Reading Wikipedia
to answer open-domain questions

Подготовил:

Пугачев Александр, 151

Open-domain questions

What is the capital of Russia?

Open-domain questions

Who won the 2018 FIFA World Cup?

Open-domain questions

How many people lived in Australia in 1968?

Question answering system

Input: question in a natural language

Output: answer to the input question

Question answering system

How many people lived in Australia in 1968?

Question answering system

How many people lived in Australia in 1968?

Question answering system

How many people lived in Australia in 1968?

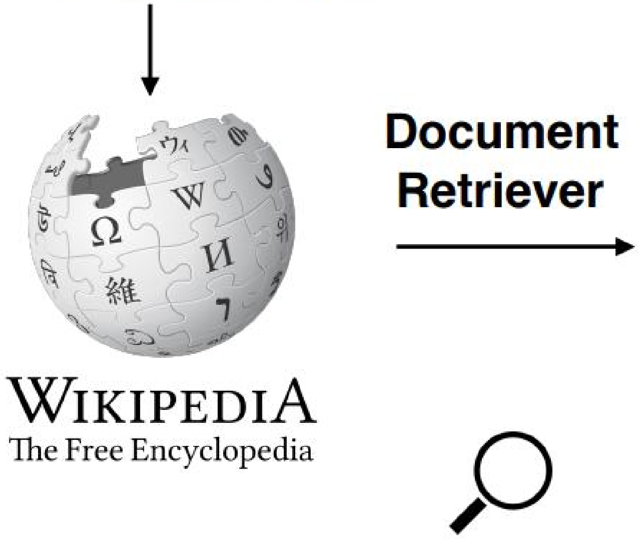
- Contains up-to-date knowledge

- Approach is generic

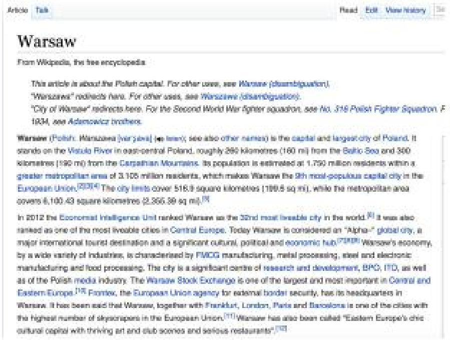
- Model is very precise while searching for an answer

DrQA

Q: How many of Warsaw's inhabitants spoke Polish in 1933?



Document Retriever



Warsaw

From Wikipedia, the free encyclopedia

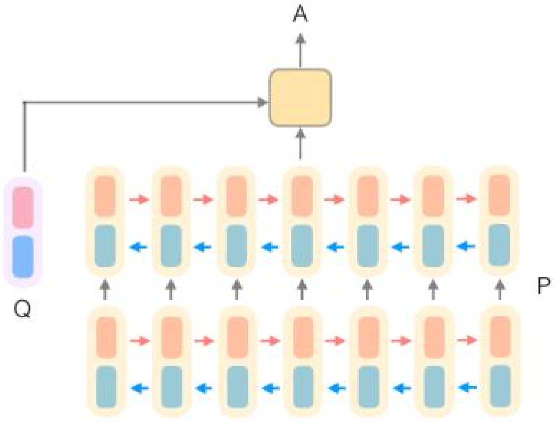
This article is about the Polish capital. For other uses, see Warsaw (disambiguation).
"Warszawa" redirects here. For other uses, see Warszawa (disambiguation).
"City of Warsaw" redirects here. For the Second World War fighter squadron, see No. 316 Polish Fighter Squadron. F
1934, see Adamowicz brothers.

