



INFORMATION  
SERVICES

Enterprise Architecture

# Building an enterprise Natural Language Search Engine with Elasticsearch and Facebook's DrQA

Louis Baligand, Debmalya Biswas

---

Berlin Buzzwords, 17 June 2019

# About

Debmalya Biswas



Louis Baligand



<https://github.com/philipmorrisintl>

The screenshot shows the GitHub profile for Philip Morris International. At the top, there is a profile picture of the company logo, the name "Philip Morris International", and location information: "Switzerland" and "https://www.pmi.com/". Below this, statistics are shown: "Repositories 15", "People 20", "Teams 4", and "Projects 0". A search bar for repositories is present, along with filters for "Type: All" and "Language: All", and a "New" button. Two repositories are listed:

- NPAModels**: R package and data for NPA models. Tags: bioinformatics, r, bioconductor, networks. Stats: 0 forks, 0 stars, 1 issue, 0 pull requests. Updated 18 days ago.
- NPA**: Network Perturbation Amplitude. Tags: bioinformatics, r, bioconductor, network-visualization, bif, npa, network-perturbation-amplitude. Stats: 0 forks, 3 stars, 0 issues, 0 pull requests. Updated 19 days ago.

On the right side, there are three sections:

- Top languages**: Python, R, Shell, HTML, Smarty.
- Most used topics**: bioconductor, bioinformatics, r.
- People**: 20 people.



*Forrester defines cognitive search and knowledge discovery solutions as*

*A new generation of enterprise search solutions that employ AI technologies such as natural language processing and machine learning to ingest, understand, organize, and query digital content from multiple data sources.*



*"The average interaction worker spends [...] nearly 20 percent (of the workweek) looking for internal information."  
-MGI Report, 2012.*

*Half (54%) of global information workers said, "My work gets interrupted because I can't find or get access to information I need to complete my tasks" a few times a month or more often. -Forrester Data Global Business Technographics Devices And Security Workforce Survey, 2016.*

# Enterprise Search vs. Web Search

## Enterprise Search

Small amount of content

Employees are the end-users

Multiple content types

Limited tagging/metadata management

Role-based content trimming

No team in charge of Search Experience

vs.

## Web Search

Enormous amount of content

WWW users

Single source (web pages)

Large investments in SEO (\*)

No visibility restrictions (public pages)

Search xxperience as core business

(\*): Search Engine Optimization

# Natural Language Search (NLS)



What is the most dangerous part of a fire?



Volume: 0/mo | CPC: \$0.00 | Competition: 0 ★

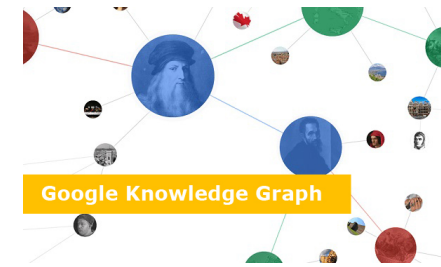
All Images News Maps Videos More Settings Tools

About 340,000,000 results (0,57 seconds)

Smoke can be the **most dangerous part** of a house **fire**. MEMPHIS, TN (WMC) - The **fire** that killed nine people Monday morning only burned 25 percent of the home, but the smoke filled the entire building, according to Memphis **Fire** Department. Sep 12, 2016

[Smoke can be the most dangerous part of a house fire](https://www.wmcactionnews5.com/.../smoke-can-be-the-most-dangerous-part-of-a-hous...)  
<https://www.wmcactionnews5.com/.../smoke-can-be-the-most-dangerous-part-of-a-hous...>

About this result Feedback



## Knowledge Graph



When was obama's wife born?



Volume: 0/mo | CPC: \$0.00 | Competition: 0 ★

All News Images Videos Shopping More Settings Tools

About 41,200,000 results (1,21 seconds)

Michelle Obama / Date of birth

January 17, 1964

age 55 years



People also search for



Barack Obama  
August 4, 1961



Donald Trump  
June 14, 1946



Melania Trump  
April 26, 1970

Feedback

# Chatbots and Natural Language Search

## (Rules based) FAQs

- Works only for specific hardcoded questions.
- The only way to scale with respect to question variants, is to extend the knowledgebase by manually adding variants of a question.

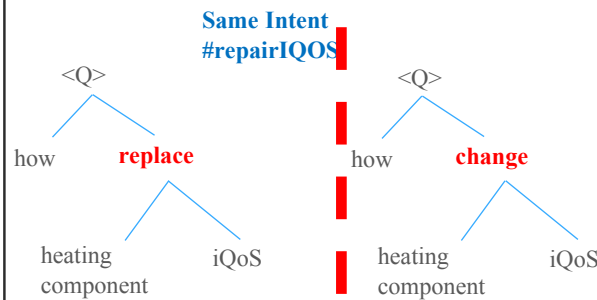
"How do I replace the heating component of my iQoS?"

