

Case 3. Patient Drug Review

Cognitive Systems for Health Technology Applications, Spring 2019

Type

Individual or pair work, 15-20 hours

Aim

The aim of this assignment is to learn to use neural networks to process text data and predict ratings associated to the text.

Task

Your task is to use recurrent and convolutional neural networks to create a predictor for a collection of patient drug reviews extracted from [Drugs.com](https://www.drugs.com). Drugs.com is a comprehensive source of drug information online. The original dataset is explained in [Grässer et al \(2018\) article](#).

The task is to answer to a question:

- Can you predict the rating of the drug based on the review?

This is a [regression analysis](#) problem, e.g. how close to the original rating your model can predict.

Use at least 3 different categories for the rating (negative, neutral, positive), most preferably 5 categories (highly negative, negative, neutral, positive, highly positive). Use for example, Grässer et al article, Table 1, Data Description, as a starting structure for the categories.

Download the KUC Hackathon Winter 2018 dataset either from the Hackathon's homepage: <https://www.kaggle.com/jessicali9530/kuc-hackathon-winter-2018/home> or use the zip-file uploaded to OMA workspace. The dataset is already split into training and test sets. Use pandas to read and explore the data.

Return

Save your results to your GitHub folder and provide a link to your Notebook(s) in OMA. Use OMA's hyperlink tool for providing the link.

Evaluation

The following categories are used for evaluation:

- Organisation
 - The code is sequential and the code cells (parts of scripts) are in right order
 - The document follows a clear structure
- Clarity

- The document (and embedded code) is clear, polished, and easy to understand
 - The code follows good coding practices and contains sufficient comments
 - The document parts support the code
- Contents
 - The background and data preprocessing are well explained
 - The models are validated
 - The results are reasonable
 - The conclusions are clearly stated and in a line with the results

max. 15 points. Late submission reduces the maximum achievable points.

Materials

- [Deep Learning with Python.](#)
 - [Chapter 6. Deep learning for text and sequences.](#)
- [Deep learning for text](#)
- [One-hot encoding of words or characters](#)
- [Using word embeddings](#)
- [Understanding recurrent neural networks](#)
- [Keras.io - Text preprocessing](#)