Document Haystacks: Vision-Language Reasoning Over Piles of 1000+ Documents

Supplementary Material

Evaluation Prompt

Task: You are an evaluator. Compare the Predicted Answer with the True Answer and determine if the Predicted Answer is Correct or Incorrect.

Instructions:

- If the Predicted Answer provides the same information or a reasonable interpretation of the True Answer, respond with `Correct`.
- If the Predicted Answer does not match or does not reasonably interpret the True Answer, respond with `Incorrect`.

Important: Answer only with `Correct` or `Incorrect`—no explanations.

Input:

- **Question:** { }
- **True Answer:** { }
- **Predicted Answer:** { }

Figure 1. **The Designed Prompt for GPT Evaluation.**

Which year CLAUD T.CARNEY worked at Windsor beet lab?

Label: 1910

Answer: Claude T. Carney worked at the Windsor beet lab in 1910.

GPT: correct **ANLS**: incorrect **Exact Match**: incorrect

Who is the president of First National Johnstown?

Label: arthur g. salberg

Answer: The president of First National Johnstown is Arthur G. Salberg.

GPT: correct **ANLS**: incorrect **Exact Match**: incorrect

Figure 2. **Zero-Shot VQA without using Task-Specific Prompt.** Without limiting the output space, traditional metrics cannot evaluate model performance even when the model's answer is correct.

A. Evaluation

GPT-based evaluation vs. traditional evaluation. In the open visual question answering (VQA) task, the model can generate the answer in diverse format and it is hard to accurately evaluate the generated answer.

Traditional evaluation metric such as, *Exact Match*, measures the percentage of questions where the predicted answer exactly matches one of the target answers, giving a score of zero even when the prediction is only slightly different from the correct answer. The issue with this evaluation is that language is inherently flexible, and there can be various texts to express the same idea (*e.g.*, "the dog is sleeping" vs. "a sleeping dog"). This flexibility often results in the *Exact Match* metric failing to capture the true capability of the model sometimes. As shown in Fig. 2, we show examples of zero-shot predictions from Qwen2-VL, generated without any specialized prompt design. It is clear that the *Exact Match* metric fails to evaluate the accuracy of responses in this setting. To address this, we seek a more reasonable evaluation metric that accounts for this linguistic flexibility. With the recent advancements in LLMs, GPT-based evaluations are growing in popularity. due to their closer alignment with human behavior in interpreting language. We carefully design an evaluation prompt for GPT to score the predicted answer against the true answer. The evaluation prompt is structured in Fig 1.

We validate the consistency of the proposed GPT-based evaluation with human judgments, aiming to demonstrate its effectiveness compared to traditional metrics, *i.e.*, *Exact Match* and *ANLS*. Note that for *ANLS*, we consider a similarity score greater than 0.8 as correct. As illustrated in Fig. 3, traditional metrics can sometimes misjudge zero-shot prediction. This limitation becomes evident in scenarios where the flexibility of responses is crucial.

Note that we explicitly ask the model to "Answer the question using a single word or phrase." [1] This prompt ensures brevity and facilitates a fair comparison between predicted and target answers.

We report the performance of different evaluation metrics on DocHaystack-1000 and InforHaystack-1000 in Tab. 1. We mannually evaluate the performance of the different evaluation metrics and report the accuracy. The results show that the introduced GPT-based metric has 100% accuracy, outperforming traditional metrics between 3.23%