=

"Tell me how to change the heating component of my iQoS"

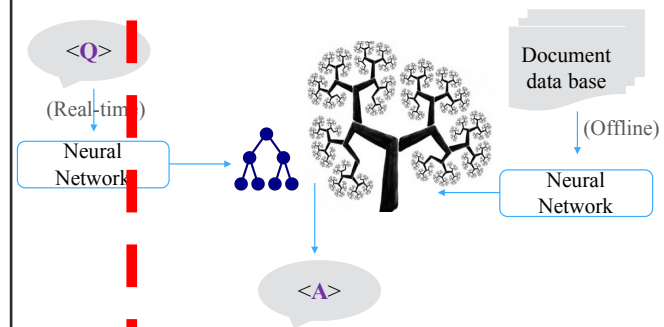
## Intent based Chatbots (Statistical Methods)

- Requires Q&A knowledge.
- Able to scale with respect to question variants by applying Statistical Clustering Methods, e.g. tf-idf, Bag-of-Words - to cluster question variants into 'intents'.



## Natural Language Search (Neural Networks)

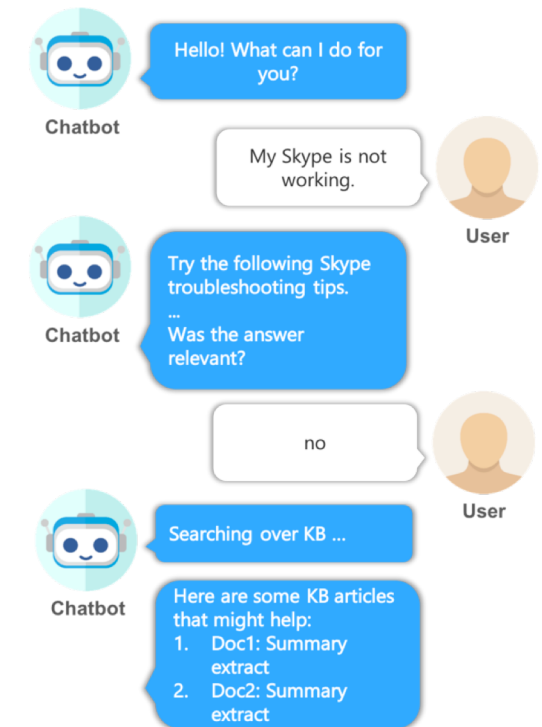
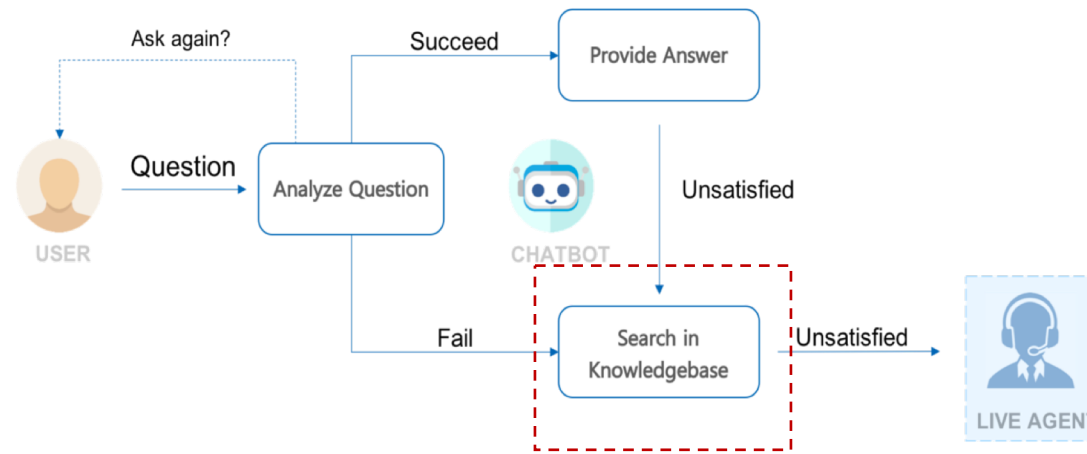
- Works on documents.
- Users can ask any question from the documents.
- Both the documents and questions are passed through the same Neural Network, producing the matching answer.



