Questions to ask about data / format, how do we obtain the relevant data?

**MEMBERS:**


- Eric Stevens - (310) 999 3742 - ericstevens26101@gmail.com
- Yunqi yqyan2010@gmail.com
- Jordan Witte - wittejm@gmail.com
- Jim Tyhurst jimtyhurst@gmail.com


What is the overall task?

- Given a pair of questions on stackexchange, are those questions duplicates?
- Or: Given a new question, is it a duplicate of some other pre-existing question?
    - 
- Dataset: a subset of stackexchange?
- 

Data needed:

- Duplicate/original pairs
    - How is this "duplicate" relationship represented in the database?
    - Not sure if I understand all the pieces yet, but this query returned 440,871 questions closed as duplicates (response returned in 1420 ms): https://data.stackexchange.com/stackoverflow/query/1025742/number-of-closed-duplicate-questions
    - However, this query returned 478,561 questions closed as duplicates (response returned in 2909 ms): https://data.stackexchange.com/stackoverflow/query/1025720/number-of-posts-closed-as-duplicates
      I cannot yet explain the discrepancy between the two.
    - See some of the SQL queries referenced here, which I used to get started: https://meta.stackexchange.com/questions/322241/how-to-find-duplicates-that-could-be-deleted
    - I experimented with variations of this query: Unmerged answered self-duplicates (list of stub Qs), which returned 8626 rows returned in 4395 ms. I clicked on a few of the links in the results. Each of those questions had a highlight at the top that said, "This question already has an answer here:" or "Possible Duplicate:" with a link to another question. I guess I never noticed this highlighted block before when searching in Stack Overflow.
    - Stack Overflow documentation (https://stackoverflow.com/help/privileges/flag-posts) says flag types include: 'Flag to close' > 'Duplicate question'. But really, this information is represented in the 'CloseReasonTypes' table, which has 'exact duplicate' and 'duplicate' items.
    - Stack Overflow documentation (https://stackoverflow.com/help/privileges/close-questions) says, "Questions with

vote to close as a duplicate will also list each possible duplicate in a tab at the top for easy reviewing."
    - Stack Overflow documentation, "Why are some questions marked as duplicate?" (https://stackoverflow.com/help/duplicates). "Moderators and anyone with 3000 reputation may vote to close a question as a duplicate by clicking the 'close' link and entering in the URL or title of the question they believe it to be a duplicate of."
- Non-trivial not-duplicate pairs (how do we obtain these?)
    - Ideally, we can find pairs that are "similar", but not duplicate to use for training a model.
    - Possible solution #1:
        i. Limit questions to some subset, e.g. only questions related to Python.
        ii. Create document term matrix for each question.
        iii. Find clusters of documents using the vectors from (ii):
            1. Using k-means unsupervised learning?
            2. k might need to be very large for the number of clusters in Stack Overflow even if limited to a subset in (i).
        iv. Within a cluster, find pairs of questions <x, y> where x is not a duplicate of y. Those pairs <x, y> are "similar", because they are in the same cluster, but they are not duplicates. We want to include those similar pairs in our training data, so that the prediction engine (whatever machine learning method is used to identify two documents as duplicates) will be tuned to identify duplicates without also mislabeling "similar" documents as duplicates.
    - Possible Solution #2
      Wei et al (2018) "Duplicate Detection in Programming Question Answering Communities" give four methods for finding candidate questions that might be duplicates. Maybe some of those candidates could be used as "similar", but not duplicates. See Sections 3.1 - 3.4.
        i. Candidate Selection with Tags
        ii. Candidate Selection with Query Likelihood Model
        iii. Candidate Selection with BM25
        iv. Candidate Selection with Topics

Data exploration!

- What do questions marked as duplicates actually look like?
  These fields seem relevant:
    - Posts.PostTypeId = 1  -- question
    - PostHistory.PostHistoryTypeId = 10  -- Post Closed
    - PostLinks.LinkTypeId = 3  -- 'Duplicate' per documentation, but there is no PostLinksTypes table, so no corresponding Name or Description fields in the database.
- Are there simple indicators like the user asking, question tags?

- How many duplicate posts are there? Over time?
- Who asks duplicate questions? New users? Veteran users?
- Are duplicate questions shorter?
- How old are duplicate questions? How long do new questions take to get marked as duplicate?

Specific data questions (with SQL query if possible) :

- Number of posts on stack exchange: 44313825

```
select count(*) from posts;
```

- Number that have been closed: 784836

```
select count(*) from posts where closeddate is not null;
```

- Number closed because duplicate:

Post similarity

- How to determine a "similarity measure" between two posts?
  - Given PostA and PostB, how similar are they?
    Response is a probability in [0, 1].
  - Python libraries that calculate similarity:
    - spaCy
      https://spacy.io/usage/vectors-similarity
    - gensim
      https://github.com/RaRe-Technologies/gensim/blob/develop/docs/notebooks/Similarity_Queries.ipynb
    - scikit-learn
      - https://markhneedham.com/blog/2016/07/27/scitkit-learn-tfidf-and-cosine-similarity-for-computer-science-papers/
      - https://www2.cs.duke.edu/courses/spring14/compsci290/assignments/lab02.html
      - https://blogs.oracle.com/meena/finding-similarity-between-text-documents
  - R packages that calculate similarity:
    - tidytext
      - https://www.tidytextmining.com/tfidf.html
      - https://www.tidytextmining.com/usenet.html#finding-tf-idf-within-newsgroups
    - text2vec
      http://text2vec.org/similarity.html
    - tm
      - https://rstudio-pubs-static.s3.amazonaws.com/266040_d2920f956b9d4bd296e6464a5ccc92a1.html
      - http://fredgibbs.net/tutorials/document-similarity-with-r.html

