TweetsKB: A Public and Large-Scale RDF Corpus of Annotated Tweets

Pavlos Fafalios[1] **Vasileios Iosifidis**[1] Eirini Ntoutsi[1]
Stefan Dietze[1]

[1]L3S Research Center
Leibniz Universität Hannover
Hannover, Germany

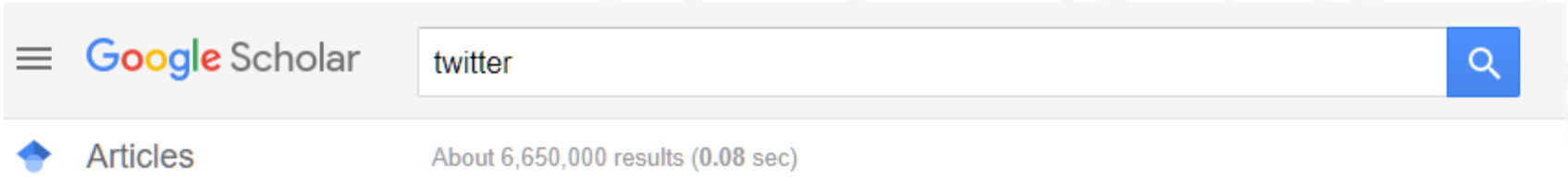
June 2018

Micro-blogging for everyday life

- Micro-blogging services are extremely popular these days.
- Combination of blogging and instant messaging.
- Huge amount of data.

Twitter

- The most well known micro-blogging platform.
- Twitter's generated data has been used in a wide area of research:
  - data science, sociology, psychology and so on and so forth.



Twitter

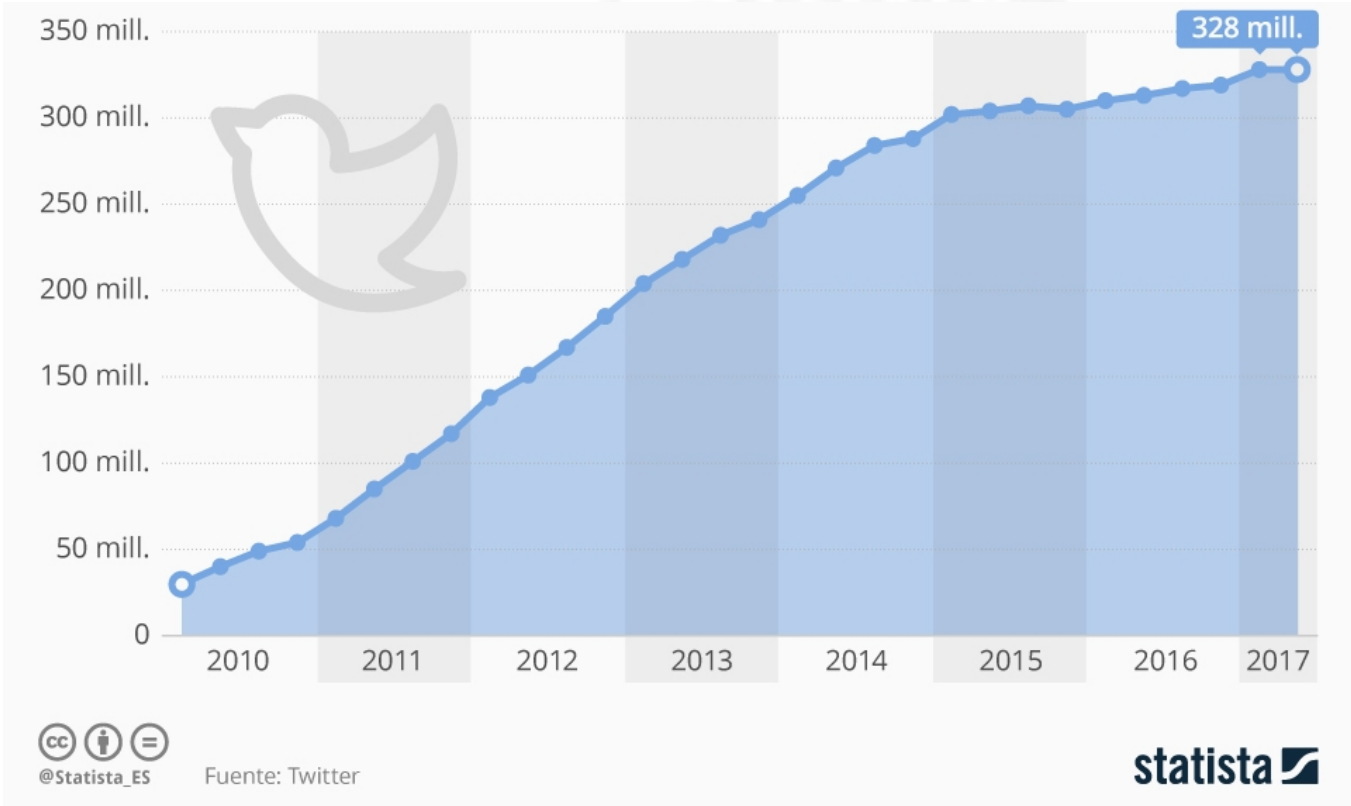
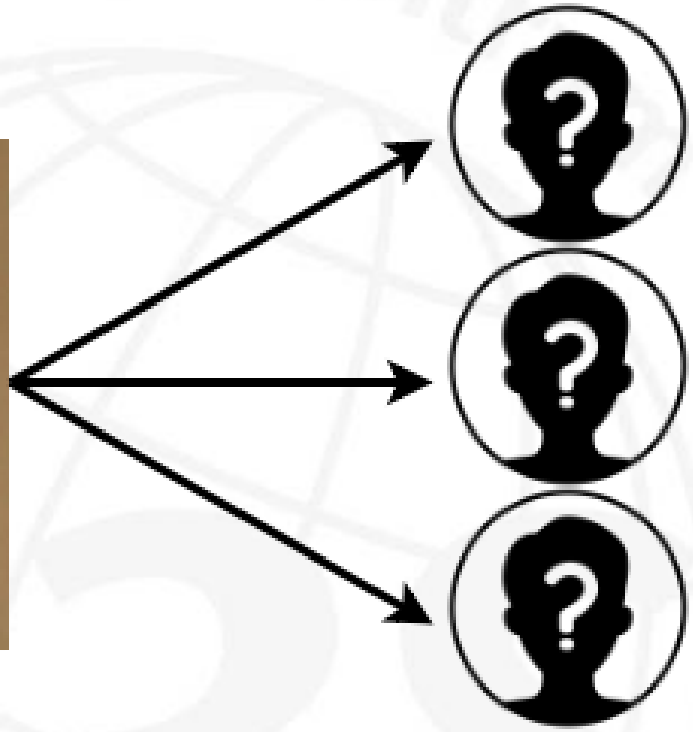