# Chatbots and Natural Language Search (2)

## 3- tier strategy:

- A Chatbot with its pre-defined Q&A set remains the entry point – think of it as the 1st line of defense.
- If the bot encounters a user query which cannot be mapped to one of its pre-configured intents, it performs a NLS over its KB. This is the 2nd line of defense.
- If the user is not satisfied even with search results, plan for a final handover to a live agent.





# Positioning vs e-commerce search

- End-user searching for products (not answer)
- Filter-Oriented
- Rates, Review

The screenshot shows the Alibaba.com search results for 'headphones'. The page features a navigation bar with the Alibaba.com logo and various utility links. The search bar contains the term 'headphones' and shows related search suggestions. Below the search bar, there are filter options for 'RELATED CATEGORY' (Earphone & Headphone) and 'FILTER RESULTS BY' (Supplier Types, Supplier Location, Sample Order, Min Order, Price). A 'SUGGESTED FILTERS' section displays icons for various product attributes like Microphone, Bluetooth, Waterproof, Portable, Music Streaming, Wireless, Wired, In-Ear, Headband, Neckband, Foldable, and Ear Hook. The main results area shows 234,995 results for 'headphones' with a price filter set to '< 100'. Several product listings are visible, each with an image, title, price, and supplier information. For example, one listing is for 'Baseus 2019 New Encok S17 Wireless...' priced at US\$26.96 per piece. Another is for 'Custom Logo OEM Sport Mini TWS Stereo Headset...' priced at US\$2.75-US\$5.25 per box. A third is for '2019 wired headphones stereo headset airplane' priced at US\$2.30-US\$3.00 per piece. A fourth is for 'Branding OEM headphone QCC 3020 chip wireless sport earphone bluetooth TWS...' priced at US\$18.00-US\$19.50 per piece. The page also includes a 'Premium Related Products' section on the right side.

# Philip Morris' Use case: Operator Trainings

- Hundreds to thousands of operators
- Long manuals with specific terminology
- A 1min downtime of a machine would lead to 20,000 cigarettes unmade
- Typical Full text Search (Boolean search, no relevancy score)
- Document Management System Manually classified
- On-boarding difficulty



# Example of fine-grained results

FILE TYPE	COUNT
PDF	798
MS-Office	434
Image	627
Video	57
Zip	73

Q. How many knives are there on the drums?

Type in your question here.

Q: How many knives are there on the drums?

A:

GD121 – AF12 OPERATORS TRAINING MANUAL OPERATION DEFINITION With always changing feeding speed by the tipping paper sliding pallet (31) and the glued tipping paper received from the glue application group (C), would be kept ready for cutting by the tipping paper drum (25). There are fourteen knives which are operating oppositely on the drums (24-25). The glue applied side of the tipping paper which has been fed upward, would be cut by the adjustable knives and would be transported to the next drums for the continuation of the process. The cut tipping papers which has been defined as waste would be transported to the waste pot located at the rear side of the machine. The necessary vacuum at the transporting moment would be provided by a separated fan. The paper should be fed at the same speed with the tipping paper drum (25) and knives during the cutting operation. At the cutting moment, the paper would be held and would be slung on the tipping paper drum vacuum plates for the reason of keeping the paper ready for the next cut. PAGE 250 PHILSA TECHNICAL TRAINING DEPARTMENT PREPARED BY:A.SEYHAN GÜRELLI Technical Trainer

Rank	Answer	
1	fourteen	<a href="#">1. Cigarette Making Equipment\5. GD 121\1.0 Operator\01.GD121-AF12 OPERATORS MANUAL.pdf (page 250 of 357)</a>
2	177	<a href="#">1. Cigarette Making Equipment\5. GD 121\1.0 Operator\01.GD121-AF12 OPERATORS MANUAL.pdf (page 5 of 357)</a>
3	15	<a href="#">1. Cigarette Making Equipment\5. GD 121\1.0 Operator\01.GD121-AF12 OPERATORS MANUAL.pdf (page 121 of 357)</a>

# Question Answering?

- Squad Dataset: a reference in Question Answering
- 100,000+ Q&A on Wikipedia articles
- State of the art is beating Human Performance

