CMPE 493 - Term Project Part 1
Dataset Preparation

Abdullatif Köksal (abdullatifkoksal@gmail.com)
Rıza Özçelik (riza.ozcelik@boun.edu.tr)
Gökçe Uludoğan (gokce.uludogan@boun.edu.tr)

DEADLINE: April 15, 09:00

Acknowledgement

We hugely thank Enes Çakır for coding the tool and enabling a smoother dataset preparation process.

1 Introduction

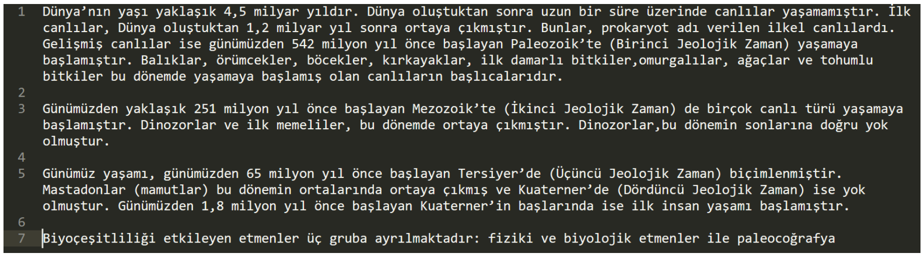
In the term project, you are expected to implement a question answering (QA) model that can answer geographic questions that are asked in Turkish free-text. Since a comprehensive Turkish QA dataset does not exist, we will create the dataset together!

A QA dataset is supposed to have two parts: A corpus that will be used to answer questions and QA pairs. To create the dataset, we will assign chapters to each one of you from high school geography books that are used by Turkish Education Ministry and available online in pdf format. These books will form our corpus. We will also provide an online tool that you can use to create QA pairs and upload your corpus. In this part of the project, you will extract the corpus from the parts you are assigned by copy/paste. The corpus must be separated to paragraphs and uploaded to the system. Then you will derive 60 QA pairs from these paragraphs and match each pair with the paragraph that contains its answer. So, we can list the steps as follows:

1. Download the book that is assigned to you. You can find the link in the ***sheet*** we have uploaded to Moodle[1].

2. Find your chapter. ***Note that page numbers refer to the page numbers written on the book itself, not to numbers written on pdf reader.***

[1] If you are not assigned with a part, contact us as early as possible



Dünya’nın yaşı yaklaşık 4,5 milyar yıldır. Dünya oluştuktan sonra uzun bir süre üzerinde canlılar yaşamamıştır. İlk canlılar, Dünya oluştuktan 1,2 milyar yıl sonra ortaya çıkmıştır. Bunlar, prokaryot adı verilen ilkel canlılardı. Gelişmiş canlılar ise günümüzden 542 milyon yıl önce başlayan Paleozoik’te (Birinci Jeolojik Zaman) yaşamaya başlamıştır. Balıklar, örümcekler, böcekler, kırkayaklar, ilk damarlı bitkiler,omurgalılar, ağaçlar ve tohumlu bitkiler bu dönemde yaşamaya başlamış olan canlıların başlıcalarıdır.

Günümüzden yaklaşık 251 milyon yıl önce başlayan Mezozoik’te (İkinci Jeolojik Zaman) de birçok canlı türü yaşamaya başlamıştır. Dinozorlar ve ilk memeliler, bu dönemde ortaya çıkmıştır. Dinozorlar,bu dönemin sonlarına doğru yok olmuştur.

Günümüz yaşamı, günümüzden 65 milyon yıl önce başlayan Tersiyer’de (Üçüncü Jeolojik Zaman) biçimlenmiştir. Mastadonlar (mamutlar) bu dönemin ortalarında ortaya çıkmış ve Kuaterner’de (Dördüncü Jeolojik Zaman) ise yok olmuştur. Günümüzden 1,8 milyon yıl önce başlayan Kuaterner’in başlarında ise ilk insan yaşamı başlamıştır.