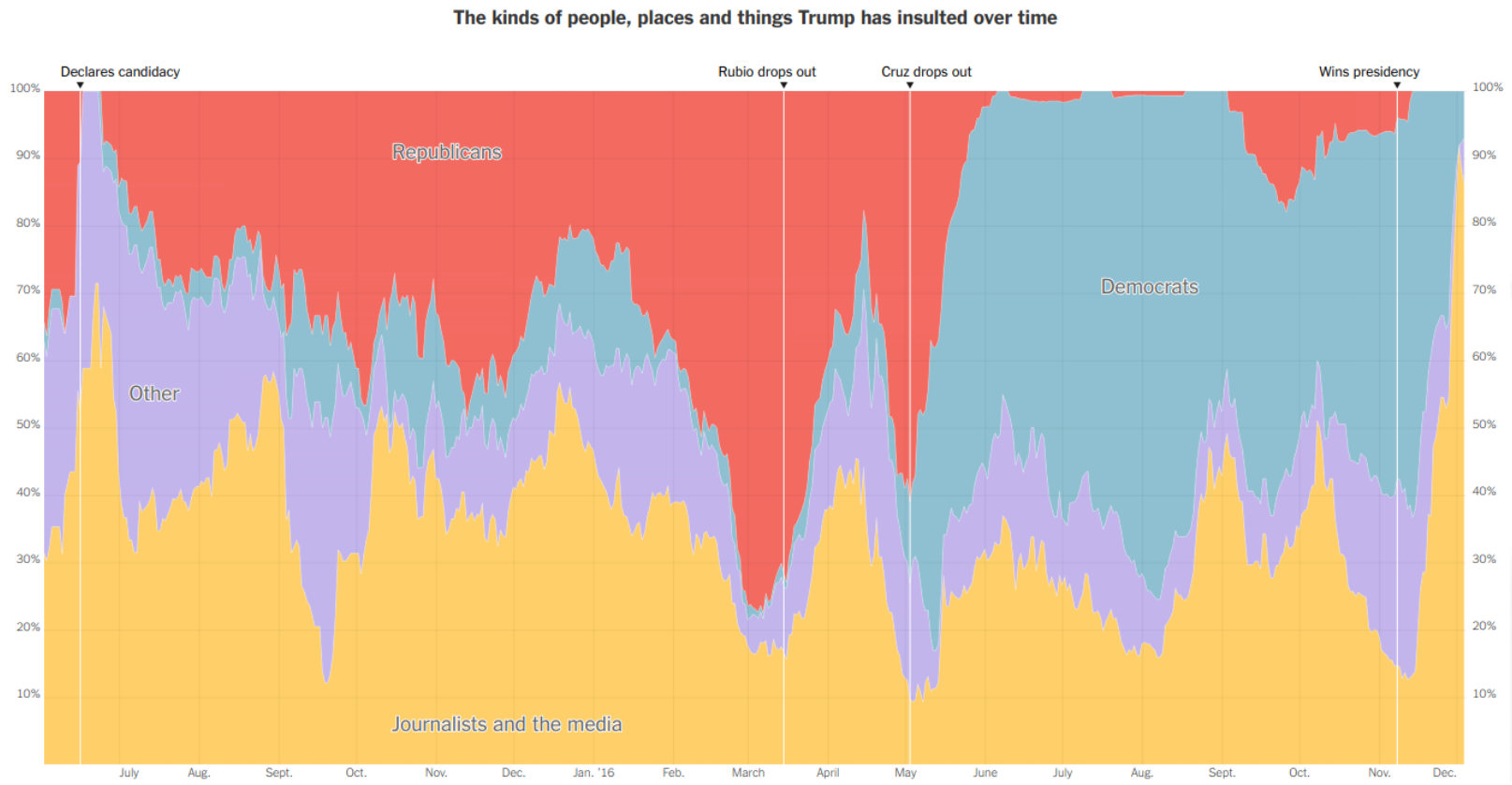
Figure: Number of tweets per month