Data available:

- Schema: https://sedeschema.github.io/
- SQL access to online database: https://data.stackexchange.com/stackoverflow/query/new
- Examples of queries: https://data.stackexchange.com/stackoverflow/queries
- Tutorial for writing queries: https://data.stackexchange.com/tutorial/
- John Burt's notebook, demonstrating how to use the Python library 'StackAPI' to access Stack Exchange data.
- Google BigQuery has "Stack Overflow posts" available publicly for query, but:
    - I do not see all of the tables listed. For example, I do not see the *Types tables (VoteTypes, CloseReasonTypes, ...).
    - Schema is different. For example, rather than one table of 'Posts' with a 'PostTypeId' field, there are multiple tables: 'posts_questions', 'posts_answers', 'posts_privilege_wiki', ...
    - The field names are different, e.g. 'creation_date' vs 'CreationDate'.
    - You can run queries at: https://bigquery.cloud.google.com/dataset/bigquery-public-data:stackoverflow
    - 31,017,889 rows in BigQuery stackoverflow_posts table, compared to 44,313,825 rows in Stack Exchange's 'Posts' table (queried on 2019-04-07).
    - SELECT
      MIN(creation_date) AS creation_date,
      MAX(last_activity_date) AS last_activity_date
      FROM
      [bigquery-public-data:stackoverflow.stackoverflow_posts]
      returns:
        creation_date: 2008-07-31 21:42:52.667 UTC
        last_activity_date: 2016-06-12 04:07:42.853 UTC
      So the data is not current. A similar query at
      https://data.stackexchange.com/stackoverflow/query/new
      on the Posts table returns:
        CreationDate: 2008-07-31 21:42:52
        LastActivityDate: 2019-04-07 05:05:22
    - The following query returned 452,253, but I am not really sure if the query is correct:
      -- Counts the number of Posts closed as a duplicate.
      SELECT COUNT(DISTINCT d.id) AS n_closed_duplicate_questions
      FROM [bigquery-public-data:stackoverflow.posts_questions] d  -- d=duplicate
      LEFT JOIN [bigquery-public-data:stackoverflow.post_history] ph ON ph.post_id = d.id
      LEFT JOIN [bigquery-public-data:stackoverflow.post_links] pl ON pl.post_id = d.id
      LEFT JOIN [bigquery-public-data:stackoverflow.posts_questions] o ON o.id = pl.related_post_id  -- o=original
      WHERE

```
d.post_type_id = 1  -- 1=Question
AND pl.link_type_id = 3  -- 3=duplicate
AND ph.post_history_type_id = 10  -- 10=Post Closed
```

- Zip files: https://archive.org/download/stackexchange/
  There are 7 files of data in XML format for StackOverflow:
  - stackoverflow.com-Badges.7z
  - stackoverflow.com-Comments.7z
  - stackoverflow.com-PostHistory.7z
  - stackoverflow.com-PostLinks.7z
  - stackoverflow.com-Posts.7z
  - stackoverflow.com-Tags.7z
  - stackoverflow.com-Users.7z
  - stackoverflow.com-Votes.7z

  Can we load these XML files into a local database easily? Or a cloud db? Is it even a good idea to try that? Maybe better to just access the Stack Exchange database through SQL queries in a web browser, even if we are (severely?) limited by how much data can be returned per query.

Approach:

- Machine Learning Algorithm:
  -
- Use question tags, topic identification (e.g. which programming language)

Tasks:

- Replicate database locally (and import subset of relevant

Related work:

- Per Runeson, et al. 2007. Detection of Duplicate Defect Reports Using Natural Language Processing. 29th International Conference on Software Engineering (ICSE'07).
  [$14.95 purchase for IEEE member; $33 for non-member]
- Quora Example:
  https://www.kaggle.com/c/quora-question-pairs/data
- Rodrigo F. G. Silva, et al. 2018. Duplicate Question Detection in Stack Overflow: A Reproducibility Study. https://peerj.com/preprints/26555v1.pdf
  Discusses two implementations in Java for detecting duplicates:
  - https://github.com/muldon/dupPredictorRep
  - https://github.com/muldon/dupeRep
- Muhammad Ahasanuzzaman, Muhammad Asaduzzaman, Chanchal K. Roy, Kevin A. Schneider: Mining duplicate questions in stack overflow, in Proceedings of the 13th International Conference on Mining Software Repositories (MSR 2016), pp. 402-412, Austin, TX, USA, May 2016.
  https://dl.acm.org/citation.cfm?id=2901770 [$10 to purchase]

[2019-04-07 Jim is an ACM member, but does not have a subscription to the digital library. He sent an email request to one of the authors <chanchal.roy@usask.ca>, asking for a free copy of the paper.]

- Wei Emma Zhang, Quan Z. Sheng, Jey Han Lau, Ermyas Abebe, and Wenjie Ruan. 2018. Duplicate Detection in Programming Question Answering Communities. ACM Trans. Internet Technol. 18, 3, Article 37 (April 2018), 21 pages. https://mafiadoc.com/37-duplicate-detection-in-programming-question-answering_5c1786ac097c47310a8b45af.html
-