NLP Project Milestone (March 3rd)

Zahra Ghods, Banafsheh Khazali

Introduction

The main purpose of the project is to find semantic text similarity between symptoms of different diseases to further group diseases with similar symptoms in a same cluster. Since our data was not in a standard format of a dataset, we planned to collect data from a website through scraping. Then after preprocessing step, we explored some approaches for feature extraction of the data. Afterwards, to make progress in our work, we used some similarity techniques such as cosine similarity. Data was also fed to a clustering model for the mentioned goal. Finally, using some visualizing graphs, the results were shown.

Solution

As for the data, this website provides the source for symptoms of different diseases that we are looking for. It includes more than forty links of illnesses infecting both children and adults. Using BeautifulSoup all the URLS for children and adults were extracted. Within each link, around four sections are found which demonstrate potential symptoms of each specific disease as well as a corner section specifying symptoms that necessitate emergency care in case of their occurrences. Hence, the data written in these sections were also extracted and documented in separate .txt files for both adults and children.

In the next step, we created a preprocessing method that takes data of each file, tokenizes, lemmatizes and removes stop words. The preprocessed data will be added in new files. Finally, all the information is stored in text documents, each includes around forty lines of informative texts regarded as symptoms which further provides around 1600 text input data to our similarity model.

Furthermore, TF-IDF vectorization is an approach used for extracting features of our text. However, more features are required to be extracted through different approaches like Bag of words BOW, Word Embeddings, Word2Vec and NER which we will consider for the comparison section. The model used to show the similarity is Cosine Similarity from sklearn.metrics.pairwise that measures the similarity between these embeddings in space based on cosine distance. The result shown in the Dendrogram representation (Fig.1), taking files as input and giving numeric values for distances as the output. More specifically it demonstrates the hierarchical clustering of the similarity matrix. However, more interpretations on the results are required in this step to identify what factors cause the similarity between text like its length. Another model used for predicting the similarities is a clustering algorithm called K-means. This model is also fed by the word embeddings of TF-IDF and clusters the input into three groups. Generally, we are going to investigate on other similarity techniques for our data.

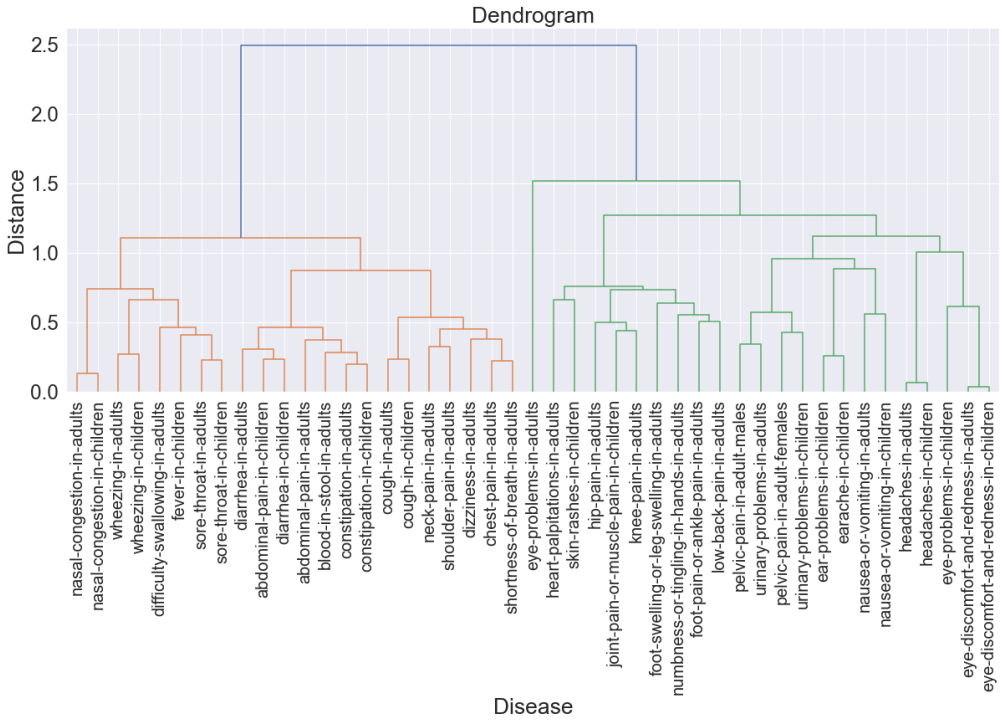

Fig.1. Dendrogram representation showing the similarity between every disease in the corpus.

Challenges or Discussion of Current Result

A problem with our data is that some general parts that do not add informative values to our data are also being crawled and saved in the stored txt files. These sections are repetitive in every page which add some bias for similarity detection. These parts need to be omitted in the preprocessing step.

Some pages prevent crawler to fetch and extract their data, and cause to missing information for some diseases. This problem is planned to be solved by creating sessions and further requesting through them.

A challenge that exists in our project is finding a way for validating our results. This means that even if the model predicts that symptoms of two diseases are very similar so these diseases are similar as well, we need to find an evaluation metric that proves this evidence by showing that these diseases are similar in the real world.

One important future update is to rich techniques used for similar semantic detection and analysis. This means, first, more features are required to be extracted using some vectorizations proposed in the second section. Second, in addition to features, more similarity detection models (e.g. more clustering methods) should be used to provide general insights of which techniques are more effective and result more accurate outputs.