chin_2017_totem_personal_tweets_summarization_on_mobile_devices

Year

2017

Author(s)

Chin, Jin Yao and Bhowmick, Sourav S. and Jatowt, Adam

Title

TOTEM: Personal Tweets Summarization on Mobile Devices

Venue

SIGIR'17

Topic labeling

Fully automated

Focus

Secondary

Type of contribution

Novel approach

Underlying technique

Graph-based TextRank algorithm Mihalcea and Tarau, 2004

Topic labeling parameters

INPUT: Set of tweets associated with a generated topic
PARAMETER(S):

- Graph reduction factor: 0.1 (i.e. top 10% of nodes by weight)

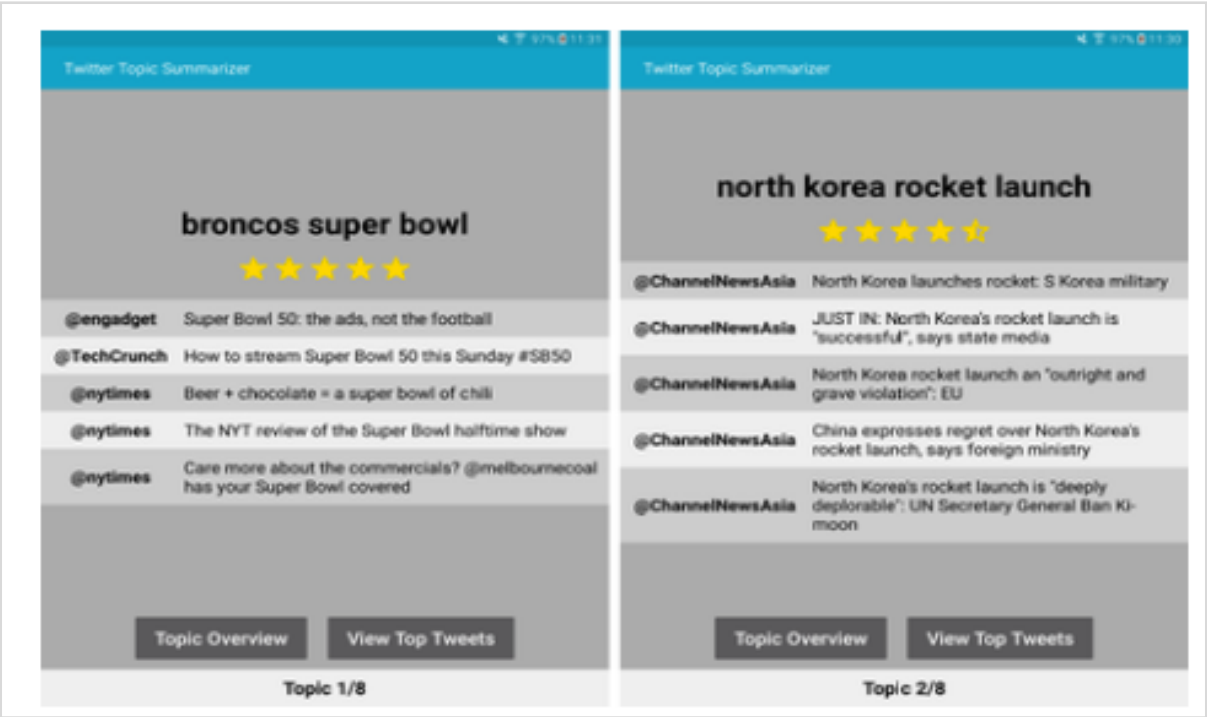
Label generation

Topic labels are extracted from graph nodes:

1. Tweets associated with a topic are used to constructs an undirected, weighted graph
    - Nodes represent words
    - Word co-occurrences are weighted edges.
2. Node values are iteratively updated (until convergence) based on the current values of its neighbours and weights of the edges connecting them, as well as the total weight of the outgoing edges of each of these neighbouring nodes.
3. Nodes are sorted in descending order of their final values, and filtered using a graph reduction factor.
4. Nodes are used as seeds to generate candidate topic labels.
    - Neighbours of the seeds are fetched and used to compute words permutations.
    - n-grams frequencies obtained based on the collection of tweets are used to evaluate the permutations.
    - Individual scores are multiplied with the edge weights connecting the nodes, and then normalised by the number of nodes (words) in the candidate label.
    - Normalization ensures that longer but not necessarily more important word sequences will not be preferred over shorter ones.

Motivation

In the context of this work, N topics are generated from a set of pre-processed tweets belonging to a given user's recent timeline.
Each topic is then associated with the top (30) most relevant tweets as computed by a dedicated Tweet Ranker Submodule.
Generating topic labels allows to produce more easily interpretable topic-tweets summaries:



Topic modeling

LDA (with symmetric Dirichlet distributions)

Topic modeling parameters

$\alpha = 0.005$

$\beta = 0.01$

$N = 8$

Nr of iterations = 2000

Nr. of topics

8 topics per timeline.

Label

Multi-word labels (2-, 3- or 4-grams) extracted from the graph generated by the set of tweets associated with a topic.