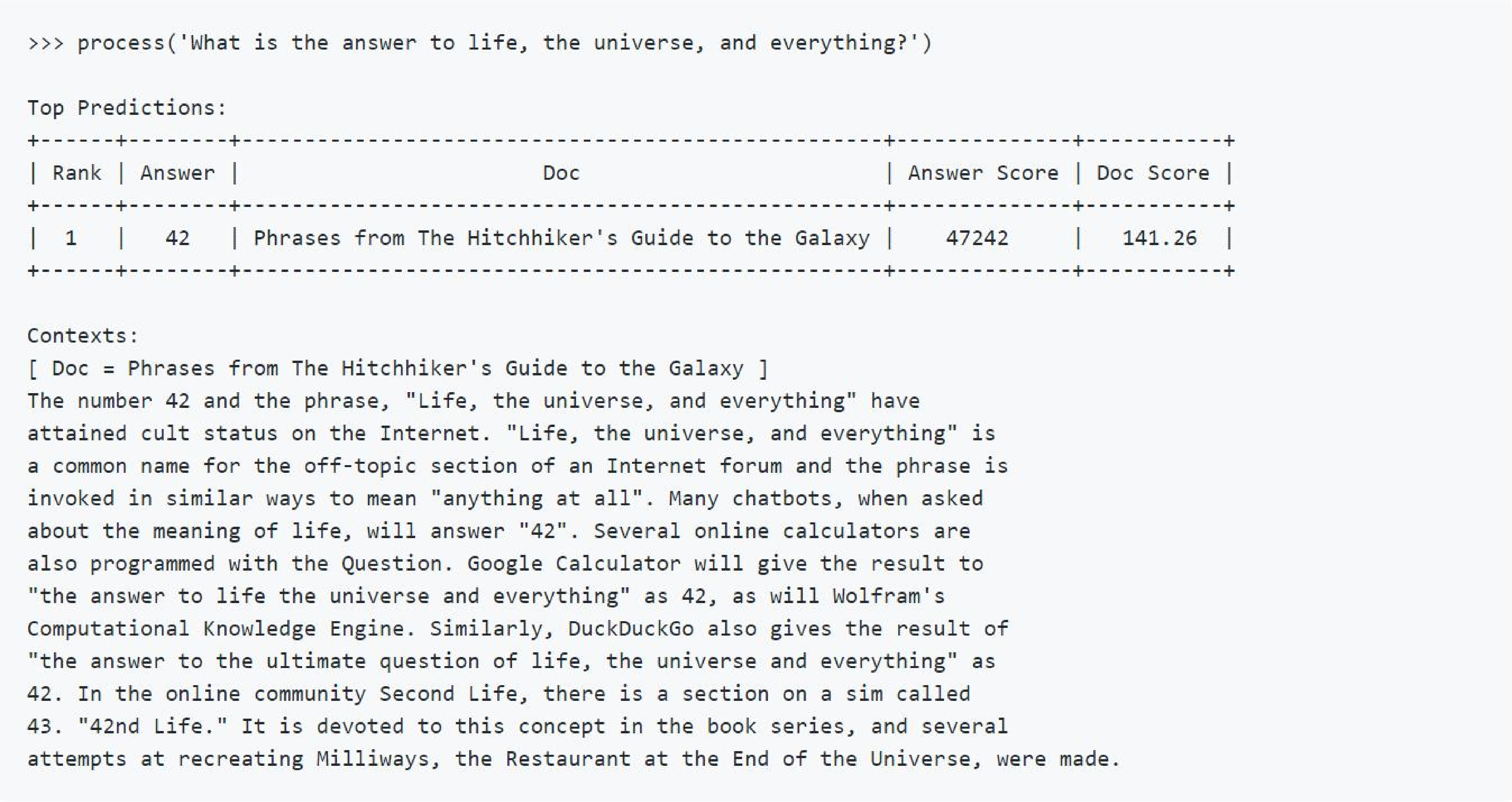
Warsaw (Polish: Warszawa [varˈʂava] (listen); see also other names) is the capital and largest city of Poland. It stands on the Vistula River in east-central Poland, roughly 260 kilometres (160 mi) from the Baltic Sea and 300 kilometres (190 mi) from the Carpathian Mountains. Its population is estimated at 1.750 million residents within a greater metropolitan area of 3.105 million residents, which makes Warsaw the 9th most-populous capital city in the European Union.[2][3][4] The city limits cover 516.9 square kilometres (199.6 sq mi), while the metropolitan area covers 6,100.43 square kilometres (2,355.39 sq mi).[5]