Entity Relatedness ?



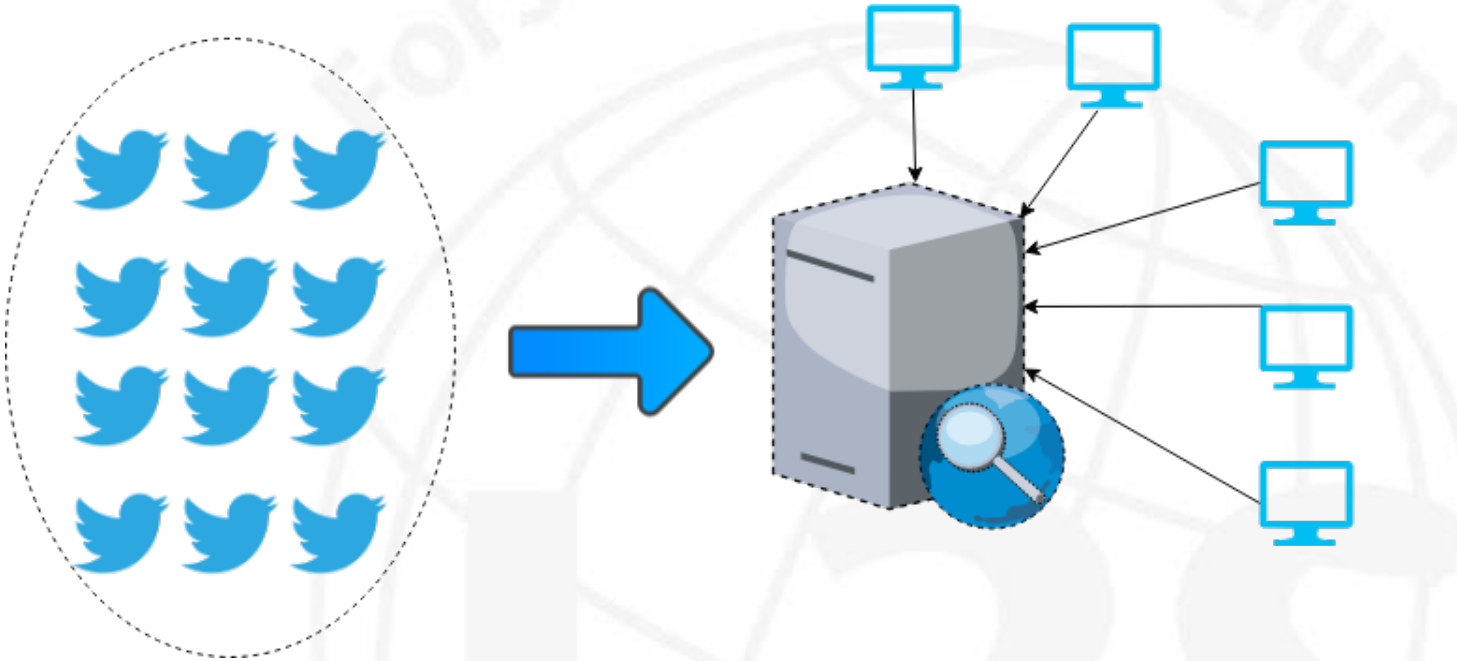
Opinions Overtime ?