[1] https://github.com/EvolvingLMMs-Lab/lmms-eval

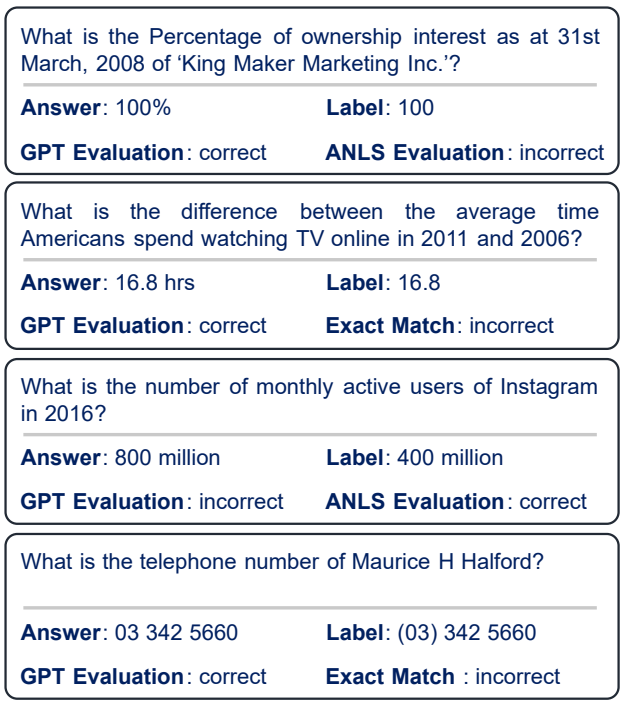


Figure 3. **Zero-Shot VQA with using Task-Specific Prompt.** Even when the output space is limited, traditional metrics sometimes fail to evaluate model performance correctly, even if the model's answer is correct.

and 6.42%. Based on these experimental results, we choose GPT-based evaluation for better accuracy.

Evaluation Metric	DocHaystack 1000	InfoHaystack 1000
Exact Match	93.6	96.8
ANLS	95.4	96.8
GPT	100	100

Table 1. **Performance of different evaluation metrics on DocHaystack-1000 and InfoHaystack-1000.**

B. Ablation on three-stage question filtering.

We ablate on the three-stage question filtering to valid the importance of selecting the unique questions. This three-stage question filtering is essential to ensure answer uniqueness and maintain the high quality of our benchmark. To assess its impact, we sampled an equal number of filtered examples and evaluated retrieval performance using V-RAG on DocHaystack-1000 and InfoHaystack-1000. As shown in Table below, the absence of filtering leads to a 39% drop in accuracy. This can demonstrate the usefulness of performing the three-stage question filtering.

	DocHaystack			InfoHaystack			Avg
w/ filtering	66%	78%	79%	65%	74%	78%	**73%**
w/o filtering	22%	30%	32%	29%	42%	46%	**34%**

Table 2. Retrieval accuracy on data without three-stage filtering.

Model	DocHaystack			InfoHaystack		
	100	200	1000	100	200	1000
Qwen2-VL-upper-bound	95.4	94.5	94.5	74.2	72.3	73.5
Qwen2-VL+V-RAG	82.6	74.3	66.1	65.8	65.8	60.0
Qwen2-VL-f.t.+V-RAG	86.2	79.8	73.4	67.1	67.7	60.0

Table 3. **The upper-bound of Qwen2-VL.** We evaluate the limits of Qwen2-VL's visual understanding using paired documents and questions on the DocHaystack and InfoHaystack benchmarks.

C. Upper-Bound Analysis

Tab. 3 presents the performance of Qwen2-VL-7B on the DocHaystack and InfoHaystack benchmarks using paired images and questions. We report the upper-bound performance of Qwen2-VL as a reference for its document understanding capabilities. As the number of retrieved documents increases (*i.e.,* retrieving the paired document from 100, 200, and 1000 documents), the Qwen2-VL's performance progressively deviates from its upper bound. This trend suggests that the primary challenge in multi-document understanding lies in accurately identifying the paired documents. The gap between the upper bound and the retrieval-augmented models highlights the current limitations in retrieval effectiveness, emphasizing the need for improved document retrieval mechanisms to enhance multi-document understanding

