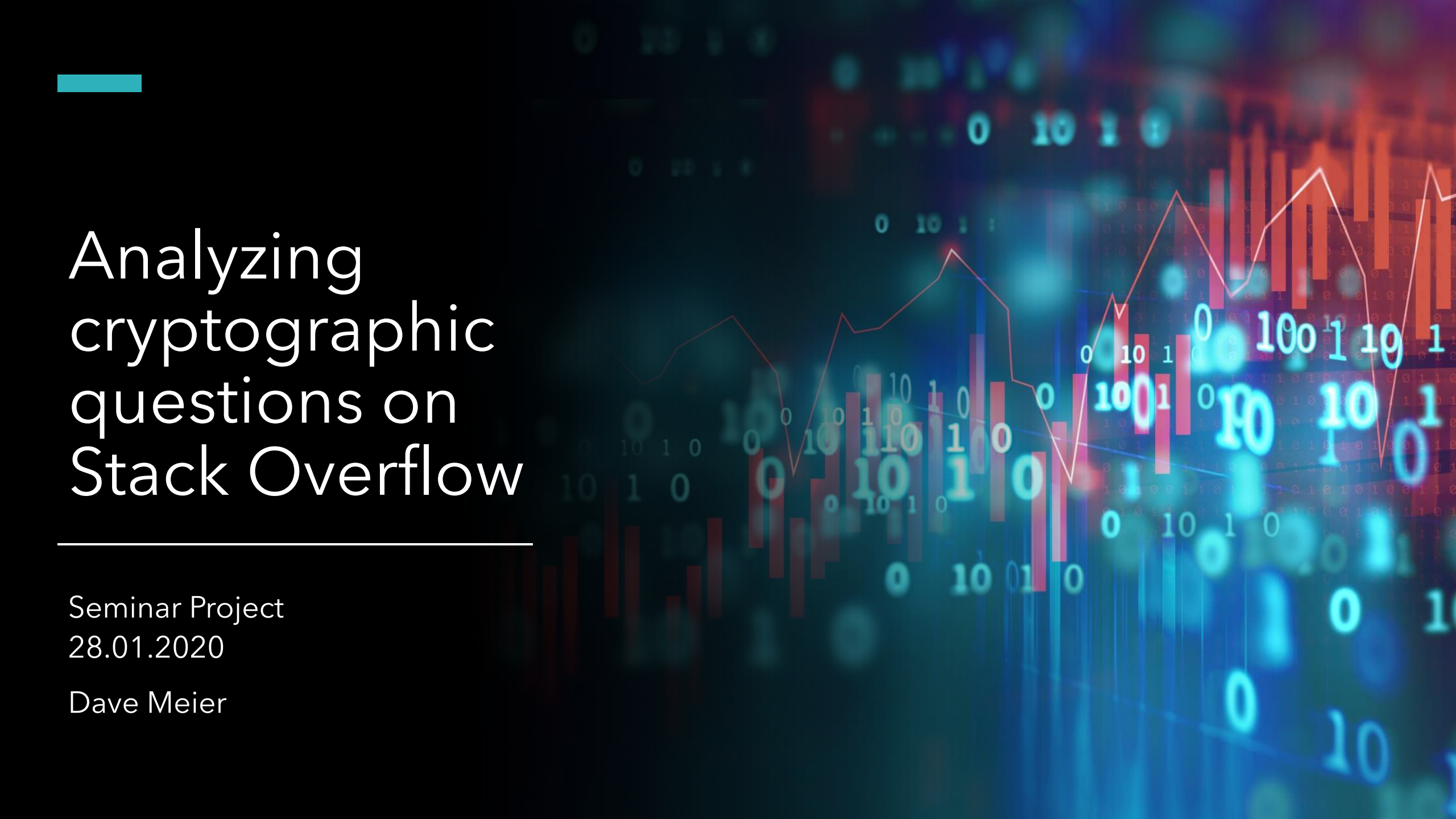
Analyzing cryptographic questions on Stack Overflow

Seminar Project
28.01.2020

Dave Meier



Research questions

- RQ1: What **topics** are covered by cryptography-related questions asked on Stack Overflow?
- RQ2: Which topics are the **most popular** among cryptography-related questions?
- RQ3: Which crypthography-related topics are the **most difficult** to answer?

