TensorFlow 2.0 Question Answering

Identify the answers to real user questions about Wikipedia page content

Project goals

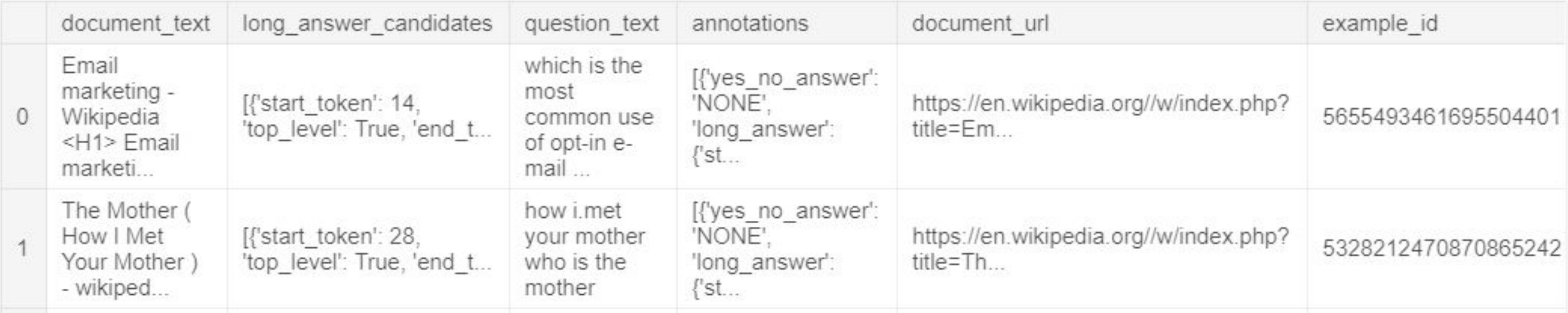
1. Develop a model that defines the answers to the questions
2. The score of the predictions must be at least a baseline equal to 0.21
3. Depending on the question it is necessary to distinguish between short and long answers to questions

Kaggle:

Dataset

The contents of dataset:

- Text
- List of long answers
- Question text
- Annotations

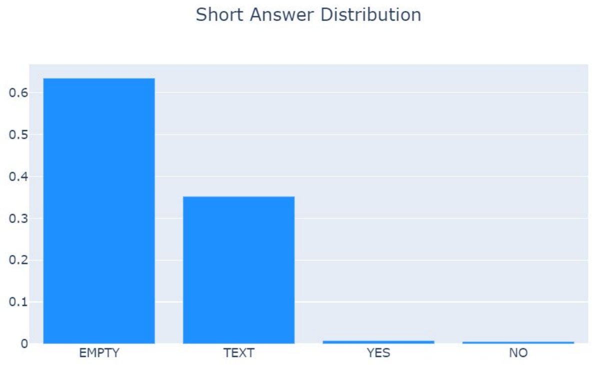


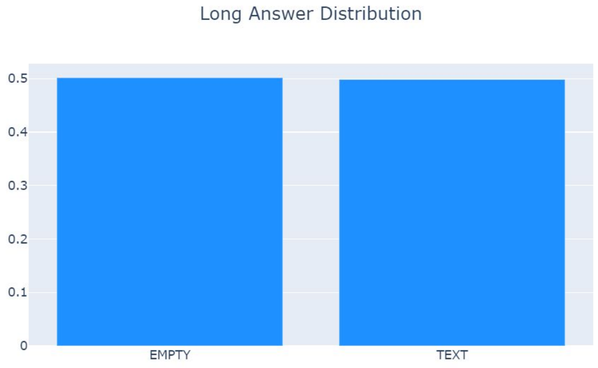
	document_text	long_answer_candidates	question_text	annotations	document_url	example_id
0	Email marketing - Wikipedia <H1> Email marketi...	[{'start_token': 14, 'top_level': True, 'end_t...	which is the most common use of opt-in e-mail ...	[{'yes_no_answer': 'NONE', 'long_answer': {'st...	https://en.wikipedia.org//w/index.php?title=Em...	5655493461695504401
1	The Mother (How I Met Your Mother) - wikiped...	[{'start_token': 28, 'top_level': True, 'end_t...	how i.met your mother who is the mother	[{'yes_no_answer': 'NONE', 'long_answer': {'st...	https://en.wikipedia.org//w/index.php?title=Th...	5328212470870865242