D. Image Retrieval

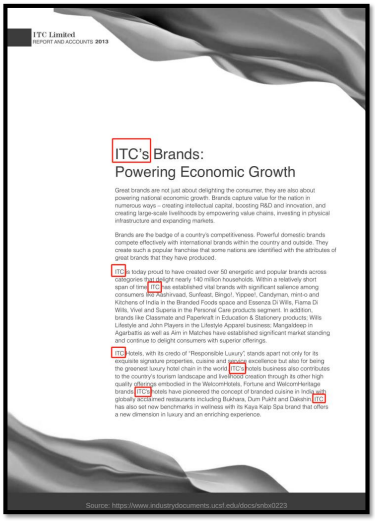
We select retrieved images based on their similarity to the question and present the three most relevant images retrieved using different methods (*i.e.*, V-RAG-based retrieval, CLIP-based retrieval, SigLIP-based retrieval, OpenCLIP-based Retrieval): Fig. 4 – 6 for DocHaystack and Fig. 7 – 9 for InfoHaystack. In these figures, the red box outside the retrieved image highlights the ground truth image paired with the question. The red box within each retrieved image shows the related information to the question. The yellow box in the ground truth image paired with the question shows the ground truth answer for the question. As can be seen in these figures, our proposed V-RAG performs well in question-related image retrieval.

Question: In which state is ITC's Watershed Development Project located?

Answer: Madhya Pradesh

Rank 1 Retrieval

Rank 2 Retrieval



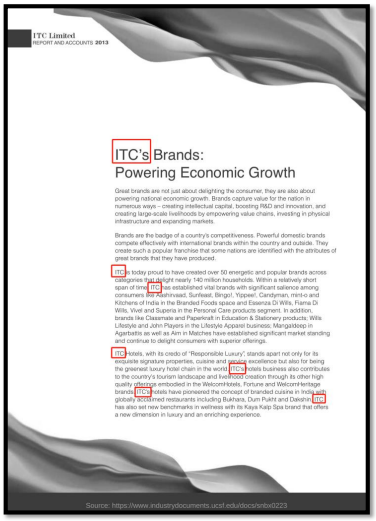
Rank 3 Retrieval

Figure 4. The three images most similar to the question retrieved using V-RAG in DocHaystack. The red box highlights the ground truth image paired with the question.

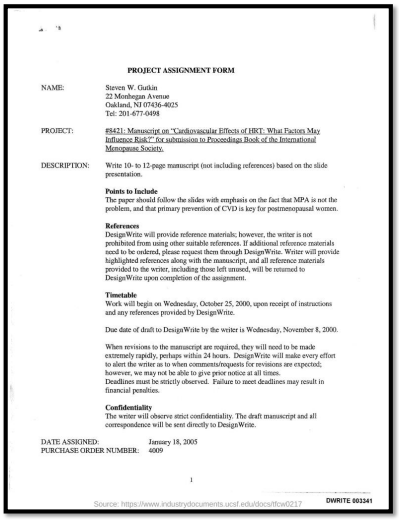
Question: In which state is ITC's Watershed Development Project located?

Answer: Madhya Pradesh

Rank 1 Retrieval



Rank 2 Retrieval



Rank 3 Retrieval

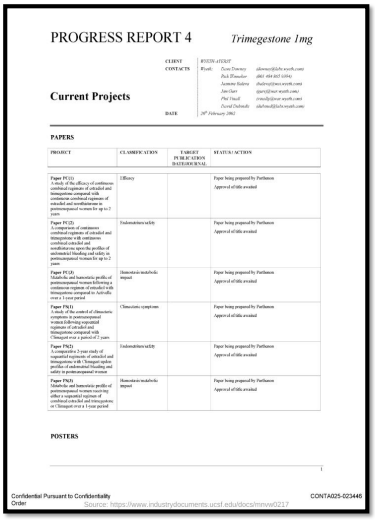


Figure 5. The three images most similar to the question retrieved using CLIP in DocHaystack. The red box highlights the ground truth image paired with the question.

Question: In which state is ITC's Watershed Development Project located?