Journalists manually analysed 14,000 tweets[1] !



[1] https://www.nytimes.com/interactive/2016/12/06/upshot/how-to-know-what-donald-trump-really-cares-about-look-at-who-hes-insulting.html

Lack of large scale public social archives



TweetsKB Overview

- Public RDF corpus of anonymized tweets.
- More than 1.5 billion English tweets.
- Spanning period: Jan-2013 till Nov-2017.
- Includes entities and sentiment annotations.
- Ideal for:
  - time-aware and entity-centric exploration.
  - data integration by existing knowledge base (like DBpedia).
  - multi-aspect and entity-centric analysis.

Generating TweetsKB: Filtering

- 6 billion tweets (Jan-13 - Nov-17).
- Retweets, non-English tweets and spam[2] tweets were removed.
- We employ some of tweets' metadata.

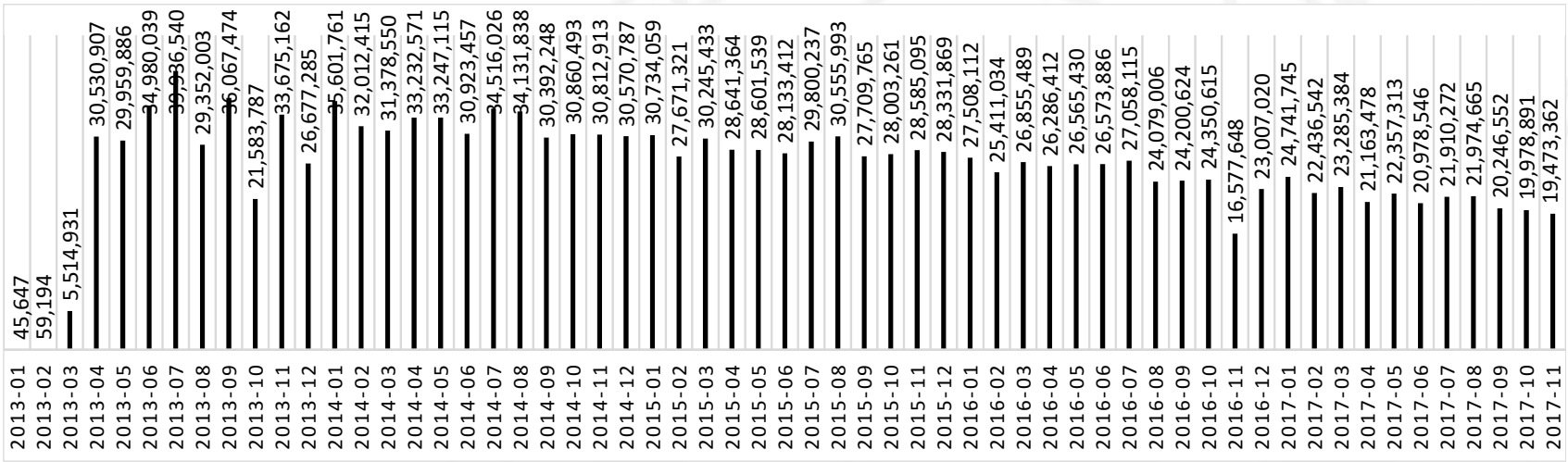


Figure: TweetsKB: Number of tweets per month

[2]for spam detection HSpam14 dataset was used [Sedhai and Sun, 2015]

