παρομοιο εχει βοηθητικο με το τρειν τεστ με το transform)

Λογικά θέλουμε κάτι τέτοιο (αντί για movie, τα genres)

Εικόνα που περιέχει πίνακας

Περιγραφή που δημιουργήθηκε αυτόματα

Από τα λινκ για κώδικα παίρνουμε κάτι τέτοιο

Εικόνα που περιέχει πίνακας

Περιγραφή που δημιουργήθηκε αυτόματα

* Using chi-squared distribution to select the most significant features
* Combine tf-idf calculations and chi-squared thesholds

|  |  |
| --- | --- |
| **Useful links** | **Not so useful** |
| <https://github.com/LinggarM/Movie-Genre-Classification-based-on-Synopsis-using-Deep-Neural-Network-and-TF-IDF-Vectorizer/blob/main/Movie_Genre_Classification_based_on_synopsis_using_Deep_Neural_Network_and_TF_IDF_Vectorizer.ipynb> (κατανοητό train, test και γραφή αλγορίθμου αλλά εδώ το κάνει μόνο με ένα είδος ταινίας και δεν έχει τα ngrams)  <https://medium.com/@kunalgupta4595/predicting-movie-genres-based-on-plot-summaries-bae646e70e04> (έχει και κάποια καλά για την περιγραφική στατιστική)  <https://github.com/Bashkeel/BoW-TFIDF/blob/master/TF-IDF%20and%20N-Grams.ipynb>  <https://towardsdatascience.com/multi-class-text-classification-with-scikit-learn-12f1e60e0a9f> (εδώ έχει και το chi-squared) | <https://www.kaggle.com/code/leomauro/movies-classifying-genres-from-the-synopsis/notebook> (εδώ είναι περίπλοκο)  <https://www.kaggle.com/code/manasdalakoti/tf-idf-hybrid-movie-recommendation-system/notebook> (εδώ κάνει αναλυτικά τον τύπο δεν έχει το transform) |

**Links for MLP after TF-IDF**

|  |  |
| --- | --- |
| **446…** | **<https://arxiv.org/pdf/1802.05322.pdf>** |
| results were obtained by combining the MLP classifier and the TF-IDF features extraction method. All combinations  were evaluated using 5-fold cross-validation and the final classification metrics were obtained by averaging the results from each fold. | fitted using the train set and then predictions were done for the test set |
|  | <https://slogix.in/source-code/python/deep-learning-samples/how-to-build-spam-detector-using-multi-layer-perceptron-in-python/>(ακριβώς το παραπάνω γίνεται εδώ) |

Table με τους συνδιασμους από τα combo με τα genres είναι αρκετα σημαντικό να γίνει