In 2012 the Economist Intelligence Unit ranked Warsaw as the 32nd most liveable city in the world.[6] It was also ranked as one of the most liveable cities in Central Europe. Today Warsaw is considered an "Alpha–" global city, a major international tourist destination and a significant cultural, political and economic hub.[7][8][9] Warsaw's economy, by a wide variety of industries, is characterised by FMCG manufacturing, metal processing, steel and electronic manufacturing and food processing. The city is a significant centre of research and development, BPO, ITO, as well as of the Polish media industry. The Warsaw Stock Exchange is one of the largest and most important in Central and Eastern Europe.[10] Frontex, the European Union agency for external border security, has its headquarters in Warsaw. It has been said that Warsaw, together with Frankfurt, London, Paris and Barcelona is one of the cities with the highest number of skyscrapers in the European Union.[11] Warsaw has also been called "Eastern Europe's chic cultural capital with thriving art and club scenes and serious restaurants".[12]

Document Reader

833,500

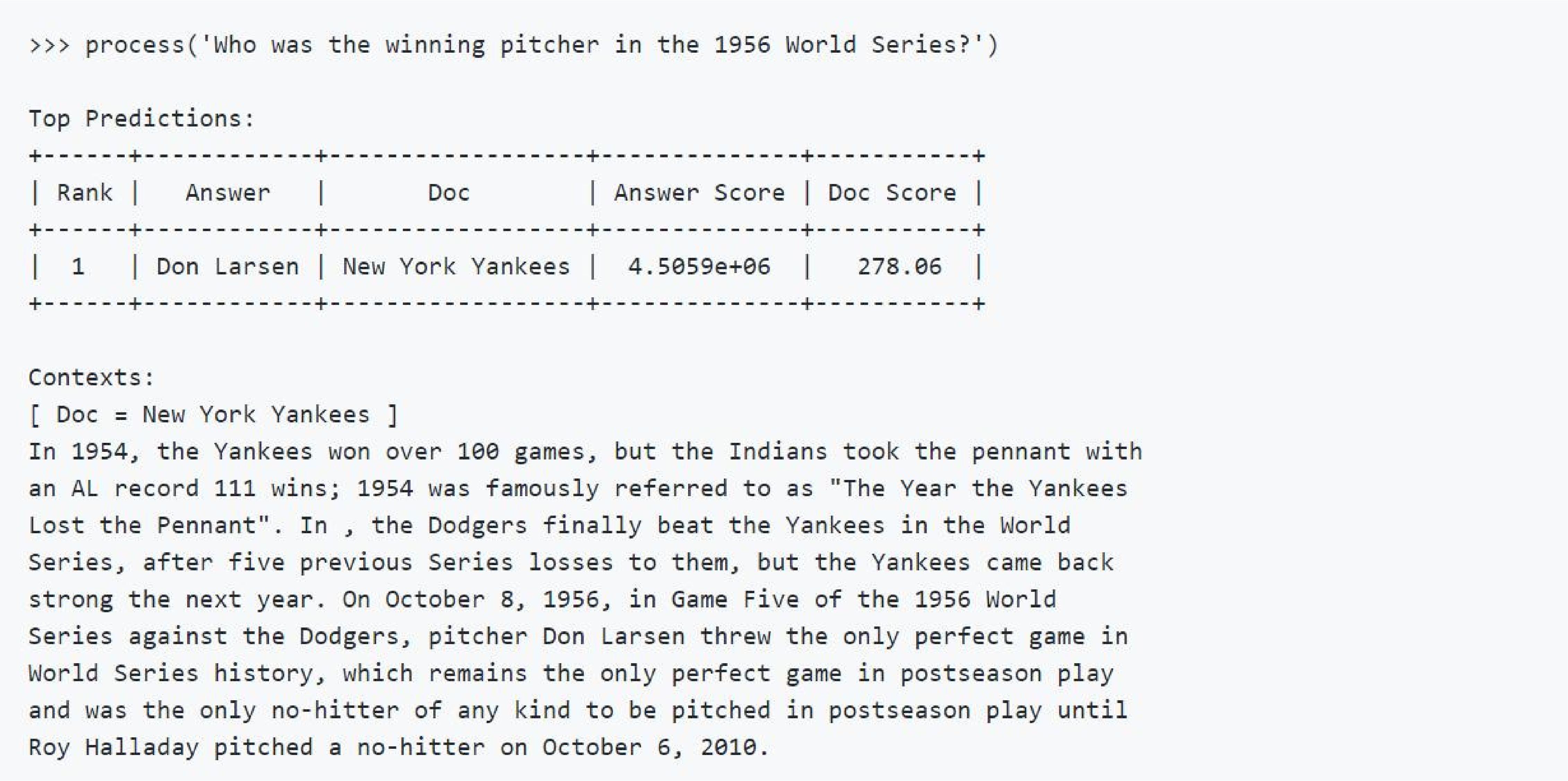




```
>>> process('What is the answer to life, the universe, and everything?')

Top Predictions:
+------+--------+---------------------------------------------------+--------------+-----------+
| Rank | Answer |                        Doc                        | Answer Score | Doc Score |
+------+--------+---------------------------------------------------+--------------+-----------+
|  1   |   42   | Phrases from The Hitchhiker's Guide to the Galaxy |    47242     |   141.26  |
+------+--------+---------------------------------------------------+--------------+-----------+

Contexts:
[ Doc = Phrases from The Hitchhiker's Guide to the Galaxy ]
The number 42 and the phrase, "Life, the universe, and everything" have
attained cult status on the Internet. "Life, the universe, and everything" is
a common name for the off-topic section of an Internet forum and the phrase is
invoked in similar ways to mean "anything at all". Many chatbots, when asked
about the meaning of life, will answer "42". Several online calculators are
also programmed with the Question. Google Calculator will give the result to
"the answer to life the universe and everything" as 42, as will Wolfram's
Computational Knowledge Engine. Similarly, DuckDuckGo also gives the result of
"the answer to the ultimate question of life, the universe and everything" as
42. In the online community Second Life, there is a section on a sim called
43. "42nd Life." It is devoted to this concept in the book series, and several
attempts at recreating Milliways, the Restaurant at the End of the Universe, were made.
```



```
>>> process('Who was the winning pitcher in the 1956 World Series?')

Top Predictions:
+------+------------+-------------------+---------------+-----------+
| Rank |   Answer   |        Doc        | Answer Score  | Doc Score |
+------+------------+-------------------+---------------+-----------+
|  1   | Don Larsen | New York Yankees  |  4.5059e+06   |   278.06  |
+------+------------+-------------------+---------------+-----------+

Contexts:
[ Doc = New York Yankees ]
In 1954, the Yankees won over 100 games, but the Indians took the pennant with
an AL record 111 wins; 1954 was famously referred to as "The Year the Yankees
Lost the Pennant". In , the Dodgers finally beat the Yankees in the World
Series, after five previous Series losses to them, but the Yankees came back
strong the next year. On October 8, 1956, in Game Five of the 1956 World
Series against the Dodgers, pitcher Don Larsen threw the only perfect game in
World Series history, which remains the only perfect game in postseason play
and was the only no-hitter of any kind to be pitched in postseason play until
Roy Halladay pitched a no-hitter on October 6, 2010.
```