Procedure

1. Gather data from Stack Overflow
2. Generate topics with LDA
3. Calculate metrics for RQ2 and RQ3

1. Gather data from Stack Overflow

2. Generate topics with LDA
3. Calculate metrics for RQ2 and RQ3

I. Download posts with tag «cryptography» (as CSV) from https://data.stackexchange.com/
II. Tag analysis to get tags related to «cryptography»
III. Download a separate CSV for each related tag

1. Gather data from Stack Overflow

StackExchange Data Explorer

Home | Queries | Users | Compose Query

Viewing Query

Enter a title for your query

edit description

stackoverflow
Q&A for professional and enthusiast programmers

```
1  SELECT questions.Id as [post_link],
2  questions.title as [title],
3  questions.body as [body_question],
4  answers.body as [body_answer],
5  answers.id as [answer_id],
6  questions.viewcount as [viewcount], questions.Tags as [tags],
7  questions.Score as [score], answers.Score as [score_answer],
8  questions.AnswerCount as [answercount], questions.LastActivityDate,
9  questions.AcceptedAnswerId, questions.CommentCount,
10 questions.FavoriteCount, questions.LastEditorUserId,
11 questions.LastEditorDisplayName,
12 questions.LastEditDate, questions.CreationDate, questions.ClosedDate,
13 answers.LastEditDate as [answer_LastEditDate],
14 answers.CreationDate as [answer_CreationDate]
15 FROM Posts questions
16 left JOIN Posts answers ON answers.parentid = questions.id
17 WHERE questions.Tags like '%<cryptography>%'
18 order by questions.viewcount DESC
```

RQ1

RQ2

RQ3

Database Schema

Posts

Id	int
PostTypeId	tinyint
AcceptedAnswerId	int
ParentId	int
CreationDate	datetime
DeletionDate	datetime
Score	int
ViewCount	int
Body	nvarchar (max)

Revisions

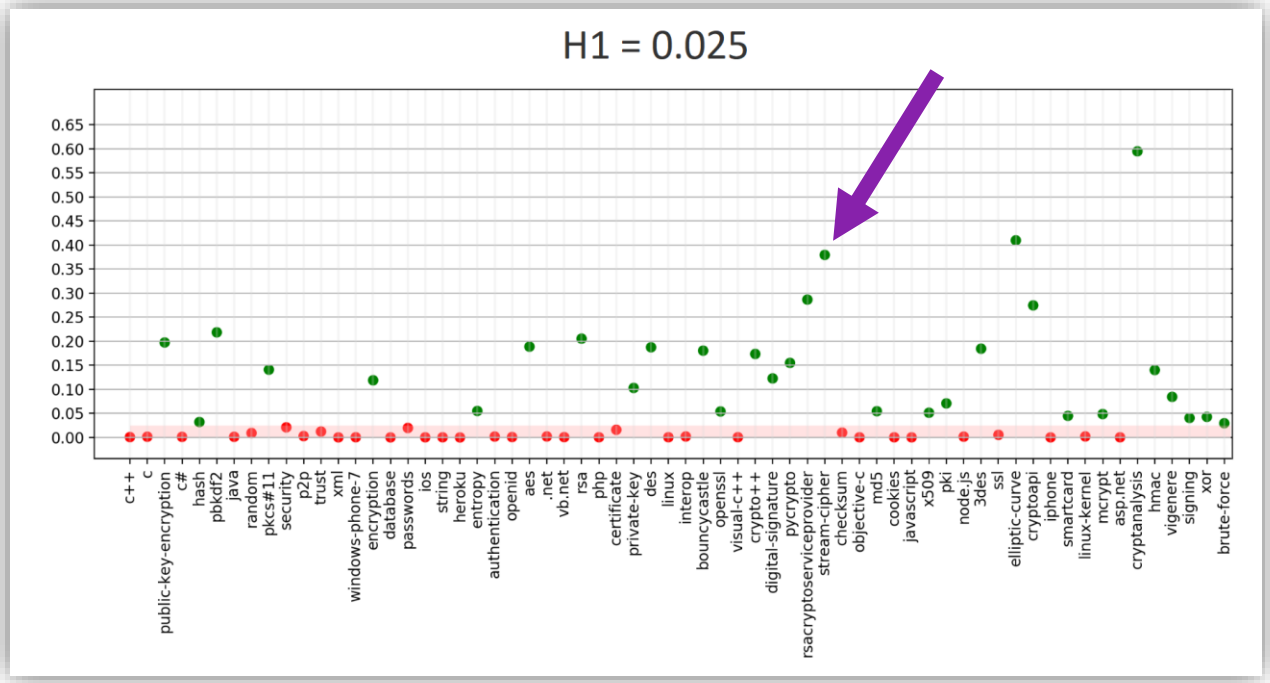
Waiting for you to make your first edit…