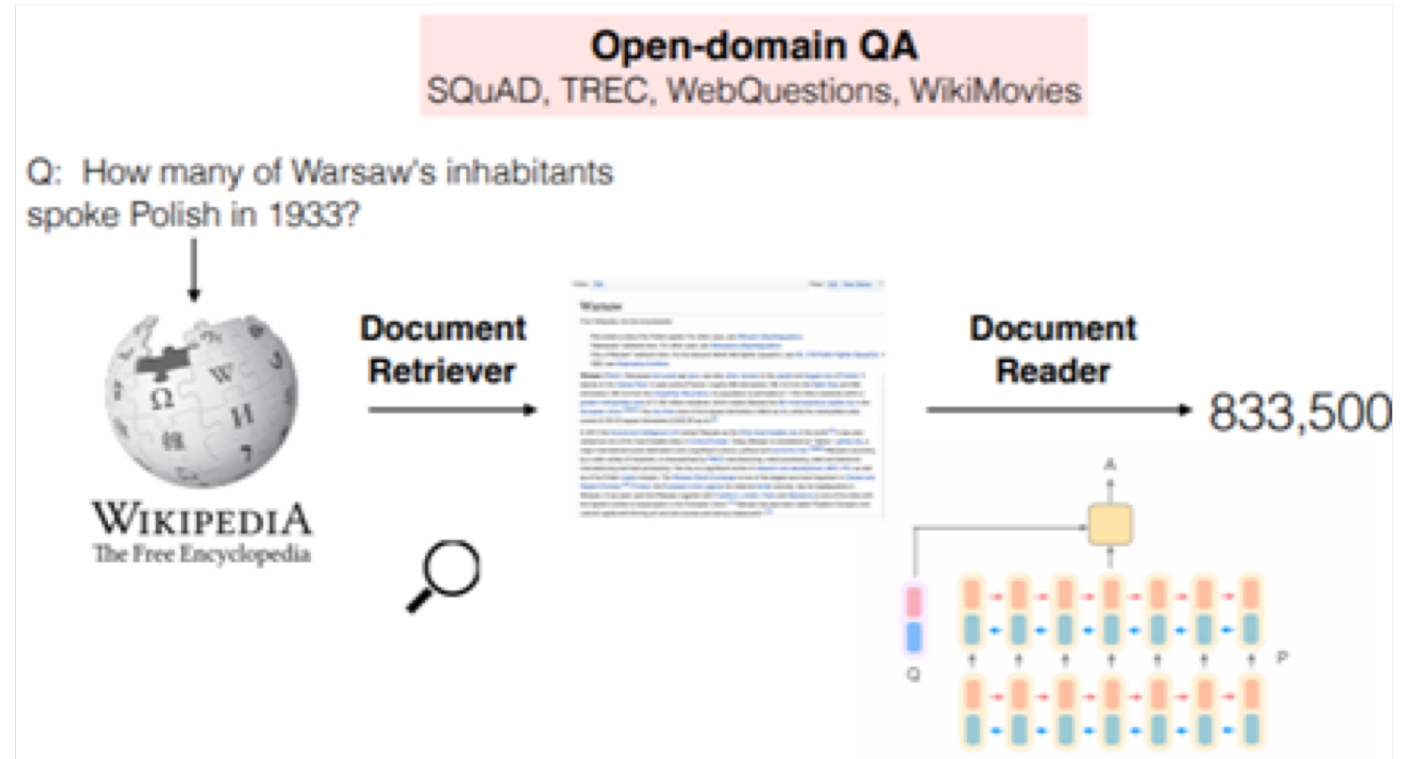
## SQuAD1.1 Leaderboard

These are the ExactMatch (EM) and F1 scores evaluated on the test set of SQuAD v1.1.

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar et al. '16)	82.304	91.221
1 May 21, 2019	XLNet (single model) XLNet Team	<b>89.898</b>	<b>95.080</b>
2 Oct 05, 2018	BERT (ensemble) Google AI Language <a href="https://arxiv.org/abs/1810.04805">https://arxiv.org/abs/1810.04805</a>	87.433	93.160
3 May 14, 2019	ATB (single model) Anonymous	86.940	92.641

