**Project Proposal: Question answering in the citizen rights domain in order to help citizens fully utilize their rights in Israel based on KolZhut website:** [**https://www.kolzchut.org.il/he/%D7%A2%D7%9E%D7%95%D7%93\_%D7%A8%D7%90%D7%A9%D7%99**](https://www.kolzchut.org.il/he/%D7%A2%D7%9E%D7%95%D7%93_%D7%A8%D7%90%D7%A9%D7%99)

Overall project idea:

Closed Book Q&A problem is where the answer to a given question is expected to be directly inferred from a provided dataset without any external knowledge or information beyond what is available in the dataset. In other words, the machine learning model must predict the answer based solely on the information provided within the context of the question-answer pairs. The model is presented with a question and a subject-specific context, and is expected to extract the information relevant to the question from the provided context. The aim of this project is to develop a Q&A model that is tailored to the domain of Israeli citizen rights. This will involve compiling a dataset in the Hebrew language by sourcing information from KolZhut. After that I will finetune adaptions of BERT for the Hebrew language. The selection of the best model will be based on its performance on a validation set.

Previous work:

* In English language:
  + WikiQA - Wiki Question Answering corpus from Microsoft.  
    The WikiQA corpus is a publicly available set of question and sentence pairs, collected and annotated for research on open-domain question answering. Papers and current top models can be found here: <https://paperswithcode.com/dataset/wikiqa>. This dataset is for a different kind of Q&A when there is a multiple-choice question. **This project in this part was to demonstrate other aspects of Q&A.**
* SQuAD - The Stanford Question Answering Dataset.  
  Reading comprehension dataset, consisting of questions posed by crowd workers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable. 150,000 Q&A. The Dataset and top models results can be found here: <https://rajpurkar.github.io/SQuAD-explorer/> .  
  Best results on the dataset are [IE-Net (ensemble)](https://worksheets.codalab.org/bundles/0x8bf592beb7804a49bdca7eecc3fa82ef) **90.939 EM 89.452 F1** .   
  From this study I adopted the **data collection format** used in this study, which involved initially splitting the content into titles, and subsequently extracting paragraphs from each title's KolZhut page. Then, I created questions related to each paragraph, based on the instructions provided to the question builders who constructed SQuAD. This format was used to train datasets with HuggingFace models and will help me write **cleaner and more comprehensible dataset to train**.
* My model will be based on a **Hebrew text**, which has a different vocabulary and sentence structure. Furthermore, my model will require **specialized domain knowledge** oppose to SQuAD which can cover a wide range of topics.
* In Hebrew language:
  + PARASHOOT - A Hebrew Question Answering Dataset.  
    The dataset follows the format and crowdsourcing methodology of SQuAD, and contains approximately 3000 annotated examples. The paper can be found here: <https://arxiv.org/abs/2109.11314> .   
    The paper best model results on the dataset are **32.0 EM 56.1 F1**.  
    I drew inspiration from the **experimental setup** used in this study, which involved fine-tuning adaptations of BERT **(mBert,AlephBert,mBert)** to establish baseline results for evaluating a question answering dataset in modern Hebrew.  
     My dataset will be relatively small as theirs but my dataset is specifically focused on rights issues and is not intended for general applicability which I hope yield better results.
  + hebwiki-qa - a translated SQUAD dataset by Technion Data and Knowledge Lab.  
    Their work can be found here: <https://github.com/TechnionTDK/hebwiki-qa?fbclid=IwAR0Xbq-s1xu2gH8BS35zgFgNCeHIJ6wVZws4gqHCZ_VucbgiIngpHNTWApU>.   
    The project best model results on the dataset are **42.6 EM 55.9 F1**.   
    I will adopt their data adjustment method to convert my data into **HuggingFace Format**, and will try their fine-tuning approach with **heBERT** (an adaptation of BERT) as a baseline for my experiments. Furthermore, I plan to split my project into two distinct parts, like the authors of the study: **dataset creation and fine-tuning of Bert.** This approach will make my work more efficient  
    My dataset origin in based on **a Hebrew text but their dataset is made by translating the English SQUAD with fixing and removing bad translations**. Origin text is considered better than translated text in accuracy by capturing the nuances and subtleties of the original text and in cultural context which can lead to misunderstandings of mistranslations. Translated text **never fully replace the precision and richness** of an original text.

While existing studies on Q&A models have general applicability whether in English or in Hebrew, I innovate by providing a dataset specialized in a specific domain. models that are specialized to a particular domain can be trained on more focused and curated data. This can result **in higher quality training data and ultimately improve the performance of the Q&A model**. In the end I will compare my results to the results presented in the PARASHOOT project and hebwiki-qa and verify whether building a model for a specific domain has resulted in better performance in **EM and F1** than theirs, and analyze the output of the model.

Dataset:

* The dataset will be created by scraping the text using the BeautifulSoup library in Pytohn.   
  I will create questions for the text using the open-source annotation tool cdQA.  
  Due to the short time until submission, I anticipate that there will be around 250 questions, which will be split into three sets: 200 for the training set, and 25 each for the validation and test sets  
  The questions will be mainly focused on worker and pension rights and the answers will be the corresponding answers that have been provided on the site.

Evaluation methods:

* F1 = 2 \* (precision \* recall) / (precision + recall) – measures harmonic mean of precision and recall.
* EM = (number of exact matches) / (total number of predictions) - measures the percentage of times the model's prediction exactly matches the ground truth.
* The advantage of using both F1 score and EM is that they provide complementary information about the model's performance. F1 score measures the tradeoff between precision and recall, and thus captures the model's ability to make correct predictions while minimizing false positives and false negatives. EM, on the other hand, measures the exact match percentage, which indicates the model's ability to provide a precise and accurate response.

The project work can be seen on: <https://github.com/tal-ladi/KolZhutQA>.