Answer: Madhya Pradesh

Rank 1 Retrieval

Rank 2 Retrieval

Rank 3 Retrieval

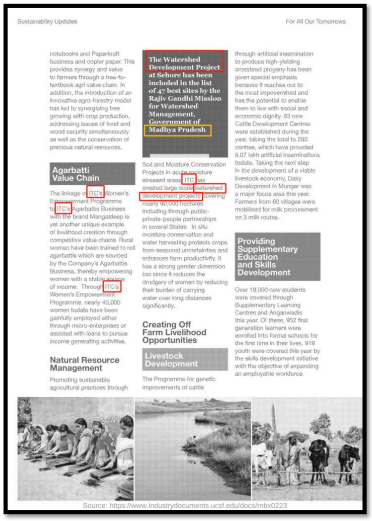


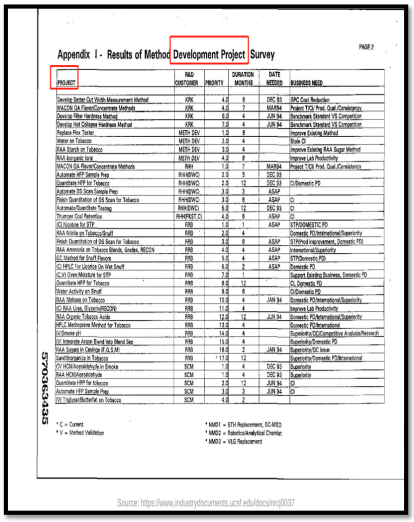
Figure 6. The three images most similar to the question retrieved using SigLIP in DocHaystack. The red box highlights the ground truth image paired with the question.

Question: In which state is ITC's Watershed Development Project located?

Answer: Madhya Pradesh

Rank 1 Retrieval

Rank 2 Retrieval



Rank 3 Retrieval

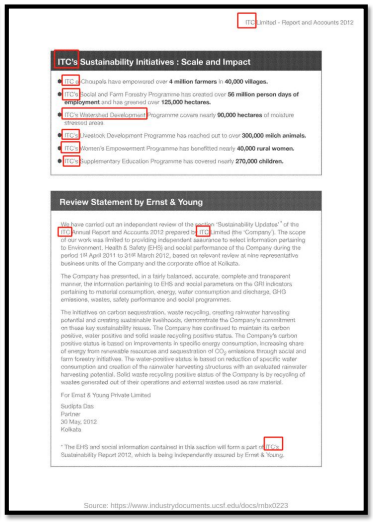
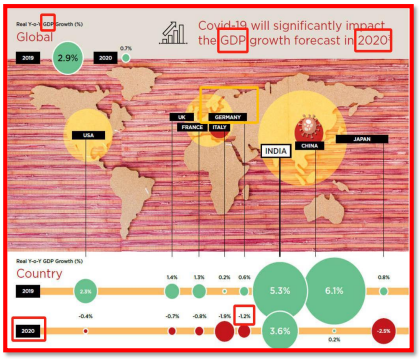


Figure 7. The three images most similar to the question retrieved using OpenCLIP in DocHaystack. The red box highlights the ground truth image paired with the question.

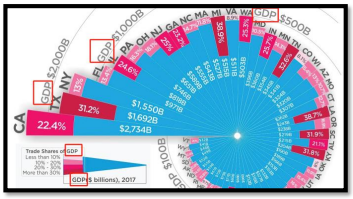
Question: Which country's GDP growth rate is -1.2% in 2020?

