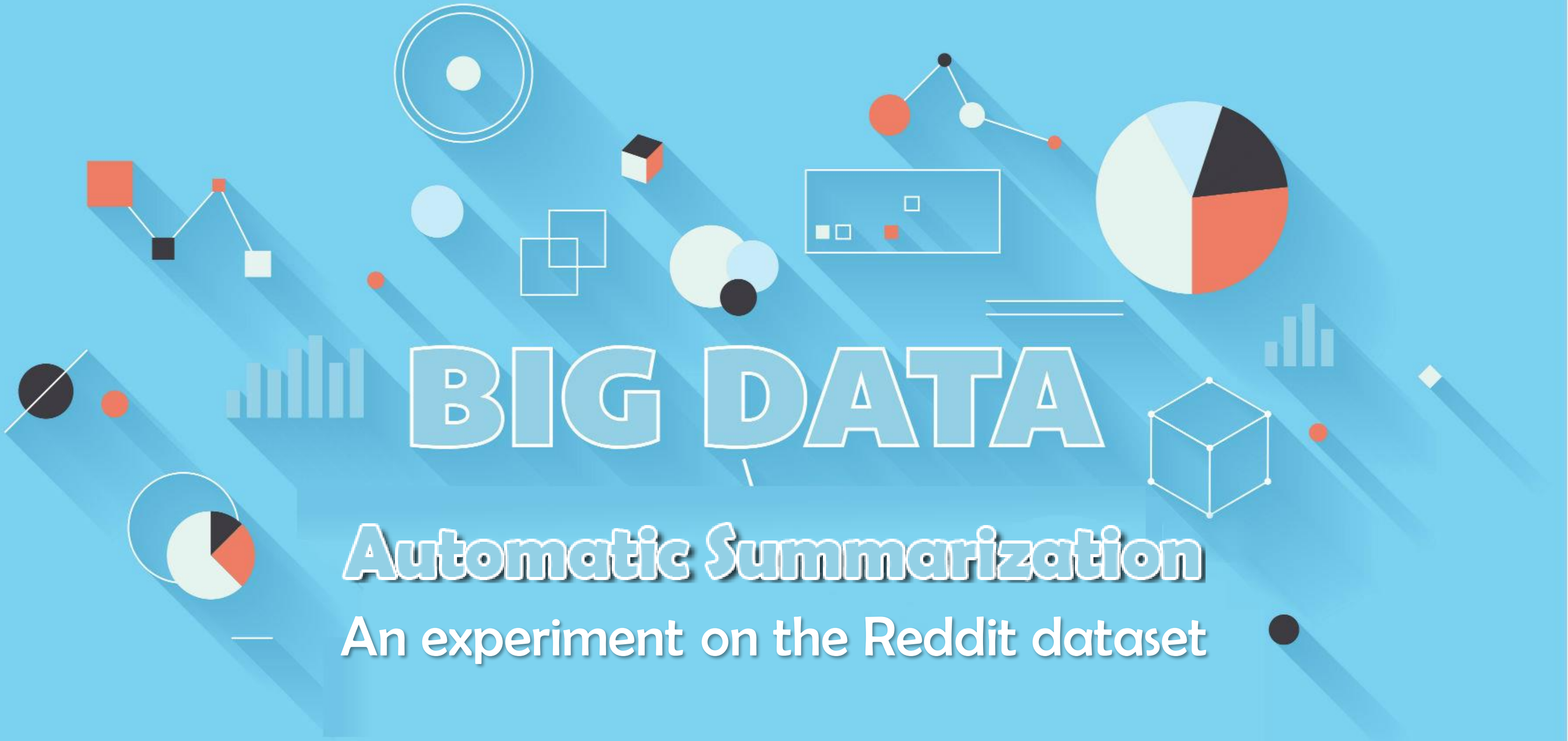


Automatic Summarization

An experiment on the Reddit dataset

Shahbaz Ahmed (115594)
Viorel Morari (115629)

