TensorFlow 2.0 Question Answering

Identify the answers to real user questions about Wikipedia page content

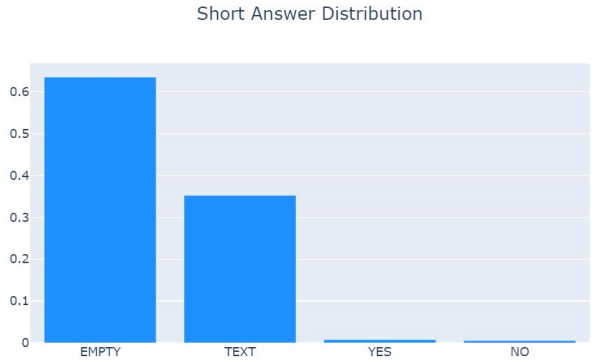
Project goals

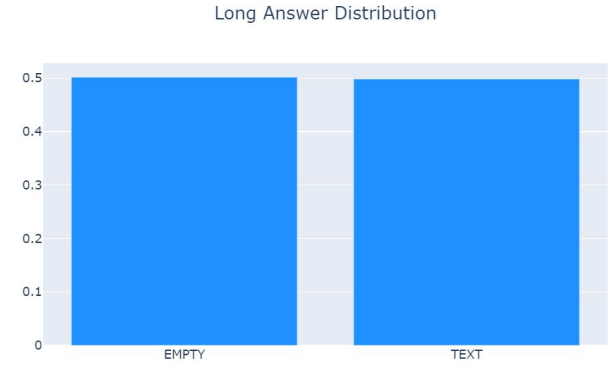
1. Develop a model that defines the answers to the questions
2. The score of the predictions must be at least a baseline equal to 0.21
3. Depending on the question it is necessary to distinguish between short and long answers to questions

Main Findings (Executive Summary)

If successful, this challenge will help spur the development of more effective and robust QA systems

- About 40 percent of the questions have short answers
- About 50 percent of the questions have long answers





Approach

Using the sinusoidal distance, the search was carried out for answers with the highest search.

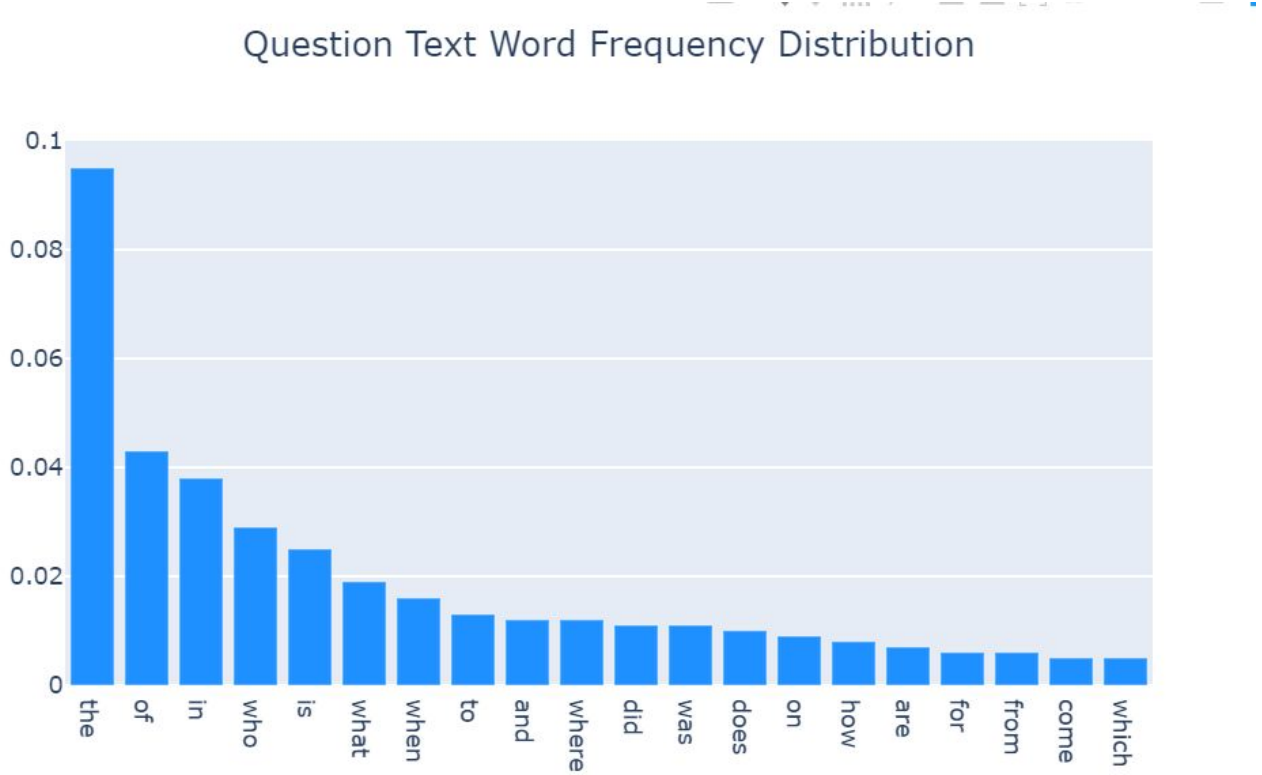
The cosine distance was considered. If the score was less than 0.15, then it was believed that there was no long answer.

The cosine distance between you and B is defined as

$$1 - \frac{u \cdot v}{||u||_2 ||v||_2}.$$

where u and v (N,) are input arrays

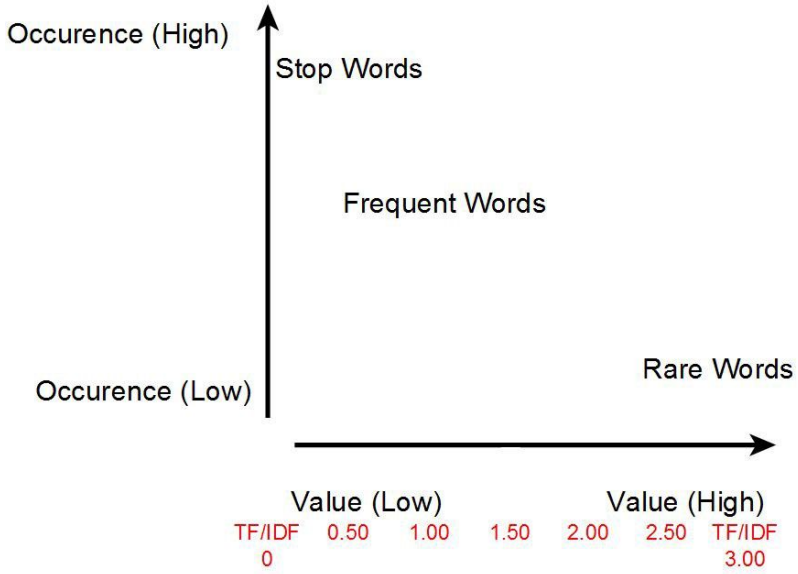
Key points


Question Text Word Frequency Distribution
0.1
0.08
0.06
0.04
0.02
0
the
of
in
who
is
what
when
to
and
where
did
was
does
on
how
are
for
from
come
which

Model description

TF - is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

IDF - is the inverse frequency of documents. It measures directly the importance of the term.



The tf - idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word. tf - idf is one of the most popular term-weighting schemes today.

Recommendations

- The creation of chat bots for support Department
- Creation of information systems

Thanks for your attention

Ruslan Astapov	email
Roman Dubatov	ds_science@gmail.com
Denis Tsapaev	phone
Elisaveta Povarova	898980895808