Entity Linking

- Yahoo's FEL tool [Blanco et al., 2015].
- FEL has been specifically designed for short texts.
- 1.4 million distinct entities.
- Evaluated on NEEL challenge dataset [Rizzo et al., 2016]:
  - Dataset consists of 9,289 English tweets from 2011, 2013, 2014 and 2015.
  - Precisions = 85% , Recall = 39% and F1 = 54%

Sentiment Annotation

- For sentiment analysis we have used SentiStrength [Thelwall et al., 2012].
- SentiStrength assigns both positive and negative scores to short texts [5, -5].
- Evaluated on TSentiment15 dataset [Iosifidis and Ntoutsi, 2017]:
  - Dataset consists of 2.5 million English tweets from 2015.
  - Accuracy = 91% , F1 = 80%

TweetKB's overall statistics

- TweetsKB is available as N3 files through Zenodo repo (DOI: 10.5281/zenodo.573852)[3]
- It's also registered at datahub.ckan.io[4]

Number of tweets	1,560,096,518
Number of distinct users	125,104,569
Number of distinct hashtags	40,815,854
Number of distinct user mentions	81,238,852
Number of distinct entities	1,428,236
Number of tweets with sentiment	772,044,599
Number of RDF triples	48,207,277,042

[3] https://zenodo.org/record/573852
[4] https://datahub.ckan.io/dataset/tweetskb

Computational Cost

- Cluster: CPUs 504, RAM 6,784 GB, Disk 2.1 PB
- Sentiment annotation: 6M tweets per minute
- Entity Extraction: 4.8M tweets per minute
- Triplification: 14M tweets per minute

RDF Schema Our schema follows established vocabularies such as SIOC, schema.org, and Onyx

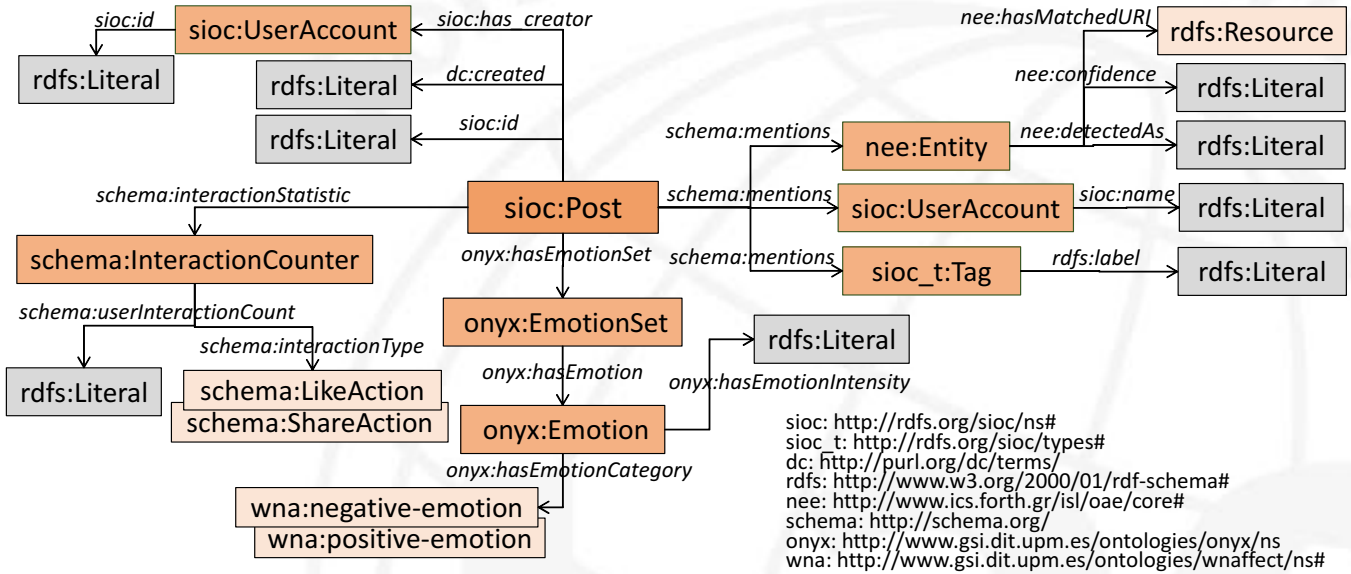


Figure: Employed RDF/S model

Tweets in our RDF/S model are associated with six types of elements:

