A blue and white logo

Description automatically generated with low confidenceGraphical user interface, text

Description automatically generated

M915 - NLG & NLU

Christina-Theano (Theatina) Kylafi

LT1200012

Extractive Summarizer

Report

## Tasks

The main task of the project comprises two subtasks:

1. Sentence Scoring
2. Single Document Extractive Summarization

## Data Processing

Kept only the extractive summaries

Split train data in 4 parts due to computational constraints

Labels -> RougeL ( with respect to gold summaries based on [SummaRuNNer: a recurrent neural network based sequence model for extractive summarization of documents](<https://arxiv.org/abs/1611.04230>) )

Chosen stats – RougeL / chosen or not

Sentence cleaning before tokenization to create the features

Features (based on Source: [Extractive Summarization using Deep Learning](https://arxiv.org/pdf/1708.04439v1) )

## Experiments

Several regression algorithms were tested and fine tuned on task 1 (sentence scoring) :

1. ElasticNet
2. LinearSVR
3. SGDRegressor
4. Lasso
5. LassoLars,

with the most appropriate turning out to be no. 3, **SGDRegressor**.

After fine tuning, the output scores were the following:

|  |  |  |
| --- | --- | --- |
| **10 Fold CV**  (mean scores) | Train 1 | Train 1 + Train 2 |
| MAE (mean absolute error) | 0.070 | 0.069 |
| MSE (mean squared error) | 0.016 | 0.016 |
| R2 (r - squared) | 0.155 | 0.156 |

A slightly better performance is observed in case of training the model with both training parts 1 & 2 (smaller error, greater R2).

## Training

Train1

Train1+2

## Evaluation

Stats and figures

## Notes