Medical Q&A using NLP

Natural language processing course project

Done by: Khalid Nimri (2140145) – Aseel Suhail (2140197)

Abstract

This project will see us create a state of the art NLP system that is specially designed to be more efficient and accurate for Arabs citizens in accessing medical information. Making use of the production and aggregation of the wide range of dataset with medical questions and answers of patients the system applies state of the art NLP procedures like the Word2Vec, TF-IDF and Cosine Similarity. These methods are the ones which are especially selected in order to overcome the specific linguistic problems of the Arabic language which are like the complicated morphology and syntax. The Word2Vec is employed to do word embeddings of text to a vector space with the purpose of highlighting the departure of medical terms. Such measure, at the same time, the TF-IDF algorithm is to assign to each term the importance (weight) for the entire corpus, saving the time and improving the relevance. The Cosine Similarity is then used to measure the closeness between the query vectors and the document vectors, thus making sure that the system matches the user queries with the most relevant and factual medical answers.

The system has immensely benefitted through these high-end competencies and is now progressing its medical advice as contextually correct and precise. The initial findings have illustrated an improvement of the system adaptability to Arabic when it comes to accuracy and relevance to responses provided.

The wider implication of our system is great and it will be very helpful in improving the health literacy for the Arabic-speaking users. Through assisting folks to access readily available medical information, the system consequently supports better health outcomes by increasing the understanding and management skills for a given health condition. This abstract reviews our breakthrough technology and the convincing results to date, stressing the system's ability to change medical information transfer to something starkly different to what the digital health landscape is now.

Introduction

The proliferation of so many more digital health sources as well as massive amounts of health data increases the demand of intelligent systems that are well designed.

However, the health information in these systems has to be managed in a way that is not only effective but also efficient. On the one hand, the fact that there was no large-scale dataset used to train the natural language processing system in the Arabic language was the reason why our team developed a new system, which will be used to intercept and respond to medical queries in Arabic. The main goal of this system also relates to efficient information retrieval but besides that, it is obvious that user's capacity to obtain credible medical advices would be increased as well within a short period of time.

Background/Related Work

Nowadays natural language processing (NLP) has revealed the majority of influences on medical sphere that occur in the prediction of physiologic state and extension of patient communication possibilities in the new context of the patient - doctor dialog. The main aim of our project is to work on Arabic medical texts, a comparatively less-studied area in computational linguistics that offers us a great opportunity and at the same time, is a source of great challenge for linguistic and semantic analysis. Learthing from the basic NLP parlysism the work further shemes into Word2Vec, TF-IDF and Cosine Similarity. These technologies have been picked for their proven track record of classifying difficult linguistic structures manually and enhancing the quality and speed of system-assisted information retrieval in the area of healthcare.

Approach

The approach to advancing an arabic bot for medical queries adopted by us is used to place accuracy and relevancy in the same sentence as user queries. We use Word2Vec to render framings of words through which we capture the complex semantics in Arabic medical texts. Through this model, we can interpret the hidden important meanings in the language. In addition, the operation of TF-IDF algorithm ensures the provision of a ranking of words so as to evaluate the significance of each word as far as the whole document corpus is concerned, hence facilitating better retrieval of documents based on user queries. Besides that, Cosine Similarity determines the correspondingness between vector models of users queries and the possible responses, thus preventing the system from giving inaccurate and irrelevant answers. This whole range of advanced NLP methods not only makes the system more user-friendly and functional but also makes the application of NLP in Arabic medicine more practical.

Experiments

Dataset: We utilized the dataset available at https://www.kaggle.com/datasets/yassinabdulmahdi/arabic-medical-q-and-a-dataset, which contains diverse medical questions and answers.

Experimental Setup: We developed our system with Python - the library we used being Gensim, for Word2Vec, as well as scikit-learn for our TF-IDF model management. The experiments entailed the training of Word2Vec and TF-IDF model with the help of dataset and the goal was to make the systems semantically similar and fast to get information.

Evaluation Metrics: We used Cosine similarity and the visual inspections to evaluate the model's performance.

Results: This signifies that the accuracy of the responses obtains an improvement while the semantic information is integrated into the model system. The metrics are succinct and detailed from the results of the experiment. They are illustrated on graphs and charts.

Comparison: The TF-IDF model outperformed Word2Vec for the structured query type, while Word2Vec was better in the complex question type because it captured the semantic relations.

Conclusion

The photo shows the perfect performance of the world popular NLP technologies like Word2Vec, TF-IDF and Cosine Similarity - the echo of queries in Arabic medical texts. For the strategy is also in the process involving making the machine to be an expert in the health field by highly the accuracy and the significance together with its capabilities.

As this application is all about AI and machine-learning – the main analysis tool, the guidelines should mirror the idea However, such entwining of data obtained from the different sources will be the case that while the appearance of the network will look as complex as the existing system, there will be small difference in the logic of the process. Although some engineering problems in the field of image development are still unresolved, it will be the basis for the level of the personalized care, that everything will be less or no mistakes at all.

References and Team Contributions

References

- **Dataset**: **https://www.kaggle.com/datasets/yassinabdulmahdi/arabic-medical-q-and-a-dataset**

- TF-IFD From LetsDataScience: https://letsdatascience.com/tf-idf/

Contributions are as follows:

Student name	Student ID	Tasks
Khalid Nimri	2140145	• Word2vec • Cleaning functions and applying them • Tokenizing question column for Word2vec • finding closest match • phase 1 • most of the final write up
Aseel Suhail	2140197	• Presentation slides • TF-IDF • Answer_user_query function. • Average word vectors function • Phase 2 • Some work on the final write up.