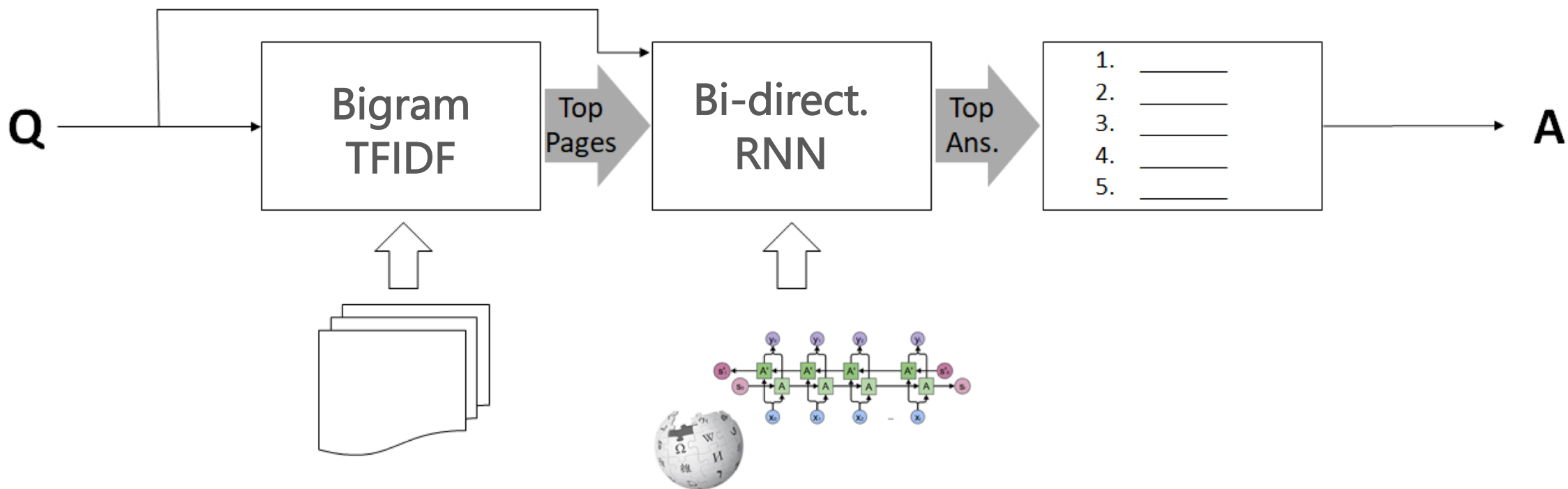
# DrQA Overview

- Facebook AI Research, ACL 2017, [Reading Wikipedia to answer Open-Domain Questions.](#)
- Open source, BSD License <https://github.com/facebookresearch/DrQA>
- Pre-trained model available



<https://github.com/facebookresearch/DrQA>

# DrQA Overview



# DrQA is easy to use on your own corpus!

```
$ python build_db.py /path/to/data /path/to/saved/db.db  
$ python build_tfidf.py /path/to/doc/db /path/to/output/dir
```