Biyoçeşitliliği etkileyen etmenler üç gruba ayrılmaktadır: fiziki ve biyolojik etmenler ile paleocoğrafya

Figure 1: A Corpus Example

3. Copy/paste the paragraphs to a file where ***paragraphs are separated with a new line***. So, each paragraph ***must*** be single line and there ***must*** exist an empty line between the paragraphs. This is your corpus.
4. Upload your corpus to the system.
5. Derive 60 QA pairs from the paragraphs. Do not forget to associate each pair with the paragraph that contains the answer.

2 Corpus Guidelines

For us to be able to process your inputs, they must satisfy certain properties. Following the points below are quite important for corpus quality.

- You will upload the paragraphs in a single file. In this file, ***each paragraph must be a single line*** and ***paragraphs must be separated with an empty line.*** Note that if you notice that your paragraphs are corrupted and upload them again, it will erase QA and paragraph matching. This means that, you have to match each pair with a paragraph all over again. So, you are strongly advised to be sure that your paragraphs are correct, prior to starting QA pair derivation. You can see an example file in Figure 1.
- Even though the chapters you were assigned may contain tables, figures or some additional parts such as *Ölçme ve Değerlendirme Soruları, Araştırma Sorusu, Saha Çalışması,* you are not supposed to upload the texts in these parts. So, ***you have to exclude them*** while extracting paragraphs from the pdf and keep only the main body of the text. This to keep corpus less noisy.
- Do not include any table or figure captions to your corpus.

3 QA Guidelines

To keep the pairs in well shape, you have to consider the followings during pair generation:

- If the answer is a noun or a noun phrase, it ***must be in lemma form***.
  Example:

  Q: Türkiyenin en uzun ırmağı hangisidir?
  A: Kızılırmak
  Note that the answer is *Kızılırmak*, not *Kızılırmaktır.*

- If the answer is a verb, it must be in the same tense with the question.

  Example:

  Q: Sıcaklık arttıkça basınç nasıl değişir?
  A: Azalır

  Q: Buzullar eriyince Dünya'daki sıcaklık nasıl değişti?
  A: Arttı

- If the answer is a date, write it in *day month year* format, whichever is available.

  Example:

  Q: Kuzey Yarımküre'de en uzun gün hangi gündür?
  A: 21 Haziran

  Q: Ülkemize en çok hangi yılda turist gelmiştir?
  A: 2018

- The answers to some question can be numeric. Yet there must be ***at most 10 such answers and 20 such questions***. By numeric, we mean numbers, percentages, dates etc.

- ***You are encouraged to utilize*** *Ölçme ve Değerlendirme Soruları* in the books. Even if your part does not contain them, please check them out in the end of the chapter.

- As stated before you are asked to generate 60 QA pairs. **At least 30 of them must be different from each other**. To complete them to 60, you can rephrase the question to the same answer. Note that adding multiple questions to one answer is possible through the tool.

  Example:

  Q: Kuzey Yarımküre'de en uzun gün hangi gündür?
  Q: Kuzey Yarımküre'de en uzun gün hangi tarihte yaşanır?
  Q: Hangi tarihte Kuzey Yarımküre'de en uzun gün yaşanır?
  A: 21 Haziran

- It is possible that the same answer can be answer to multiple questions. In such a case, they are counted as distinct questions.

  Example:

  Q: Türkiyenin en uzun ırmağı hangisidir?
  A: Kızılırmak

  Q: Sınırlarımızdan doğup Karadeniz'e dökülen en uzun ırmak hangisidir?
  A: Kızılırmak

- Do not use any punctuation on the answers.

4 Submission & Grading

Since we have also access to the system you will not submit anything extra. Yet, data preparation part is quite important for the remaining of the project and should end as soon as possible. ***So, the deadline is sharp and no slack days are allowed as opposed to assignments.*** Moreover, you will be graded for this part as well and if your QA pairs or corpus does not obey the abovementioned rules, **you will be expected to correct them yourselves**. Here are the grading criteria of the term project and their tentative grades:

- Dataset Preperation - *35 pts*
- Final model - *50 pts*
- Presentation - *15 pts*
- Bonus - *10 pts.* To obtain bonus, you have to prepare more questions than 60. You will earn 1 points for each extra 6 questions, where 3 of them are unique. If you are willing to and need extra chapters for this purpose, contact us.