- metadata: $sioc : Post \rightarrow tweet$, $sioc : UserAccount \rightarrow user$
- entity mentions: for each entity, we store surface form, URI, confidence score (NEE).
- user mentions: $sioc : UserAccount \rightarrow user$
- hashtag mentions: $sioc\_t : Tag$
- sentiment scores: $onyx : EmotionSet$
- interaction statistics:
  $schema : LikeAction \rightarrow favorite\ count$,
  $schema : ShareAction \rightarrow retweet\ count$,

Example

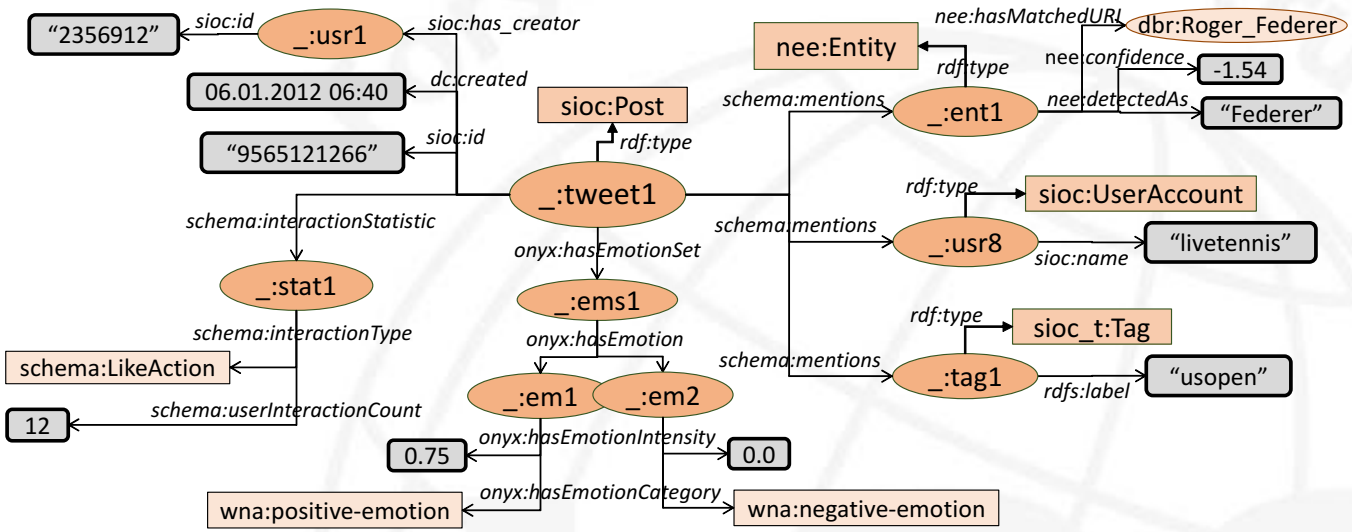


Figure: Tweet example

Analysis of the top 100K entities and hashtags

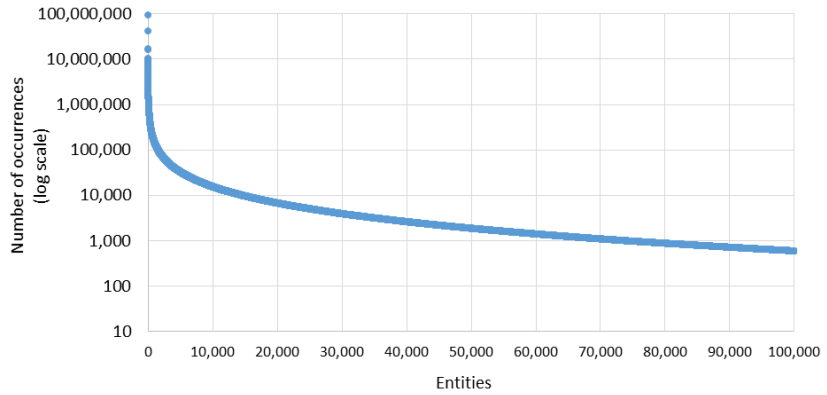


Figure: Distribution of top-100,000 entities.