Document Retriever

- Articles and questions are compared as TF-IDF vectors

- Local word order is taken into account with n-gram features (bigrams perform best)

- Hashing is used for preserving speed and memory efficiency

Document Reader

- Question $q = \{q_1, \ldots, q_\ell\}$

- Set of documents with n paragraphs in total

- Paragraph $p = \{p_1, \ldots, p_m\}$

Paragraph encoding

Each token p_i in paragraph p
is represented as feature vector $\widetilde{p}_i \in \mathbb{R}^d$

Paragraph encoding

Word embedding

$$f_{emb}(p_i) = \mathbf{E}(p_i)$$

- 300-dimensional GloVe word embeddings
- Fine-tune the 1000 most frequent question words

Paragraph encoding

Exact match

$$f_{exact\_match}(p_i) = \mathbb{I}(p_i, q)$$

- p_i exactly matches a word in q
- p_i is a lowercased word from q
- p_i is lemma form of a word from q

Paragraph encoding

Token features

$$f_{token}(p_i) = \big(POS(p_i), NER(p_i), TF(p_i)\big)$$

- POS – Part of Speech
- NER – Named Entity Recognition
- TF – term frequency

Paragraph encoding

Aligned question embedding

$$f_{align}(p_i) = \sum_j a_{i,j}\mathbf{E}(q_j)$$

$$a_{i,j} = \frac{\exp\left(\alpha(\mathbf{E}(p_i) \cdot \alpha(\mathbf{E}(q_j))\right)}{\sum_{j'} \exp\left(\alpha(\mathbf{E}(p_i) \cdot \alpha(\mathbf{E}(q_{j'}))\right)}$$