II. Tag analysis to get tags related to «cryptography»

- $\boldsymbol{H1}\ of\ tag$ "x" $= \frac{\#posts\ containing\ "x"\ AND\ "cryptography"}{\#posts\ containing\ "x"}$

- Example: H1 of «**stream-cipher**»
  - 11'130 posts with "cryptography"
  - 29 posts with "stream-cipher"
  - 11 posts with both

  H1(**s.c.**)= $\frac{11}{29}$ = 0.379310

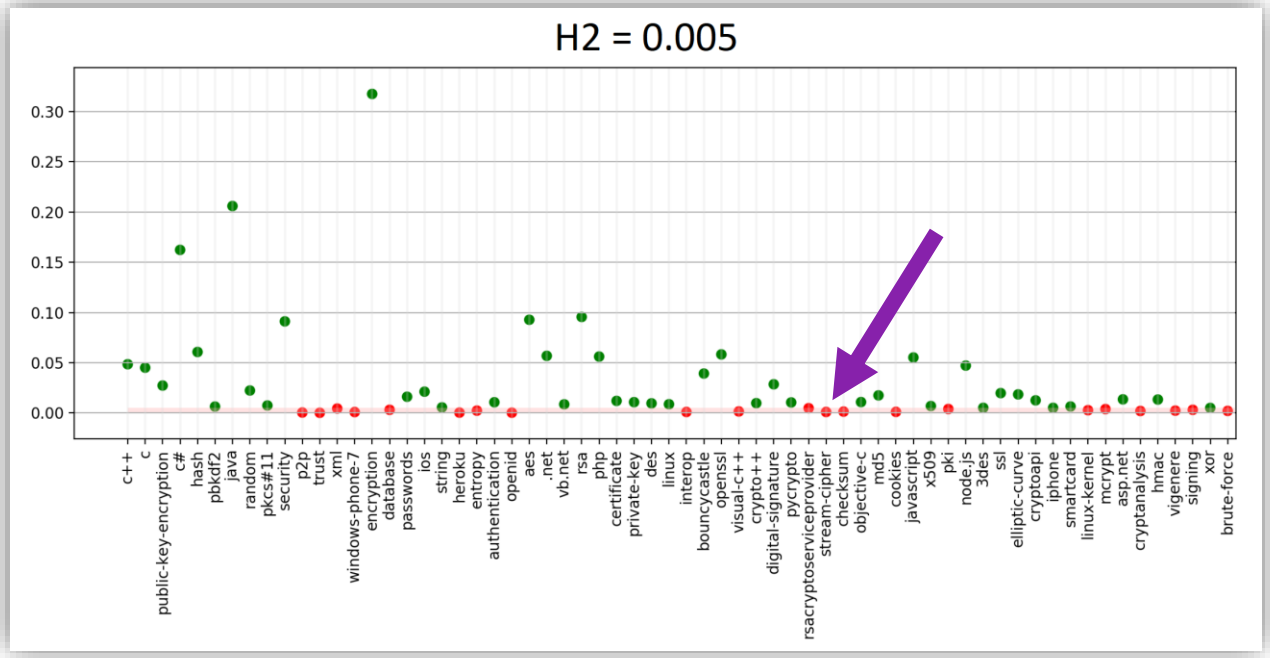


II. Tag analysis to get tags related to «cryptography»

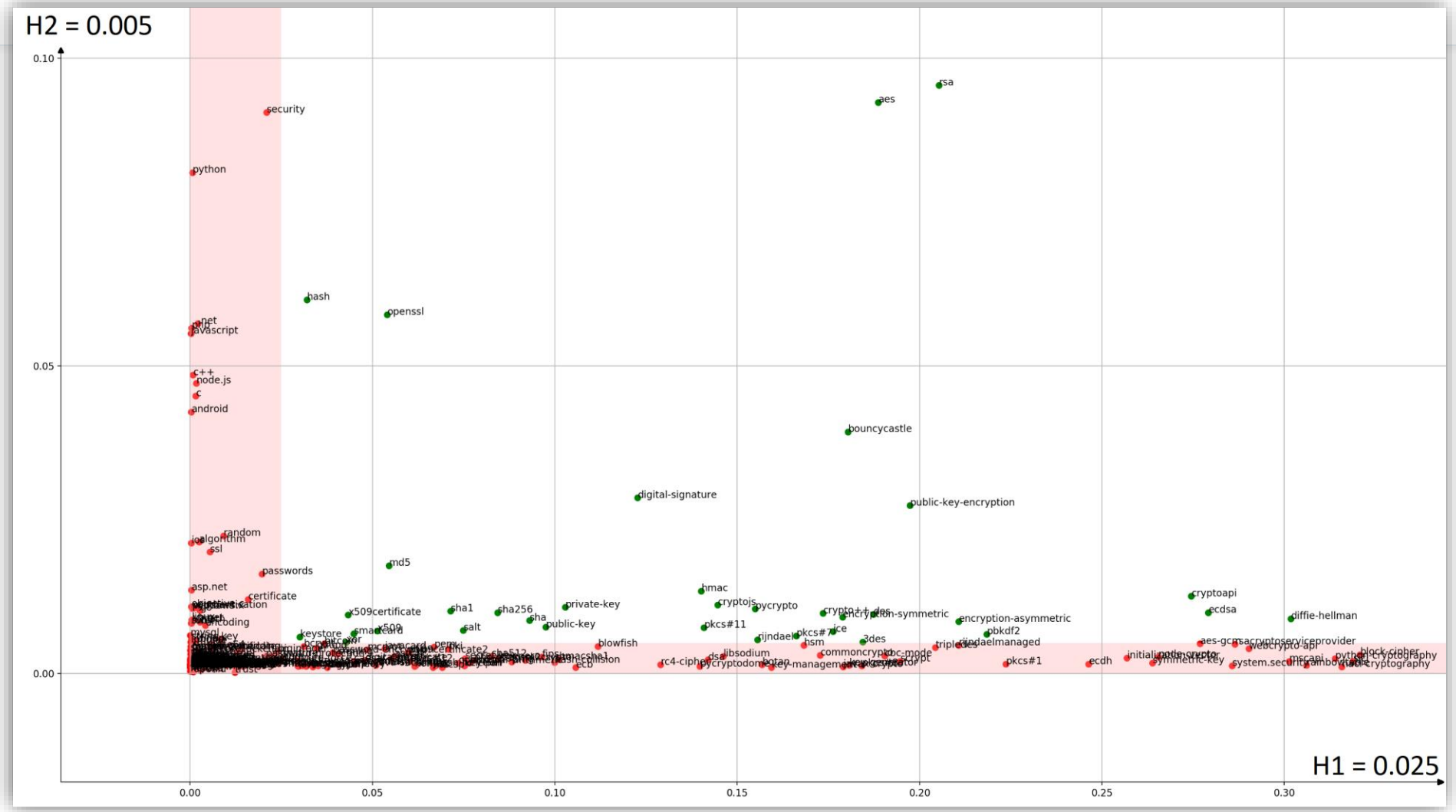
- $\boldsymbol{H2}\ of\ tag\ \text{"x"} = \frac{\#posts\ containing\ \text{"x"}\ AND\ \text{"cryptography"}}{\#posts\ containing\ \text{"cryptography"}}$

- Example: H2 of «**stream-cipher**»
  - 11'130 posts with "cryptography"
  - 29 posts with "stream-cipher"
  - 11 posts with both

  H2(**s.c.**)= $\frac{11}{11'130}$ = 0.0009883



II. Tag analysis to get tags related to «cryptography»



1. Gather data from Stack Overflow

2. Generate topics with LDA
3. Calculate metrics for RQ2 and RQ3

I. Download posts with tag «cryptography» (as CSV) from https://data.stackexchange.com/

II. Tag analysis to get tags related to «cryptography»

III. Download for each related tag a separate CSV

1. Gather data from Stack Overflow
2. Generate topics with LDA
3. Calculate metrics for RQ2 and RQ3

Demo

1. Gather data from Stack Overflow

2. Generate topics with LDA

3. Calculate metrics for RQ2 and RQ3

Optimizing LDA

- With GridSearch
  - «Number of topics»
  - «learning_decay»: Controls learning rate
  - «doc_topic_prior»: alpha
  - «topic_word_prior»: beta
- By hand with pyLDAvis