	Docs	
Terms	0.06	0.02
	0.03	0.08

Pre-trained model open sourced

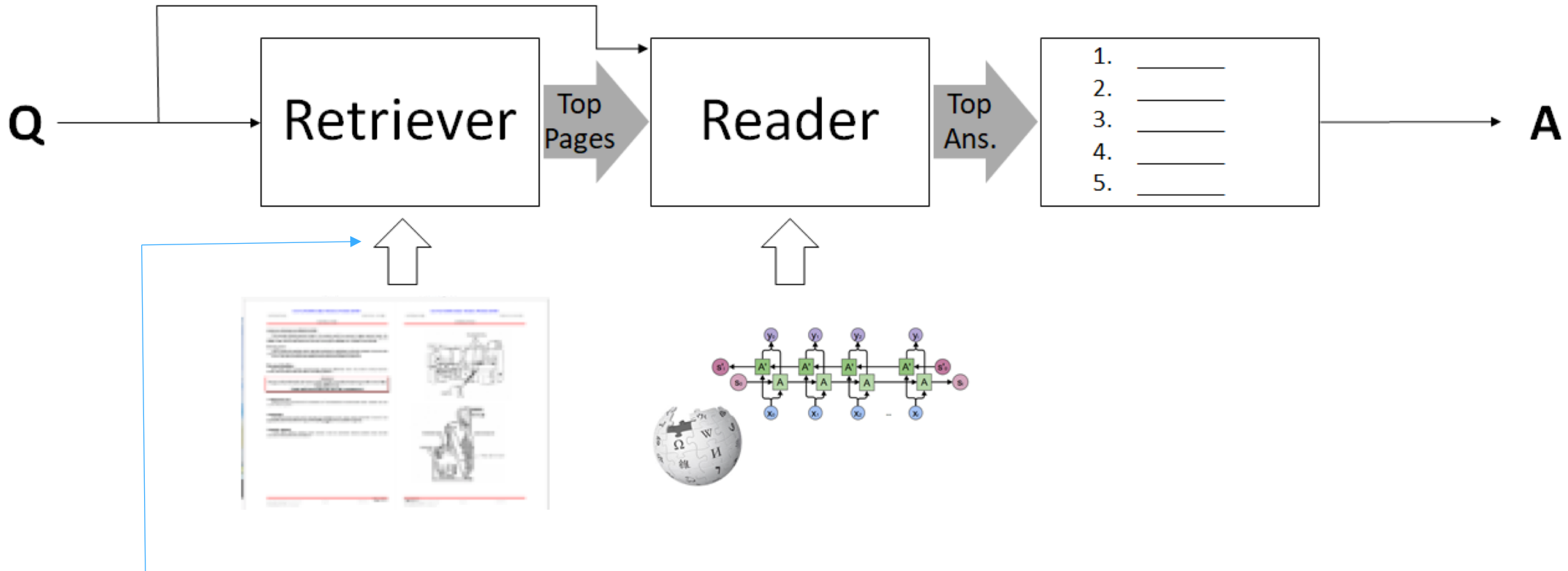
```
$ python interactive.py -reader-model multitask.mdl -retriever-model path/to/tfidf -doc-db path/to/saved/db.db
```

```
>>> 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

# DrQA to answer Operator's questions?



- Java toolkit to extract text + metadata from DOCX, PPT, XLS, PDF, JPEG, etc...
- Apache Software Foundation
- OCR

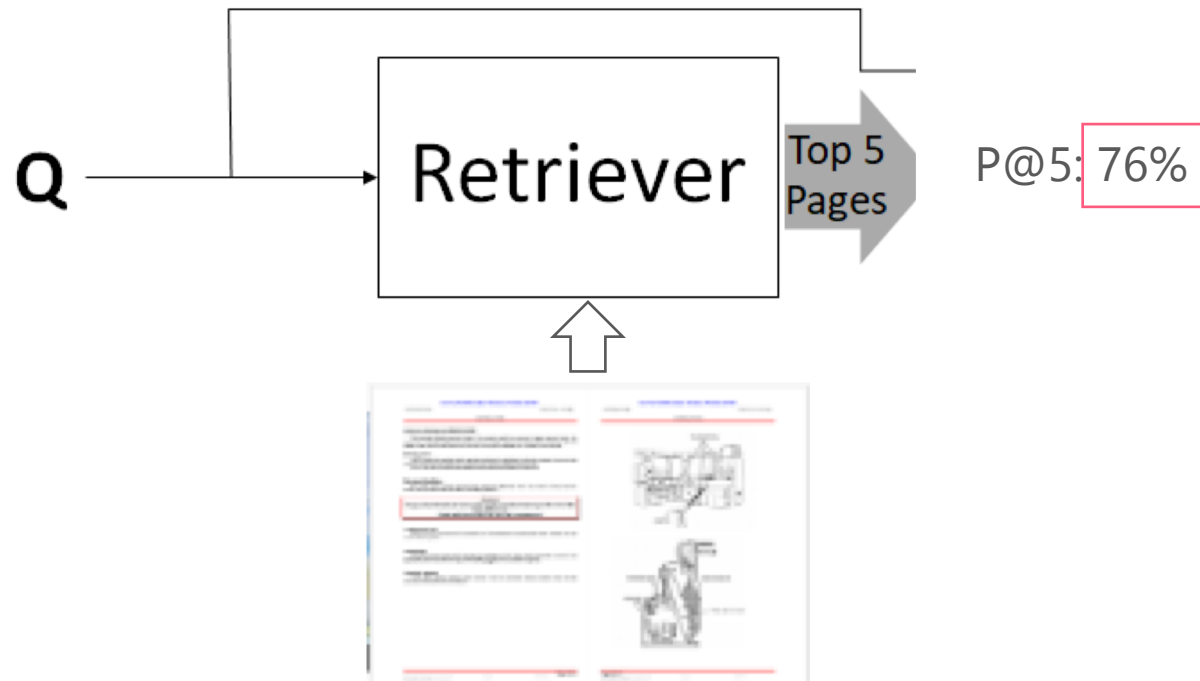


# DrQA to answer Operator's questions?

Model	SQuAD P@5	CuratedTREC P@5	WebQuestions P@5	WikiMovies P@5	Size
TF-IDF model	78.0	87.6	75.0	69.8	~13GB

*P@5 here is defined as the % of questions for which the answer segment appears in one of the top 5 documents.*

<https://github.com/facebookresearch/DrQA>



- Not a voice assistant
- End user needs at least ~95%
- Full control on the retriever
- First stage to prioritize

# Introducing Elasticsearch

---



- Open source distributed
- Highly scalable
- RESTful API on top of Lucene capabilities
- Support for Full Text search (best of breed)
- Easy to configure + extend
- Seamlessly manage conflicts
- Active community & popular

# Integrating Elasticsearch to DrQA's pipeline

facebookresearch / DrQA

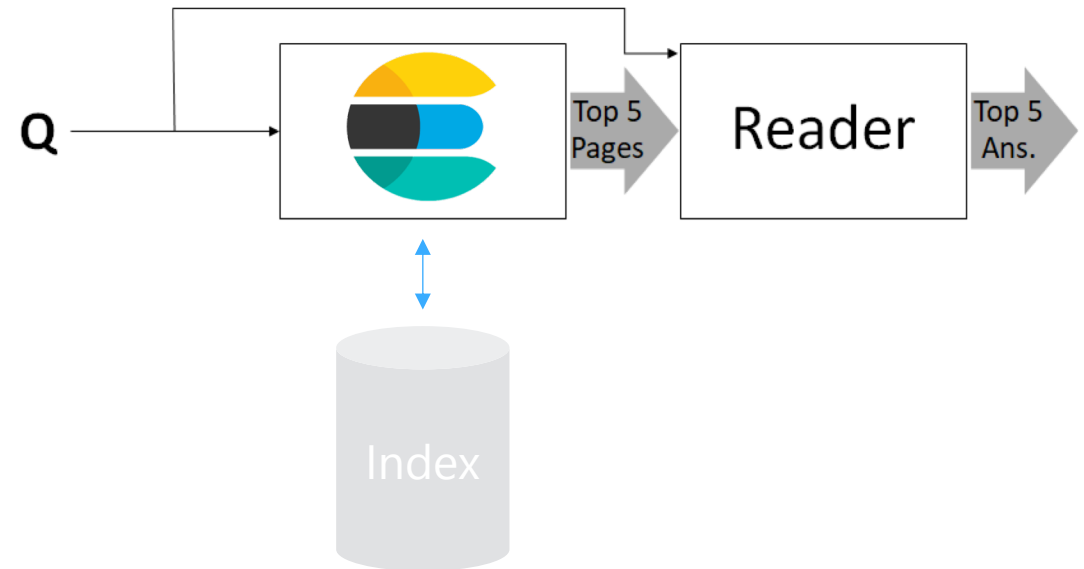
Watch 156 Star 3,239 Fork 683

Code Issues 37 Pull requests 1 Projects 0 Security Insights

Filters is:pr Labels 12 Milestones 0 New pull request

Clear current search query, filters, and sorts

1 Open	23 Closed	Author	Labels	Projects	Milestones	Reviews	Assignee	Sort