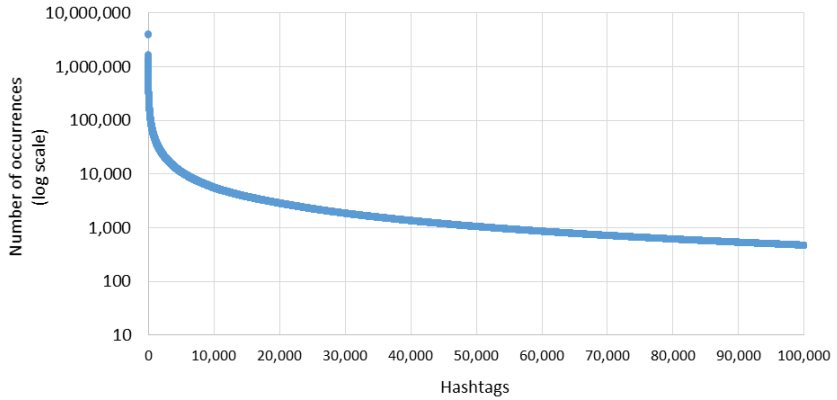


Figure: Distribution of top-100,000 hashtags.

Analysis of the top 100K entities

DBpedia type	Number of distinct entities
http://dbpedia.org/ontology/Person	21,139 (21.1%)
http://dbpedia.org/ontology/Organisation	14,815 (14.8%)
http://dbpedia.org/ontology/Location	8,215 (8,2%)
http://dbpedia.org/ontology/Athlete	5,192 (5.2%)
http://dbpedia.org/ontology/Artist	3,737 (3.7%)
http://dbpedia.org/ontology/City	2,563 (2.6%)
http://dbpedia.org/ontology/Event	510 (0.5%)
http://dbpedia.org/ontology/Politician	208 (0.2%)

Table: Overview of popular entity types of the top-100,000 entities.

Data Exploration and Integration

SPARQL queries can be used to integrate information from external knowledge bases such as DBpedia.

```
SELECT DISTINCT ?tweetID ?sentNegScore ?retweetCount ?politician ?birthPlace WHERE {
  SERVICE <http://dbpedia.org/sparql> {
    ?politician dc:subject dbc:German_politicians ; dbo:birthPlace ?birthPlace }
  ?tweet a sioc:Post ; dc:created ?date ; sioc:id ?tweetID FILTER(year(?date) = 2016) .
  ?tweet schema:mentions ?entity . ?entity a nee:Entity ; nee:hasMatchedURI ?politician .
  ?tweet schema:interactionStatistic ?stat . ?stat schema:interactionType schema:ShareAction .
  ?stat schema:userInteractionCount ?retweetCount FILTER(?retweetCount > 100) .
  ?tweet onyx:hasEmotionSet ?emotSet . ?emotSet onyx:hasEmotion ?emot .
  ?emot onyx:hasEmotionCategory wna:negative-emotion ;
        onyx:hasEmotionIntensity ?sentNegScore FILTER (?sentNegScore >= 0.75) }
```

Figure: SPARQL query for retrieving popular tweets in 2016 mentioning German politicians with strong negative sentiment.

Temporal Entity Analytics

Given an entity and a time period one can make temporal queries such as:

```
SELECT ?month xsd:double(?cEnt)/xsd:double(?cAll)
WHERE {
{ SELECT (month(?date) AS ?month) (count(?tweet) AS ?cAll) WHERE {
   ?tweet a sioc:Post ; dc:created ?date FILTER(year(?date) = 2015)
  } GROUP BY month(?date) }
{ SELECT (month(?date) AS ?month) (count(?tweet) AS ?cEnt) WHERE {
   ?tweet a sioc:Post ; dc:created ?date FILTER(year(?date) = 2015) .
   ?tweet schema:mentions ?entity .
   ?entity a nee:Entity ; nee:hasMatchedURI dbr:Alexis_Tsipras
  } GROUP BY month(?date) }
} ORDER BY ?month
```

Figure: SPARQL query for retrieving the monthly popularity of *Alexis Tsipras* (Greek prime minister) in 2015

Entity Recommendations

Entity co-occurancies indicate entity relatedness.