1. Gather data from StackOverflow

2. Generate topics with LDA

3. Calculate metrics for RQ2 and RQ3

Results:

Topic 1:
hash password md5 use salt valu user sha1 function store (password, hash)

Topic 2:
encrypt string decrypt code use key array byte tri (encrypt, decrypt)

Topic 3:
certif key use openssl file public privat server sign (certificate, openssl)

1. Gather data from Stack Overflow
2. Generate topics with LDA

3. Calculate metrics for RQ2 and RQ3

RQ2: Find most popular topics

Average view-count

Average number of comments

Average number of «marked as favorite»

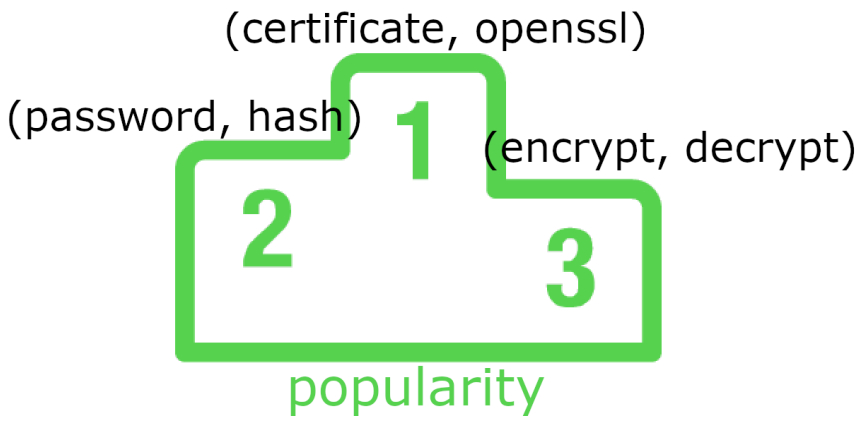
Average score of the question

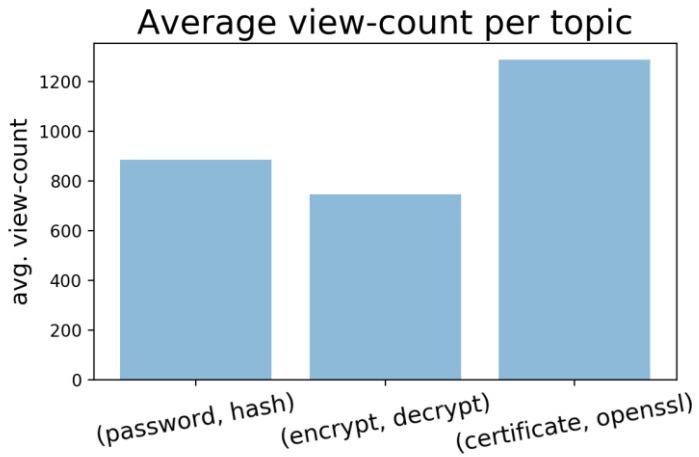
1 Gather data from Stack Overflow

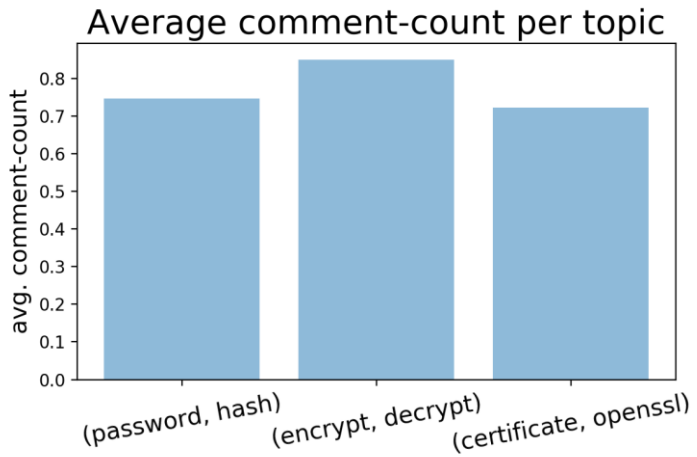
1. Gather data from Stack Overflow
2. Generate topics with LDA