Update setup.py to fix setup bug	CLA Signed	#211 by GoMapur was closed on 2 Apr	6					
Fix prediction on CPU in DocReader.	CLA Signed	#206 by ousou was merged on 13 Mar	1					
Move to PyTorch 1.0	CLA Signed	#204 by ajfish was merged on 8 Mar						
Updated Code of Conduct File	CLA Signed	#201 by LakshKD was closed on 13 Mar	1					
Add code of conduct and contributing statements.	CLA Signed	#197 by stephenroller was merged on 14 Feb	3					
Update to new s3 locations.	CLA Signed	#195 by stephenroller was merged on 17 Jan • Approved	1					
Elasticsearch integration	CLA Signed	#191 by pykcel was merged on 8 Mar • Approved	8					
Master			2					



# Integrating Elasticsearch to DrQA's pipeline

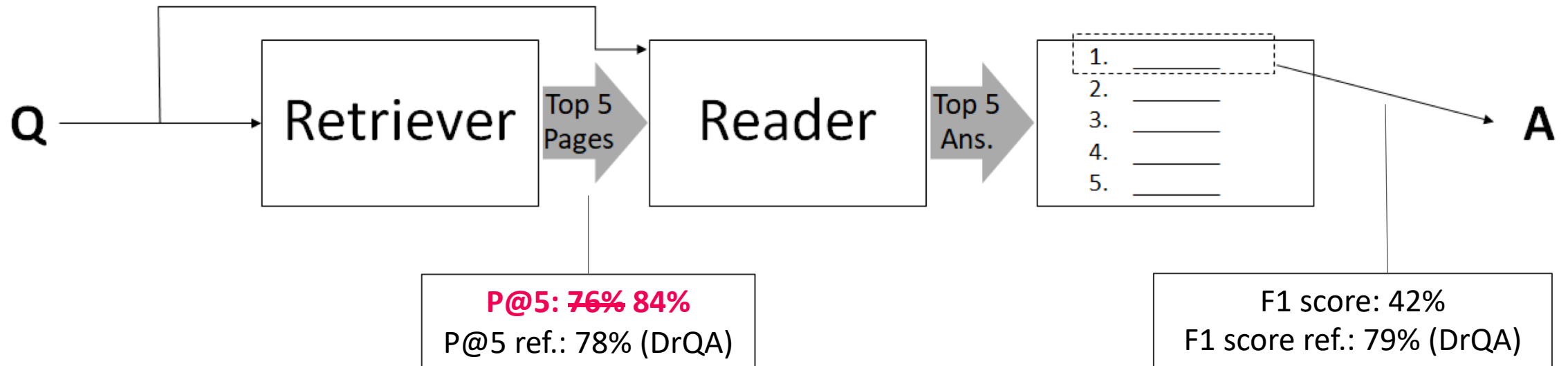
```
>>> from drqa.pipeline import DrQA
>>> from drqa.retriever import ElasticDocRanker
>>> model = DrQA(reader_model='reader_model.mdl',
                 ranker_config={'class':ElasticDocRanker,
                               'options':{'elastic_url':'127.0.0.1:9200',
                                         'elastic_index':'mini',
                                         'elastic_fields':'content',
                                         'elastic_field_doc_name':['file', 'filename'],
                                         'elastic_field_content': 'content'}})
>>> model.process('How the tensioning of the V-belts should be done?')
```

Directly point to your server hosting Elastic

Enable to search in any fields, e.g. uni-grams, bi-grams, title, metadata, etc...