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


Short Answer Distribution


Long Answer Distribution

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

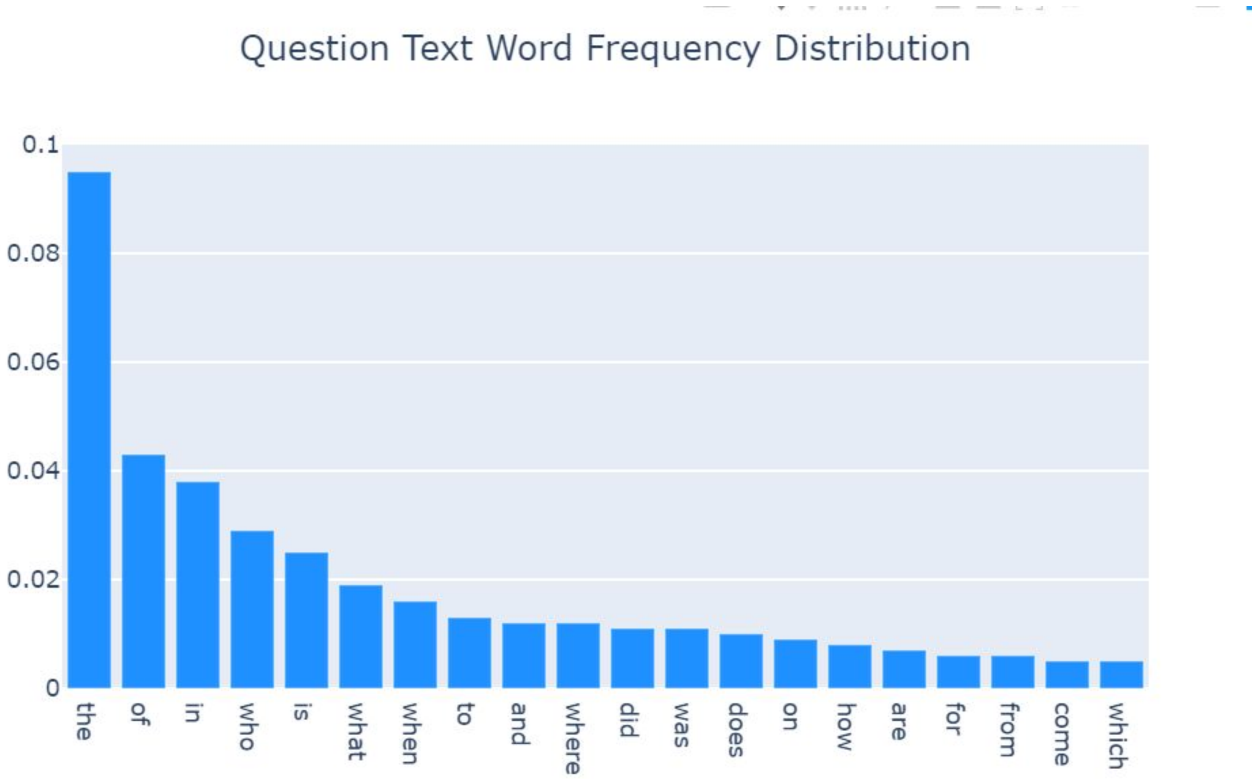
where u and v (N,) are input arrays

Python:

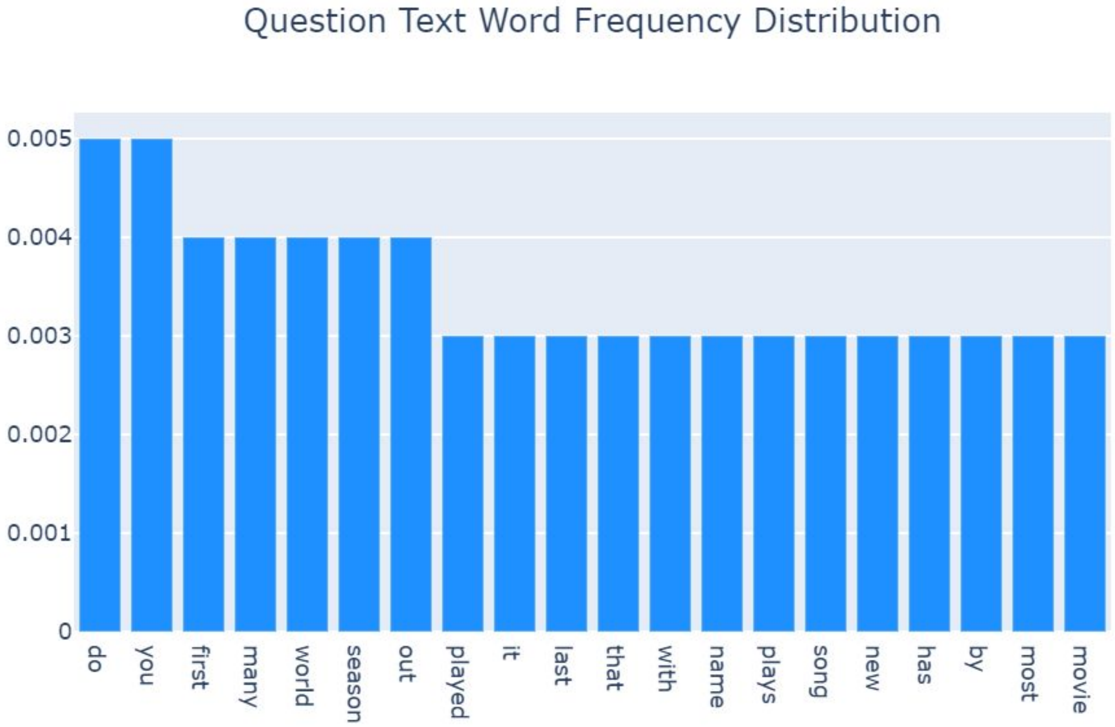
scipy.spatial.distance.cosine

scipy.spatial.distance.cosine (u, v, w = None)