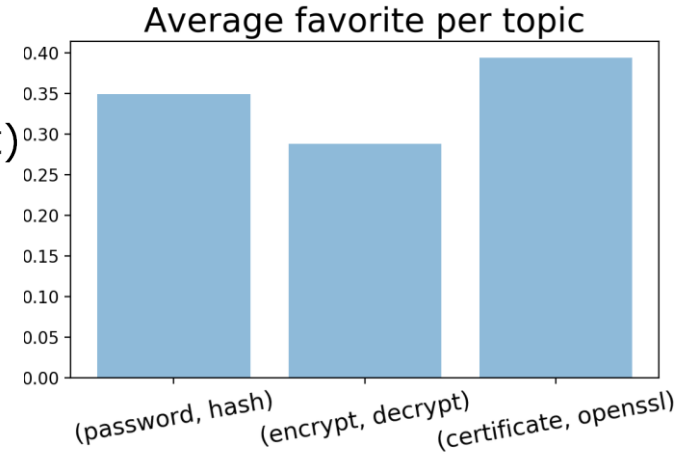
3. Calculate metrics for RQ2 and RQ3

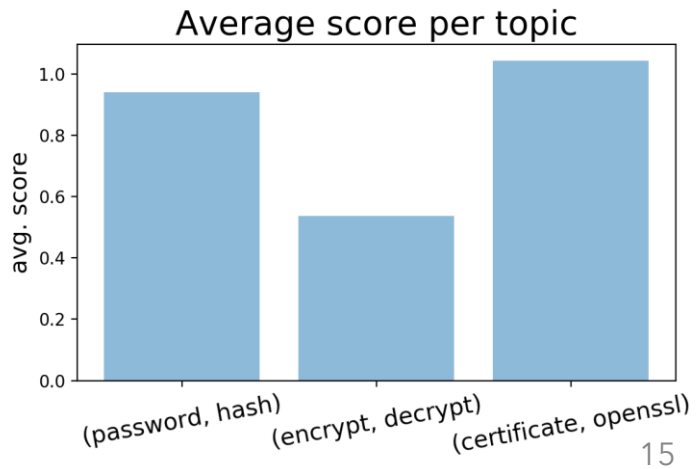
Results for RQ2:







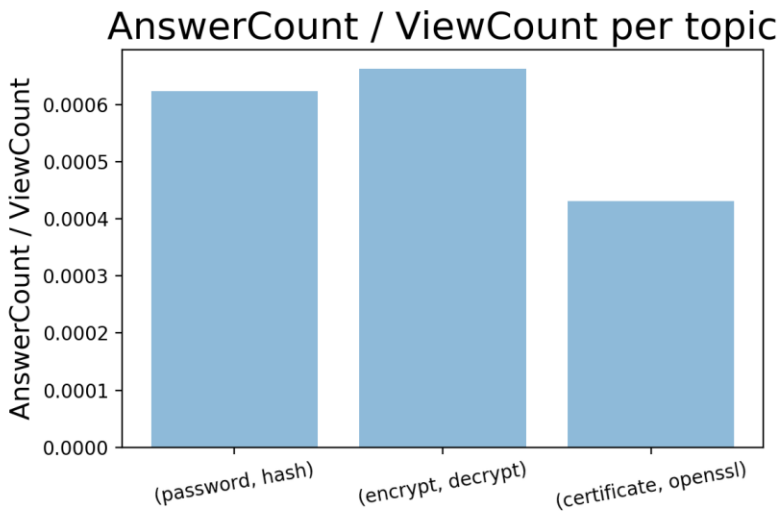




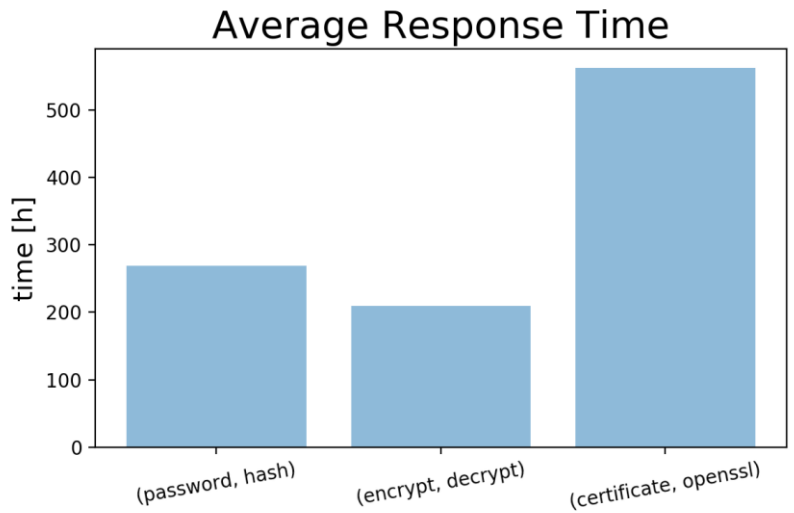
1. Gather data from Stack Overflow
2. Generate topics with LDA

