INNOMATİCS
RESEARCH LABS

Text Data Creation

OVERVIEW

The dataset consists of data with 1000 sentences collected from different sources which are related to movies for Text Analytics. The main goal for this project is to create a target column with two classes whether the text is good or bad.

GOALS

1. Write a logic to create the target column with two classes good or bad.
2. Find the top 100 good words and 100 bad words.
3. Find the similarities between sentences.
4. Find the best 10 similar sentences.
5. Extract Most Important sentences from the whole dataset.
6. Create two datasets with top 100 good and bad words with most repeated previous and next words.
7. Do the basic preprocessing like text normalization, text segmentation (stemming / lemmatization) and if required apply stopwords removal for future problems.
8. Apply any Classification Model to classify the sentence is good or bad.

MILESTONES

Dataset with target

Data

Link to Data -
https://raw.githubusercontent.com/innomaticsresearchlab/Datasets/master/Datasets/text.csv

All the best from team Innomatics.