A Twitter Search engine

Surya Teja Chavali

Houqi Li

Main Idea

- Transform unstructured data to structured data
- Use Spark to build a database
- Allow queries with operators
- Rank the results

Example

- Keyword Search
    - action:drink object:cola
    - Compiled in SQL

```
SELECT * FROM TWEETS
WHERE
ACTION LIKE ‘drink’ AND OBJECT LIKE ‘cola’
```

The Data

- ‘Spritzer’ version of Twitter Data from Internet Archive, dated 2018-10-01
- Random sample of 1% of tweets for that day
- 52 GB
- 417633 users

Data Cleaning and Tuple Extraction

- TweetCleaner
  - Parse tweet JSON into form ingestible by Spark
- NLP using SpaCy
  - Sentence Extraction
  - Triplet Extraction[1] using SpaCy's 'en' module into

    `(tweet_id, subject, predicate, object)`
  - Lemmatization of verbs

[1] Rusu et. al., 'Triplet extraction from sentences'

Search

- Query of the form `(subject, verb, object)`
- Some can be missing.
- Run the query

```
SELECT (*)
FROM TWEETS
WHERE SUBJECT LIKE ‘subject’
AND OBJECT LIKE ‘object’
AND VERB LIKE ‘verb’
```

Ranking Measures

- FollowerRank

$$\mathrm{FR}\left(a\right) = \frac{i(a)}{i(a)+o(a)}$$

- LengthRank

$$f_{LR}(t,q) = \frac{l(t)}{\max\limits_{s \in \mathcal{T}_q^k} l(s)}$$

- URLRank

$$f_{UR}(t,R) = \begin{cases} c & t \text{ contains a URL} \\ 0 & \text{else} \end{cases}$$

From Nagmoti et. al., ‘Ranking Approaches for Microblog Search’

Combining the Ranking Measures

$$f_{FLR}(t,q) = f_{FR}(t,q) + f_{LR}(t,q)$$

$$f_{FLUR}(t,q) = f_{FLR}(t,q) + f_{UR}(t,q)$$

From Nagmoti et. al., ‘Ranking Approaches for Microblog Search’

Test Set

- 21 queries - 3 each of 7 classes
  - subject only, object only, verb only present
  - Two out of three present
  - All three present
- The queries are the 3 most frequent ones in each category.

Evaluation Metric

- Our metric of evaluation is **precision@5**
- **Precision@k** corresponds to the number of relevant results present in the top k search results
- See if our top 5 results feature in top 500 results of Twitter Advanced Search
    - This is ‘fair’ because we have only 1% of the data

Results

Average p@5	0.05
Max p@5	0.4
Min p@5	0.0

Next Steps

- Run on larger dataset - see if ranking is better.
- Better queries to evaluate on
    - Queries near the median or 80th percentile of the data rather than top few

References

- Internet Archive. Internet archive search: collection:twitterstream.
- Rusu, D., Dali, L., Fortuna, B., Grobelnik, M., & Mladenic, D. (2007, October). Triplet extraction from sentences. In Proceedings of the 10th International Multiconference” Information Society-IS (pp. 8-12).
- Nagmoti, R., Teredesai, A., & De Cock, M. (2010, August). Ranking approaches for microblog search. In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01 (pp. 153-157). IEEE Computer Society.
- Jeong, J. W., Morris, M. R., Teevan, J., & Liebling, D. (2013, June). A crowd-powered socially embedded search engine. In Seventh International AAAI Conference on Weblogs and Social Media.