

Design + Code-Walkthrough

Search-engine for Q&A

AppliedAICourse.com

Agenda:

1. Pre-req

2. Problem definition & solution requirement

3. Dataset [+code]

4. High level design & Data-structures

5. Docker containerization [+setup]

6. Sentence-vectors [code]

7. Elastic Search [installation]

8. Index data [code]

9. Search [code]

10. Deployment [code]

11. Extensions & optional assignment

References:

Blog: <https://www.elastic.co/blog/text-similarity-search-with-vectors-in-elasticsearch>

Code: <https://github.com/jtibshirani/text-embeddings>

Terms related to ElasticSearch:

<https://www.elastic.co/guide/en/elasticsearch/reference/current/glossary.html>

Simple tutorial on Python+ES: <https://medium.com/naukri-engineering/elasticsearch-tutorial-for-beginners-using-python-b9cb48edcedc>

Pre-req:

1. Python programming

2. Inbuilt DS in python : lists, dict, tuples, sets

We will NOT be able to dive deep into
all the concepts

Problem - def:

→ Q^i
 A_1^i
 A_2^i
 \vdots
 A_K^i

→ fast (< 500 ms)

The screenshot shows the Stack Overflow website with a search bar containing the text "how to install pip". The search results are displayed on the right side of the page, showing 500 results. The first three results are visible:

- 2534 votes**, **37 answers**: **Q: How to install pip on Windows?**
pip is a replacement for easy_install. But should I **install pip** using easy_install on Windows? Is there a better way? ...
asked Jan 20 '11 by mit
- 552 votes**, **21 answers**: **Q: How to install pip with Python 3?**
I want to **install pip**. It should support Python 3, but it requires setuptools, which is available only for Python 2. **How** can I **install pip** with Python 3? ...
asked Jul 5 '11 by deamon
- 1077 votes**, **12 answers**: **Q: How to install packages using pip according to the requirements.txt file from a local direct...**
packages according to requirements.txt from the local archive directory. source bin/activate **pip install -r /path/to/requirements.txt -f file:///path/to/archive/** I got some output that seems to ... anyjson==0.3 ... I have a local archive directory containing all the packages + others. I have created a new virtualenv with bin/virtualenv

- high precision & recall
- low computational / server costs $\left[\begin{array}{l} \text{QPS:} \\ \text{\#servers:} \\ \text{cost per server:} \end{array} \right]$
- quick to build & deploy [< 7 days]
a basic version

Dataset

→ fields/columns

STACK SAMPLE: <https://www.kaggle.com/stackoverflow/stacksample>

→ DOWNLOAD

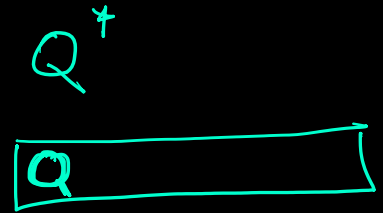
→ UNZIP xxxxx.zip

→ count # questions & answers [readData.py]

HLD + Dalā Structures

Q^i : id, title, body, ...

A^i : id, body, parentId, ...



design choices: Q^i title \rightarrow simplest
+ Q^i body
+ A^i body

$Q^i_{\text{title}}: w^i_1 w^i_2 w^i_3 \dots w^i_d$

Search \Rightarrow hashing $[O(1)]$

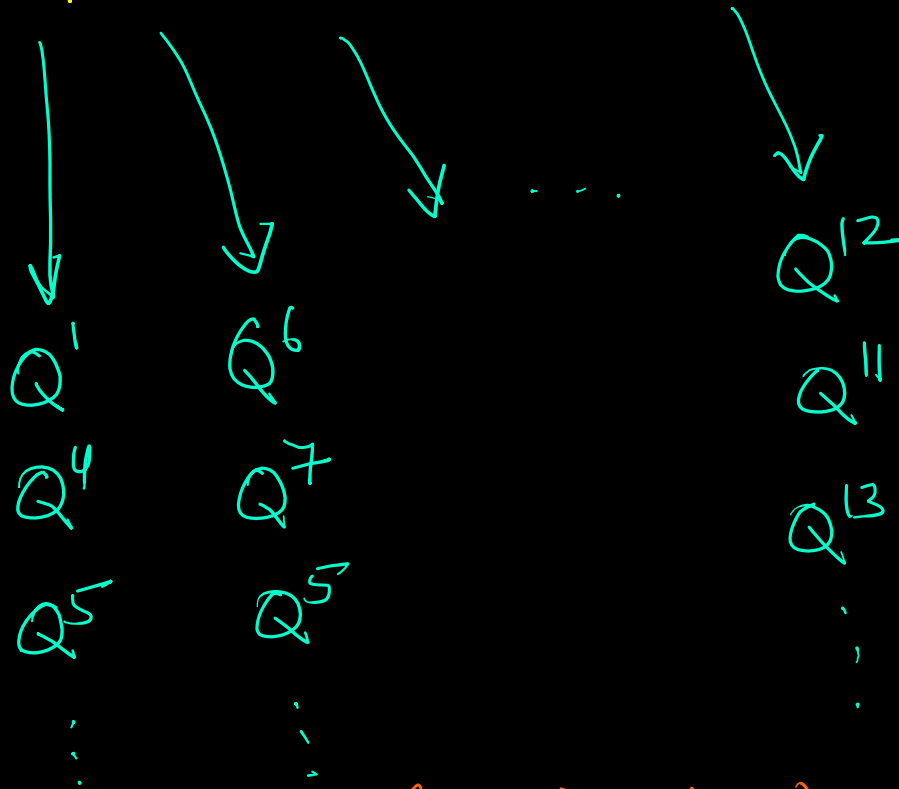
Inverted-Index:

pip \rightarrow $Q_2^1, Q_1^4, Q_1^{46}, Q_3^{32}, Q_1^{61}$ [5] \rightarrow # occurrences

of \rightarrow $Q_{10}^1, Q_{11}^2, Q_6^3, Q_{18}^4, \dots, Q_{12}^{68}$ [68] \rightarrow # docs containing the term

...

$$Q^+ : w_1^+ w_2^+ w_3^+ w_4^+ \dots w_k^+$$

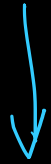


$$S = \{Q^1, Q^2, \dots, Q^k\}$$

hashing-based

$\text{score}(Q^*, Q^i)$

$\forall Q^i \in S$

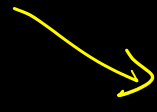


TF-IDF based



Term-freq

occurrences



Inverse doc-freq

eg: of, the, if, of ...

pip, install, ...

Elastic Search: Inverted Index, scoring, distributed,
realtime,

