Recitation 1

Soheil

Team_4_cs953

In this project, we are trying to predict the odds of dying each character in the famous TV show, Game of Thrones. In order to do so, we use various sources of information, such as three datasets available in Kaggle.com, and the text in the books.

This is a Data Science project, and as in other data science settings, we use so many different tools from Machine Learning (ML) to Knowledge Graphs and Information Retrieval. Because of that, tools from different platforms and programming languages have brought together to produce the most accurate results.

For the first stage of this project, I tried various ML classifiers in order to train a model for the characters, based on their various features, such as; gender, title, date of birth, mother, father, etc.

The datasets are split by their index. Odd indecies are used for testing and evens are used for training.

For evaluation, some well-known classification metrics such as *f1-score* and *confusion matrix* are used. For instance, f1-score for logistic regression and SVC using sigmoid function are as follows:

Method	F1-score
Logistic Regression	0.667
SVM (sigmoid)	0.75

In order to do the IR side of the project, the tool `Lucene` is considered. This is a powerful tool to do tokenization and indexing of the big corpses and make similarity scores based on the searched query. These information will later on be used to train a more accurate classifier.