Answer: GERMANY

Rank 1 Retrieval



Rank 2 Retrieval



Rank 3 Retrieval

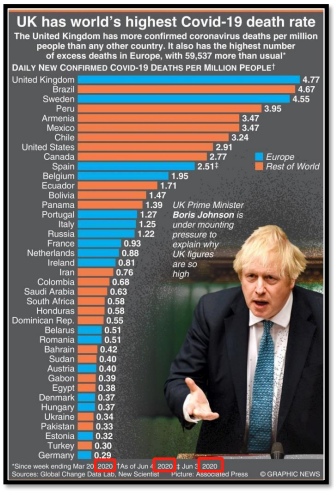
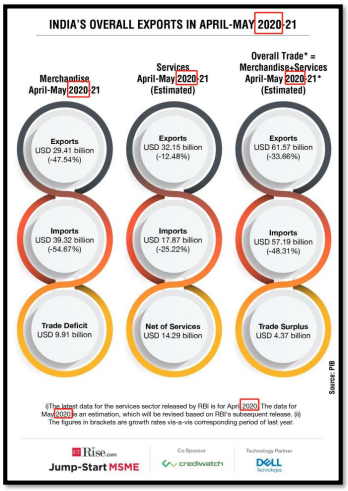


Figure 8. The three images most similar to the question retrieved using V-RAG in InfoHaystack. The red box highlights the ground truth image paired with the question.

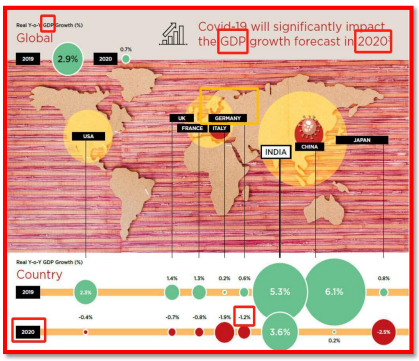
Question: Which country's GDP growth rate is -1.2% in 2020?

Answer: GERMANY

Rank 1 Retrieval



Rank 2 Retrieval



Rank 3 Retrieval

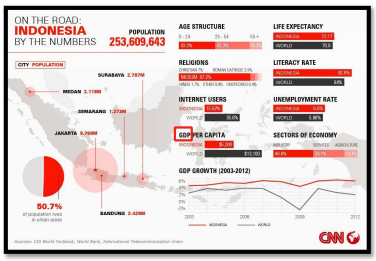
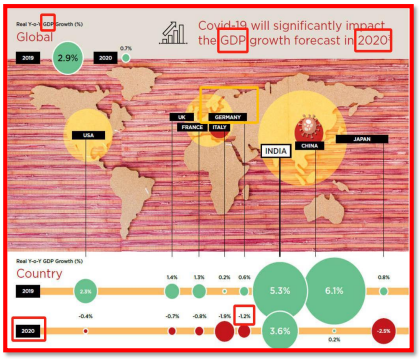


Figure 9. The three images most similar to the question retrieved using CLIP in InfoHaystack. The red box highlights the ground truth image paired with the question.

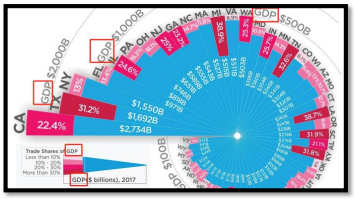
Question: Which country's GDP growth rate is -1.2% in 2020?

Answer: GERMANY

Rank 1 Retrieval



Rank 2 Retrieval



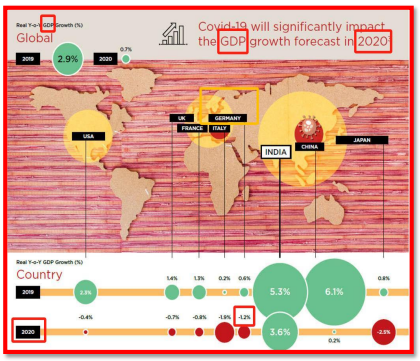
Rank 3 Retrieval

Figure 10. The three images most similar to the question retrieved using SigLIP in InfoHaystack. The red box highlights the ground truth image paired with the question.