3. Calculate metrics for RQ2 and RQ3

RQ3: Find most difficult topics



The **smaller** the more difficult

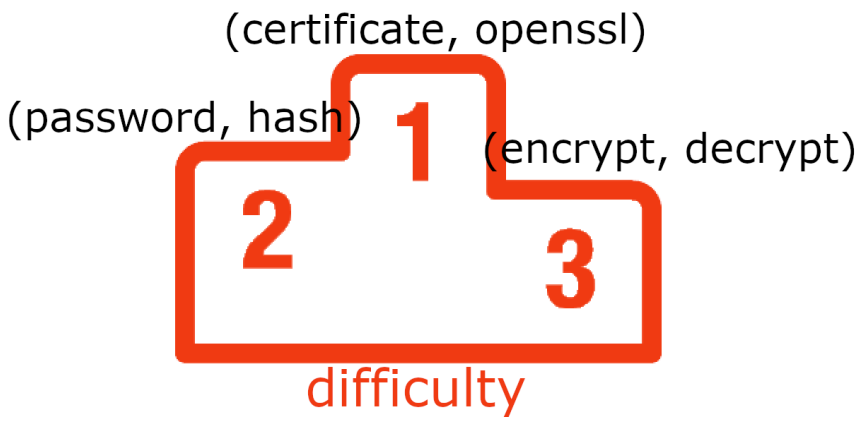


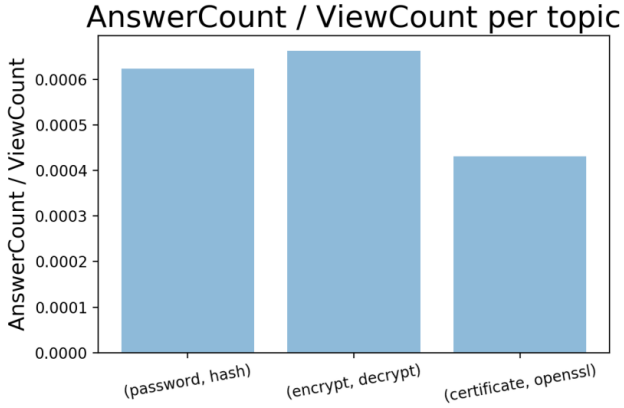
The **bigger** the more difficult

1. Gather data from Stack Overflow
2. Generate topics with LDA

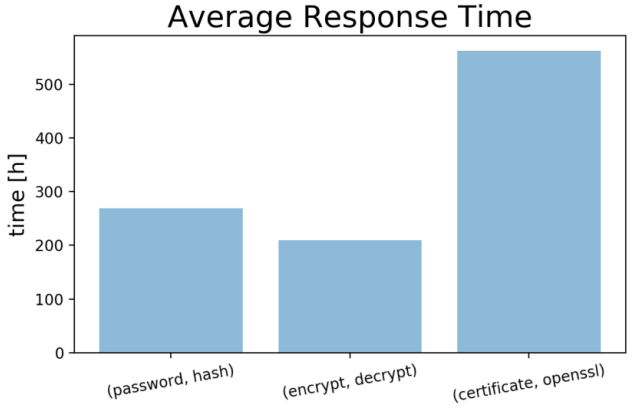
3. Calculate metrics for RQ2 and RQ3

RQ3: Find most difficult topics



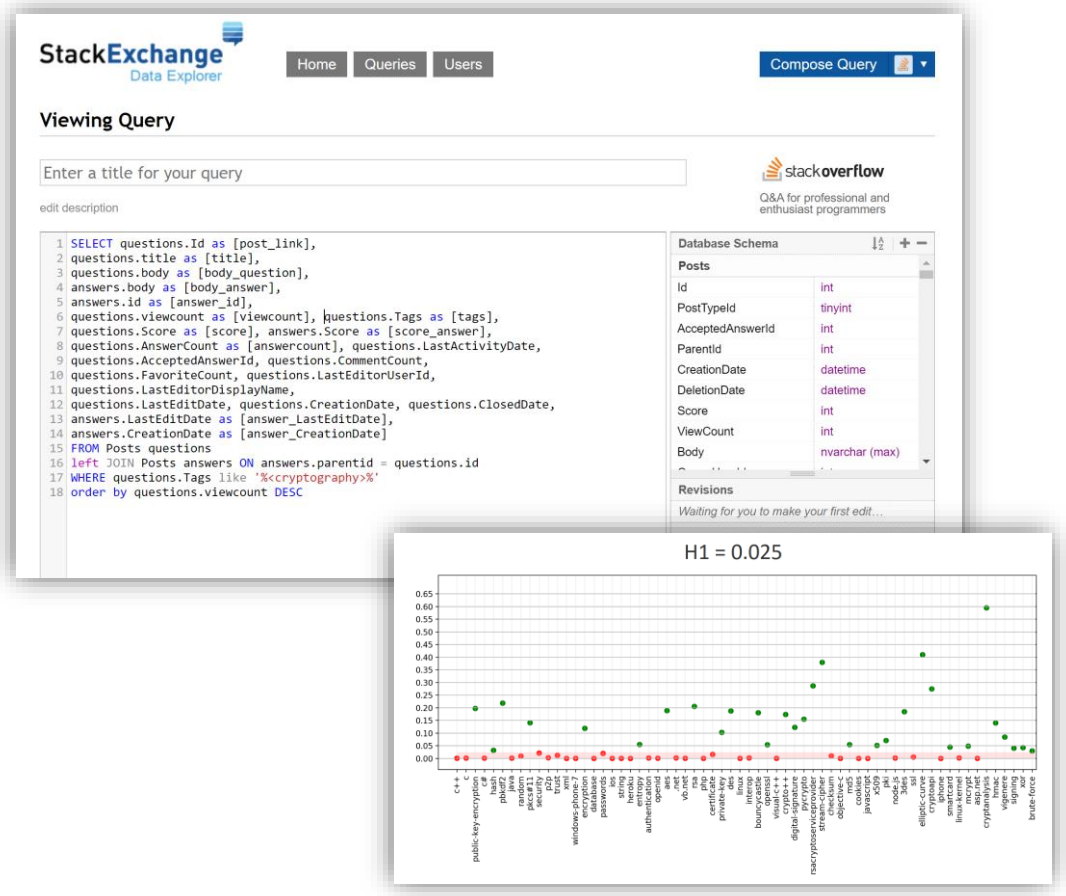