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

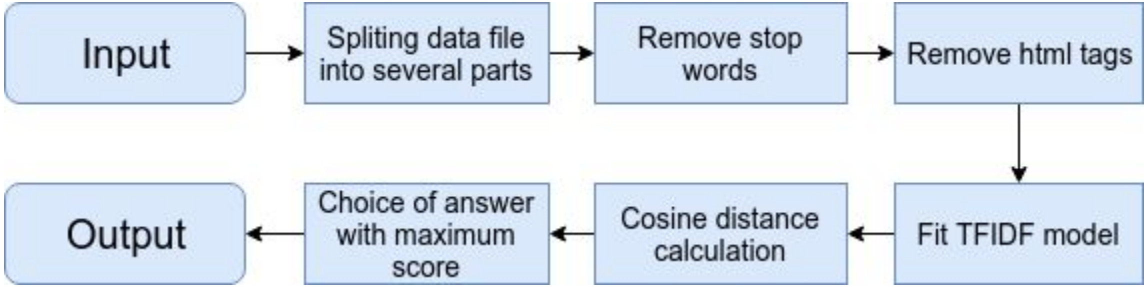
Key points


Question Text Word Frequency Distribution
0.005
0.004
0.003
0.002
0.001
0
do
you
first
many
world
season
out
played
it
last
that
with
name
plays
song
new
has
by
most
movie

Algorithm details

Input data contained 600 sample texts. The output formed the indices of the most appropriate answer.

In the process of preparing the text, lemmatization and normalization algorithms were used.



Input

Spliting data file into several parts

Remove stop words

Remove html tags

Fit TFIDF model

Cosine distance calculation

Choice of answer with maximum score

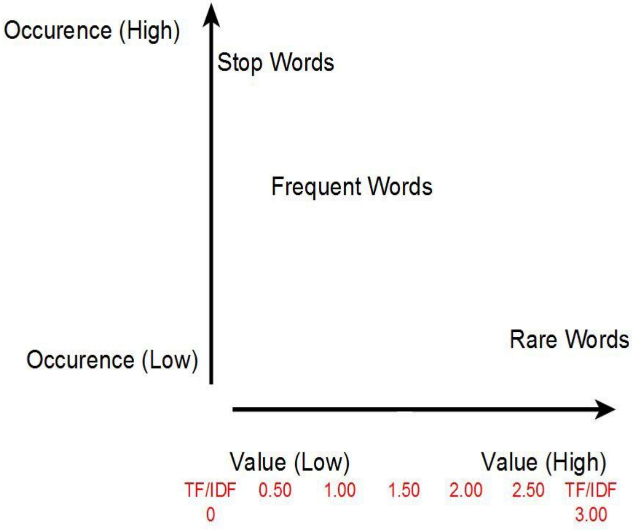
Output

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.



$TFIDF\ score\ for\ term\ i\ in\ document\ j = TF(i,j) * IDF(i)$

$where$

$IDF = Inverse\ Document\ Frequency$

$TF = Term\ Frequency$

$$TF(i,j) = \frac{\text{Term i frequency in document j}}{\text{Total words in document j}}$$

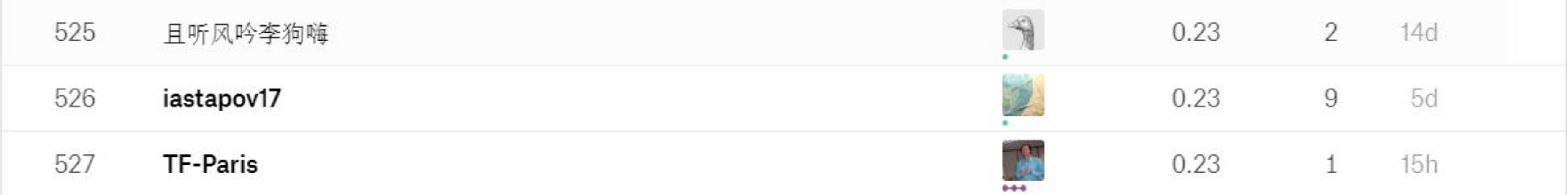
$$IDF(i) = \log_2\left(\frac{\text{Total documents}}{\text{documents with term i}}\right)$$

and

$t = Term$

$j = Document$

Result



525	且听风吟李狗嗨		0.23	2	14d
526	**iastapov17**		0.23	9	5d
527	**TF-Paris**		0.23	1	15h

Recommendations

- If tag <P> contained the answer, it increased score
- Using BERT for word processing
- Choose the optimal number of n_grams in TFIDF model

Thanks for your attention

Command 3+1

Ruslan Astapov Roman Dubatov Denis Tsapaev

Elisaveta Povarova