Label selection

/

Label quality evaluation

/

Assessors

/

Domain

Social media (Twitter)

Corpus

Set of recent tweets gathered from the timeline of a given user (800 per timeline).

Document

Textual content of a single tweet.
Single tweet made up of up to 140 characters.

Pre-processing

- Textual contents are converted to lowercase.
- URLs, embedded media, accents, emoticons, non-printable ascii characters are removed.
- Acronyms and abbreviations are converted to its original form (e.g., "govt" to "government", "SOA" to "service oriented architecture") by leveraging on a conversion dictionary (http://www.noslang.com/dictionary/)
- Bigrams that have better semantic meaning when treated as a single entity (e.g. "new year", "rocket launch", and "surface water") are joined (using an underscore character). Common bigrams are taken from Segaran and Hammerbacher, 2009
- The cleaned text is tokenised.
- Tweets with less than three tokens and near-identical tweets are discarded

---

```
@inproceedings{chin_2017_totem_personal_tweets_summarization_on_mobile_devices,
author = {Chin, Jin Yao and Bhowmick, Sourav S. and Jatowt, Adam},
title = {TOTEM: Personal Tweets Summarization on Mobile Devices},
year = {2017},
isbn = {9781450350228},
publisher = {Association for Computing Machinery},
address = {New York, NY, USA},
url = {https://doi.org/10.1145/3077136.3084138},
doi = {10.1145/3077136.3084138},
abstract = {Tweets summarization aims to find a group of representative tweets
for a specific topic. In recent times, there have been several research efforts
toward devising a variety of techniques to summarize tweets in Twitter.
However, these techniques are either not personal (i.e., consider only tweets
in the timeline of a specific user) or are too expensive to be realized on a
mobile device. Given that 80% of active Twitter users access the site on mobile
devices, in this demonstration we present a lightweight, personalized, on-
demand, topic modeling-based tweets summarization engine called TOTEM, designed
for such devices. Specifically, TOTEM summarizes most recent tweets on a user's
timeline and enables her to visualize and navigate representative topics and
associated tweets in a user-friendly tap-and-swipe manner.},
booktitle = {Proceedings of the 40th International ACM SIGIR Conference on
Research and Development in Information Retrieval},
pages = {1305–1308},
numpages = {4},
keywords = {mobile device, summarization, personal, tweets, topic modeling},
location = {Shinjuku, Tokyo, Japan},
series = {SIGIR '17}
}

@misc{noslang_website,
  title = {{Internet & Text Slang Dictionary}},
  author = {{NOSLANG}},
  howpublished = "\url{http://www.noslang.com/dictionary/}",
  note = "[Online; accessed October 2022 - June 2023]",
  year = 2023
}
```

```
@book{SegaranHammerbacher2009,
  abstract = {In this insightful book, you'll learn from the best data
practitioners in the field just how wide-ranging -- and beautiful -- working
with data can be. Join 39 contributors as they explain how they developed
simple and elegant solutions on projects ranging from the Mars lander to a
Radiohead video. Explore the opportunities and challenges involved in working
with the vast number of datasets made available by the Web; learn how to
visualize trends in urban crime, using maps and data mashups; discover the
challenges of designing a data processing system that works within the
constraints of space travel; learn how crowdsourcing and transparency have
combined to advance the state of drug research; understand how new data can
automatically trigger alerts when it matches or overlaps pre-existing data;
learn about the massive infrastructure required to create, capture, and process
DNA data.},
  added-at = {2017-05-08T14:09:48.000+0200},
  address = {Beijing},
  biburl = {https://www.bibsonomy.org/bibtex/2945136d61eb1f53fa7fd05abf3e9b506/
flint63},
  editor = {Segaran, Toby and Hammerbacher, Jeff},
  file = {O'Reilly eBook:2009/SegaranHammerbacher09.pdf:PDF;O'Reilly Product
page:http\://shop.oreilly.com/product/9780596157128.do:URL;Amazon Search
inside:http\://www.amazon.de/gp/reader/0596157118/:URL},
  groups = {public},
  interhash = {7a69c4856e84bc8a56679e09c7a622bf},
  intrahash = {945136d61eb1f53fa7fd05abf3e9b506},
  isbn = {978-0-596-15711-1},
  keywords = {01841 105 book shelf safari data processing management
information analysis graphics multimedia design zzz.big},
  publisher = {O'Reilly},
  timestamp = {2018-04-16T12:27:16.000+0200},
  title = {Beautiful Data: The Stories Behind Elegant Data Solutions},
  url = {http://norvig.com/ngrams/count_2w.txt},
  username = {flint63},
  year = 2009
}

@inproceedings{mihalcea_2004_textrank_bringing_order_into_text,
    title = "{T}ext{R}ank: Bringing Order into Text",
    author = "Mihalcea, Rada  and
```

```
      Tarau, Paul",
    booktitle = "Proceedings of the 2004 Conference on Empirical Methods in
Natural Language Processing",
    month = jul,
    year = "2004",
    address = "Barcelona, Spain",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/W04-3252",
    pages = "404--411",
}
```

#Thesis/Papers/Initial