The **smaller** the more difficult

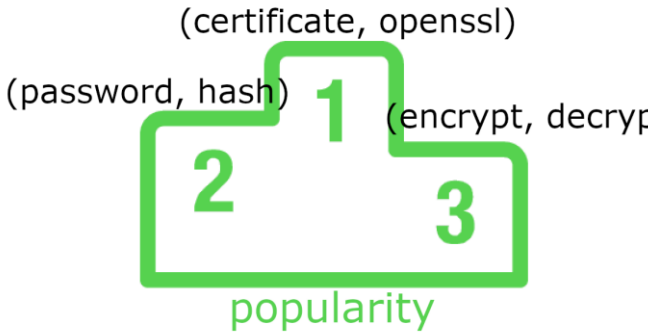
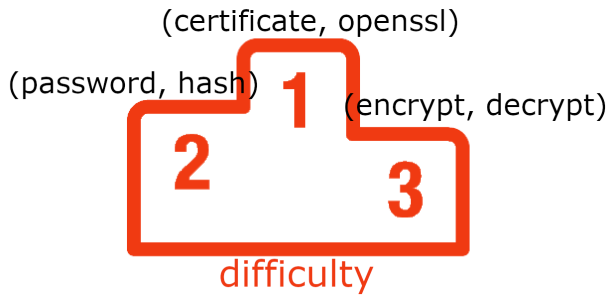


The **bigger** the more difficult

Summary



Topic 1:
hash password md5 use salt valu user sha1 function store (password, hash)

Topic 2:
encrypt string decrypt code use key array byte tri (encrypt, decrypt)

Topic 3:
certif key use openssl file public privat server sign (certificate, openssl)