Data Science Challenge

Theodoros Georgiopoulos

Homework

INF342: Domain Classification Challenge

Firstly, I divided the problem in 2 parts. The first part was to make predictions for the test set nodes which were part of the graph and there was text for them available (171 nodes), and the second part was to make predictions for the test set nodes which were part of the graph but there was not available text for them (29 nodes).

So, for the first nodes I had to use information from both sources. For the text part, I tried with the fastext greek embeddings and with the AUEB embeddigs and the later were superior. Moreover, I downloaded the nltk greek stopwords and I saw that many words were missing. So, I downloaded more stopwords from the github repo https://github.com/xtsimpouris/gr-nlp-law/tree/master/Greek20%Stopwords , I converted them to lowercase and I concatenated them with the nltk stopwords. I removed the URLs from all the texts and I calculated the centroid vector for all the texts from the training hosts that had text (677 nodes).

For the graph part, I tried Deepwalk algorithm to learn the embeddings vectors and I used all the training hosts because they were part of graph (801 nodes).

I trained both with MLPs and I made predictions for the same 171 nodes. Because I had 2 predictions from every node, I took the average from them. For the training phase I had to choose a small number of epochs to keep the loss =50%. I achieved loss ¡3% but the kaggle results weren't good. This is because the loss function that we had (log-loss) punishes the correct answers that we give very small probability. If I trained the MLPs with small loss, the predictions were flat, for example one site had 99% probability to have sports content, but when I let the loss big enough the same site had 80% probability to have sports content and the remaining probability went to the other contents. So, if the prediction was wrong, the log loss addition was small.

For the senond nodes I had to use information from graph only. So I used deepwalk again with all the training hosts available (801 nodes). Now the test set was 29 nodes and I made predictions for them using the same MLP as above.

I reconstructed the test set and I submitted it.

I tried many things without good results. Because the training set was small (801 nodes) I think that I had to make it bigger. I took all the same content texts, like sports, and I calculated the centroid for them. After this, I iterated all the texts and I found those with the biggest cosine similarity. I labeled them and I added to the training set, but the results were bad. Moreover, for training I used logistic regression, random forests and MLPs and the results were better for the MLPs. For the similarity between texts I tried with TF-IDF but the results were not good as with the centroid of embeddings. The classes

were balanced: 'athlitismos' = 125, 'diaskedasi-psyxagogia' = 138, 'eidiseis-mme' = 138, 'katastimata-agores' = 130, 'pliroforiki-diadiktyo' = 146, so I did not try upsampling or something else.