```
29     "_index": "mini",
30     "_type": "doc",
31     "_id": "ae123782374bf93209fg",
32     "_score": 2.34598412931,
33     "_source": {
34         "filename": "manual_machine_example.pdf",
35         "last_modified": "2019-02-29T12:09:02.261+0000",
36         "author": "Louis Baligand",
37         "content": "Follow these steps to do the tensioning of the V-belts:"
```

# The pipeline performance



- DrQA span +/- 10 tokens: **94%** of 1<sup>st</sup> result contains true answer

# Take aways

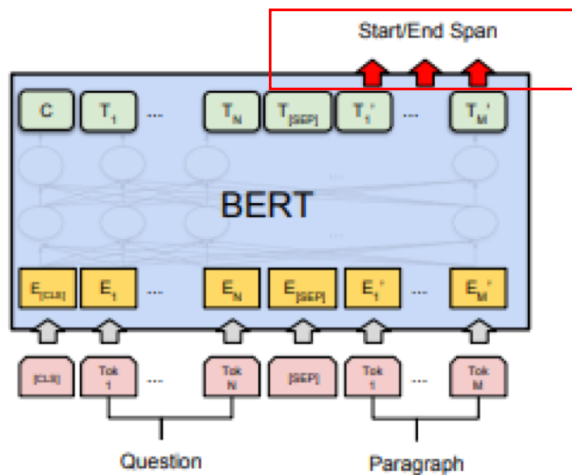
- Address pain points by combining Document Retrieval with Question Answering
- If not answered, it will provide much more granular insights of the data
- User elicitation & user experience: a top down approach
- End user does not know what to ask

The screenshot displays the 'Operations AI Search' interface. At the top, a search bar contains the query 'What is the capacity of the glue jet tank in Protos 70?'. Below the search bar, there are filters for 'Category' (listing options like Packaging Equipment, Filter Making Equipment, Tobacco Processing, and OPEN) and 'Extension' (listing file formats like pdf, doc, docx, ppt, zip, pptx, JPG). The search results section shows a single result with a score of 0.19. A detailed preview of the document is shown on the right, containing technical information about the 'OPERATION PROTOS 70 PM' and 'Gluing jet' system. The preview includes a table with the following data:

Read	Answer	Score
1	13 liters	0.04
2	approx. 13 liters	0.19
3	1) to glue feed nozzle under dead weight. Tank capacity is approx. 13 liters	0.02
4	10) Rotary piston Functional description The gluing system realizes gluing with gravitational glue feed. The glue flows from tank (1) to glue feed nozzle under dead weight. Tank capacity is approx. 13 liters	0.01
5	13	0.01

# Future work – Extend pipeline with BERT\*

- A general-purpose architecture to train models for multiple NLP tasks (sentiment analysis, etc...)
- State of the art for SQuAD
- Open source, published in Oct. 2018 by Google AI Research
- High memory required: GPU with at least 12GB of RAM (Base model)
- Enable to multi-language queries



- Add one layer to compute  $P_{start}(\text{"token"})$  &  $P_{end}(\text{"token"})$  for each tokens
- Find the best pair by maximizing  $P_{start}(\text{"token1"}) * P_{end}(\text{"token2"})$

\*<https://arxiv.org/abs/1810.04805>, <https://github.com/google-research/bert>

Thank you.