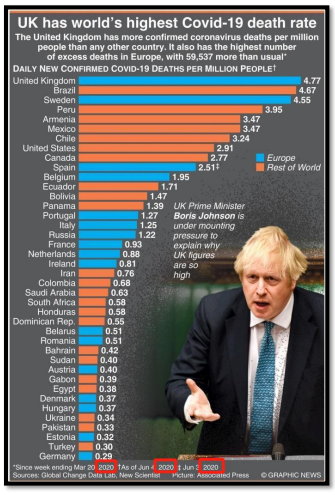
Question: Which country's GDP growth rate is -1.2% in 2020?

Answer: GERMANY

Rank 1 Retrieval



Rank 2 Retrieval



Rank 3 Retrieval

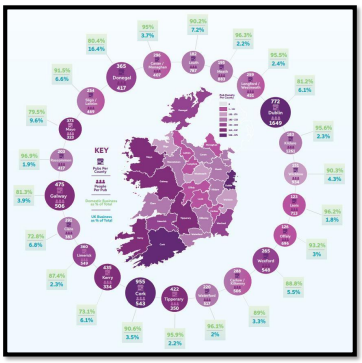


Figure 11. The three images most similar to the question retrieved using OpenCLIP in DocHaystack. The red box highlights the ground truth image paired with the question.

E. Question Filtering Pipeline

General questions can typically be answered from multiple documents, with several possible correct answers. For example, the question "Who wrote the letter" is a general question that can be answered by any document containing a letter. Generic Knowledge refers to information or facts that are widely accessible and can be answered using general world knowledge, often independent of specific visual or contextual cues from accompanying content, such as images. For example, the question "How many events were featured in the 2014 Winter Olympics?" is a generic knowledge that can be answered without accessing any image. In visual question answering tasks such as DocVQA and InfographicVQA, generic knowledge introduces a language bias when large language models (LLMs) rely on pre-existing knowledge rather than visual content, thereby undermining the focus on image-based reasoning. Therefore, it is important to exclude such questions to evaluate the true image-based reasoning capability of models. In this section, we show how we filter the data to extract the specific question.

E.1. General Question LLM Filtering

First, we leverage an LLM to filter out the general questions. This approach allows for the automatic filtering of numerous general questions. We give some filtered general questions by LLM here.

DocHaystack
- What does C stand for?
- What is the receiver number?
- What is the zip code?
- What is plotted along the x axis ?
- Who wrote the letter?
- What is the Fund No.?
- What type of report is this?
- Who is the client?
- What is the description?
- What is the name of the company?

InfoHaystack
- Who is the player in this picture?
- How many salary caps are mentioned?
- What percentage are not children?
- What percentage are not Americans?
- How many resources are listed?
- How many employers were surveyed?
- How countres are listed here in total?
- In which school did he study?
- What is the second last solution given?
- What is written in the yellow circle?

E.2. General-question manual review

However, the LLM-based filtering is not entirely accurate. Therefore, in the second stage, we involve manual filtering, where annotators are tasked with filtering out any general questions that were missed by the LLM. The filtered general question by human are as follow.

DocHaystack
- What time is the 'coffee break'?
- What is the year of publication?
- What is the name of the person on the from?
- Which is the root node in the chart?
- What is the no of cut tobacco?
- What is the name in the letter head?
- What is the exit date from china?
- What is the first person name marked in CC?
- What is the progress Report number?
- In which country is the company located?

InfoHaystack
- Where is open carry not permitted?
- How many points should protection services include?
- How many sharing tools mentioned in this infographic?
- What percentage of the survey participants are female?
- How many products are associated with blue color?
- What is the color mode used for the Web?

- Who are the swimming players in the list?
- Who made this infographic?
- How many symptoms are shown?
- Who is represented by green colour?

E.3. Generic knowledge filtering

After filtering out the general questions, we leverage GPT to filter out the remaining questions that can be answered by generic knowledge. The following is a list of some filtered questions.