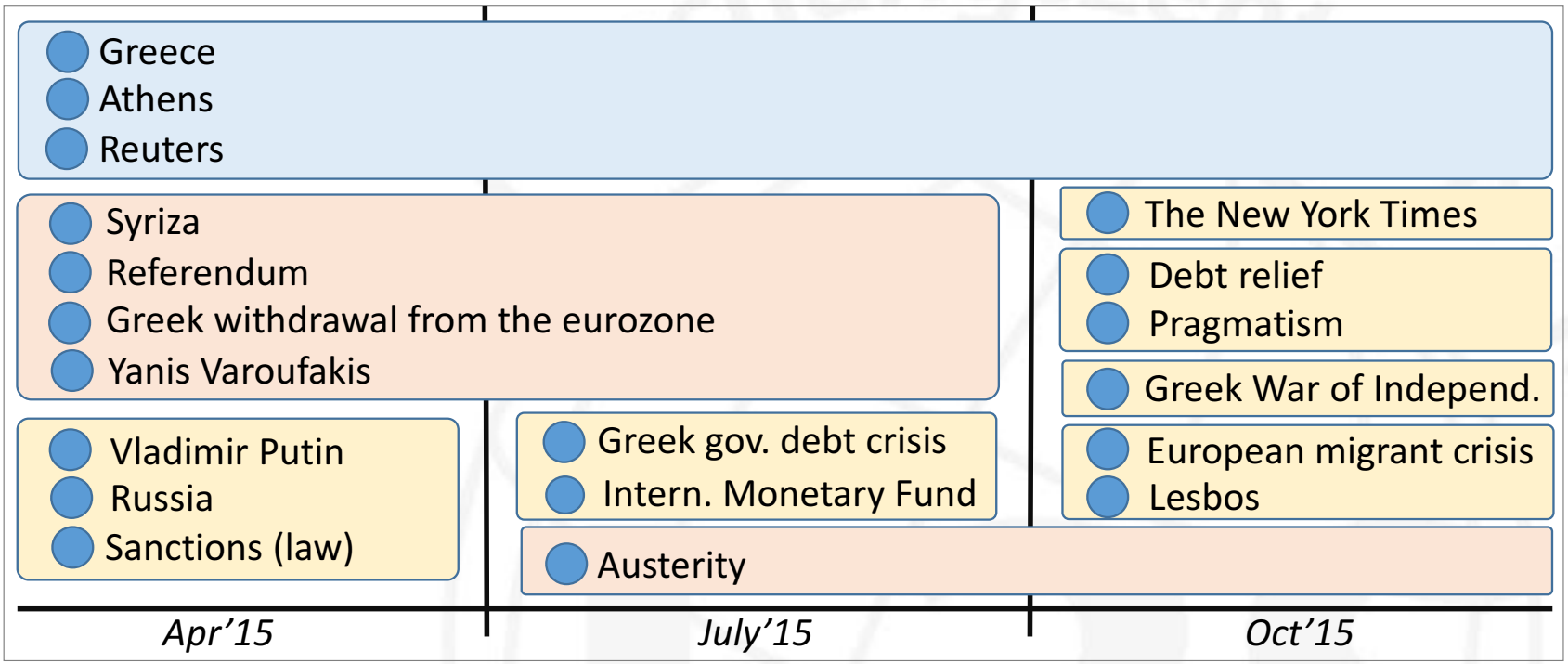


Figure: Co-occuring entities of Tsipras (Greek prime minister) during Apr'15 - Dec'15

Sustainability and Maintenance

- Archive is deposed on Zenodo and registered at datahub.ckan.io
- Integrated into AFEL[5] and ALEXANDRIA[6] projects.
- Endpoint[7] which currently facilitates 5% of the whole corpus.
- Periodically update TweetsKB (every 6 months) with recent information.

[5] http://afel-project.eu/
[6] http://www.alexandria-project.eu
[7] http://l3s.de/tweetsKB/endpoint/

Conclusions

- Largest annotated RDF corpus of tweets.
- Advanced entity-centric exploration.
- Data integration with other knowledge bases (at query execution time).
- Analytics for different disciplines and research problems.

Thanks

Questions?

Contact: {fafalios, iosifidis}@L3S.de

AFEL