$\alpha(\cdot)$ is a single dense layer with ReLU

Paragraph encoding

$$\widetilde{p}_i = \big(f_{emb}(p_i), f_{exact\_match}(p_i), f_{token}(p_i), f_{align}(p_i)\big) \in \mathbb{R}^d$$

$$\{\pi_1, \ldots .\pi_m\} = LSTM\big(\{\widetilde{p}_i, \ldots, \widetilde{p}_m\}\big)$$

Question encoding

$$\{\varphi_1, \ldots, \varphi_\ell\} = RNN\big(\{q_1, \ldots, q_\ell\}\big)$$

$$\{\varphi_1, \ldots, \varphi_\ell\} \to \varphi$$

$$\varphi = \sum_j b_j \varphi_j$$

$$b_j = \frac{\exp(w \cdot \varphi_j)}{\sum_{j'} \exp(w \cdot \varphi_{j'})}$$

Train and Prediction

- Input: $\{\pi_1, \ldots, \pi_m\}, \quad \varphi$
- Output: $\text{for each token } i: \; P_{start}(i), P_{end}(i)$

Choose (i, i'):

$$i \leq i' \leq i + 15$$

$$\arg\max P_{start}(i) \times P_{end}(i')$$

Data

- *Wikipedia* for answering questions

- *SQuAD* dataset for training and testing Document Reader

- *CuratedTREC*, *WebQuestions*, *WikiMovies* for training and testing full QA system

Distantly Supervised Data

1) Run Document Retriever and retrieve 5 Wikipedia articles
2) Discard all paragraphs without exact match of the answer
3) Discard all paragraphs shorter than 25 and longer than 1500 chars
4) Discard all paragraphs without name entities from question
5) For remaining paragraphs score all positions that match answer using overlap between question and 20-token window
6) Save Top 5 paragraphs with highest overlap

Experiments

Finding relevant articles

Dataset	**Wikipedia Search Engine**	**Document Retriever**
SQuAD	62.7	**77.8** ↑
CuratedTREC	81.0	**86.0** ↑
WebQuestions	73.7	**75.5** ↑
WikiMovies	61.7	**70.3** ↑

Numbers show the ratio of questions for which answers appear in Top 5 articles returned by each system

Reader evaluation on SQuAD

Method	Exact Match	F1 Score
Dynamic Coattention Networks	65.4	75.6
Multi-Perspective Matching	66.1	75.8
BiDAF	67.7	77.3
DrQA	**69.5** ↑	**78.8** ↑

Ablation analysis of features

Features	F1 Score
Full	78.8
No f_{token}	78.0 ↓
No f_{exact_match}	77.3 ↓
No $f_{aligned}$	77.3 ↓
No $f_{aligned}$ and f_{exact_match}	59.4 ↓

Full Wikipedia Question Answering

Three versions of DrQA:

- SQuAD: A Document Reader model is trained only on the SQuAD training set

- Fine-Tune: A Document Reader model is pre-trained on SQuAD dataset and then fine-tuned for each dataset using DS

- Multitask: A Document Reader model is trained on the SQuAD dataset and all the DS sources

Full Wikipedia Results

Dataset	**YodaQA**	**DrQA**		
		SQuAD	Fine-Tune	Multitask
All Wikipedia	n/a	27.1	28.4	29.8
CuratedTREC	**31.3** ↑	19.7	25.7	25.4
WebQuestions	**39.8** ↑	11.8	19.5	20.7
WikiMovies	n/a	24.5	34.3	36.5

Numbers show exact-match accuracy

Bibliography

- Chen, Danqi; Fisch, Adam; Weston, Jason; Bordes, Antoine (2017). "Reading Wikipedia to Answer Open-Domain Questions". arXiv: 1704.00051

- “Question answering” – Wikipedia [Electronic resource], URL: https://en.wikipedia.org/wiki/Question_answering

- “DrQA” – GitHub [Electronic resource], URL: https://github.com/facebookresearch/DrQA