DocHaystack

- What is the PO box no. of Biomet Orthopedics, Inc.?
- What is the Fax number of 'Brookstown Inn'?
- What is the location for Endocrine society-ENDO 2004 meeting?
- Which 'meeting ' was held at New Orleans, LA in January 2004?
- For which groups, WEFA, Inc. conducted study in April 1998?
- When was Argonne National Laboratory study for Department of Energy conducted?
- Where was the NAMS(North American Menopause Society)'s 14th Annual Meeting?
- What is ITC's brand of Agarbatti?
- Which 'meeting ' was held at Miami Beach, FL in May 2003?
- What is the brand name of ITC's snack food?

InfoHaystack

- What percent of world's adults have a bank account in the year 2014?
- What was India's score in the 2011 cricket world cup final?
- What percentage of Apple's revenue comes from iPhone in 2016?
- How many teams participated in the 2011 ICC Cricket World Cup?
- How many events were featured in the 2014 Winter Olympics?
- Which state is the second-largest producer of the Christmas tree in the U.S. in 2008?
- How many atomic bomb attacks were made by the U.S. in Japan in 1945?
- How many times did Hilary Mantel win the Booker prize?
- When was the 2011 cricket world cup final?
- What was Microsoft's net income in 2018?

E.4. DocHaystack Final 20 Random Questions

We randomly sample 20 questions in our final list of DocHaystack as below.

- Which was CLAUD T.CARNEY's high school?
- Which ITC Brand has 'Liquid Crystal Freezing Technology'?
- When was the study of Charles River Associates done?
- In which office does Michael Shapiro work?
- In which state is ITC's Watershed Development Project located?
- Who is 'presiding' TRRF GENERAL SESSION (PART 1)?
- How many nomination committee meetings has S. Banerjee attended?
- What is the number of Investor Services Committee meetings attended by A. V. Girija Kumar?
- How many children does George E. Wilber. Jr. have?
- Who is the R&D customer for the project "Water on Tobacco"?
- What is the Box number of "University of Florida"?
- What is the phone number of CARR SMITH?
- Who is the president of CPC International Inc?
- Who is the senior vice president and general counsel of RJR tobacco company?
- Which year CLAUD T.CARNEY worked at Windsor beet lab?
- What the location address of NSDA?
- How much was the 1988 estimated expenditure committed for System buy-out- PGA Tour in the VANTAGE GOLF OPERATIONS?
- What is the percentage of families in Poverty in Henry county?
- Who is the Chairman of 'Wembley Western Australia"?
- Who is the author for publication "Climacteric"?

E.5. InfoHaystack Random 20 Final Questions

We randomly sample 20 questions in our final list of InfoHaystack as below.

- What was the only media for watching Team USA events live in 2008?
- what is the total runs scored by Pietersen and Collingwood for England in 2007?
- What was the ratio of the U.S. population to bank branches in 1970?
- What is the number of Flickr users worldwide as of Nov. 15, 2012?
- How many lesbian & bisexual women (per 1,000 population) in Canada experienced violence in 2014?
- How many U.S. personnels were killed during the attack at Pearl Harbor?
- What percent of analytics jobs in India requires more than 5 years of experience according to the 2017 study?
- what is the number of cosmetic procedures done in Japan in millions in 2011?
- How many nonlethal gunshot wound cases were reported in America in 2009?
- What percentage of people in the U.S use social media

several times a day in 2009?
- What was the number of employees in Hemlow in 2002?
- How many actors acted in the series ”How to make it in America”?
- What is the number of tickets sold (in millions) in the 2012 London Olympic Games?
- How many US households were accessing bank accounts online as per the online banking report in Jan, 2012?
- What is the number of monthly active users of Instagram in 2016?
- In which two years did J. G. Farrell win the Booker prize?
- How many countries tested their first nuclear bomb after 2000?
- How many Florida soldiers died in the Afghanistan & Iraq war were men?
- What percentage of people in the U.S have a social networking profile in 2010?
- Length of what is specified for MQ-8 Fire Scout?