The need for Summarization

- Goal

  - to capture the important information contained in large volumes of text, and present it in a brief, representative, and consistent summary

- TL;DR

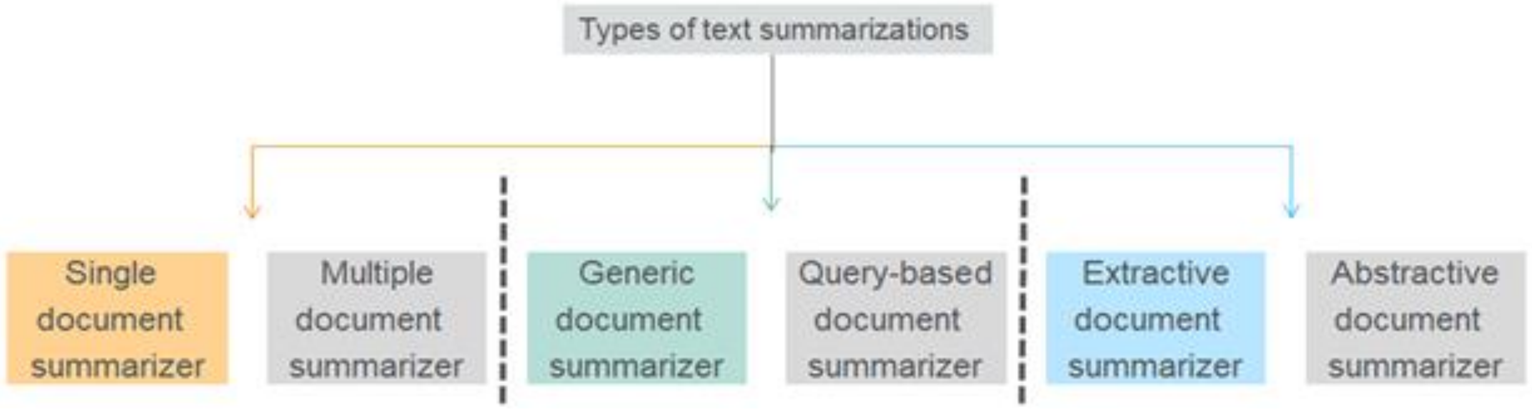
  - TLDR acronym expression stands for "Too Long, Didn't Read"

TRANSITION HOLISTIC TECHNOLOGIES EXTEND PROACTIVE E-COMMERCE WITH ...
ROBIN, GET TO THE POINT!
BATMANCOMIC.INFO

Types of Summarization

- Automatic summarization

  - reducing a text document or a larger corpus of multiple documents into a short set of words or paragraph that conveys the main meaning of the text



Types of text summarizations

Single document summarizer

Multiple document summarizer

Generic document summarizer

Query-based document summarizer

Extractive document summarizer

Abstractive document summarizer

Extractive vs. Abstractive

- <u>Extractive methods</u>

  - work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary

<u>Abstractive methods</u>

- build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate

Example

The Army Corps of Engineers, rushing to meet President Bush's promise to protect New Orleans by the start of the 2006 hurricane season, installed defective flood-control pumps last year despite warnings from its own expert that the equipment would fail during a storm, according to documents obtained by The Associated Press.

Extractive Summary : "Army Corps of Engineers," "President Bush," "New Orleans," and "defective flood-control pumps"

Abstractive Summary: "political negligence" or "inadequate protection from floods."

Existing work

- *Gupta and Lehal, 2010* – single document summarization
- *Goldstein et al. 2000* - summarization of multiple documents on the same topic **(~200 documents)**
- *Cselle, Albrecht, and Wattenhofer, 2007* - summarizing discussions such as email conversations **(~200 comments)**
- *Hu, Sun, and Lim 2007* – blogs summarization **(~1500 blog posts)**
- *Chakrabarti and Punera 2011* – tweets summarization **(440K tweets; over 150 games)**
- *Brody and Elhadad 2010* – reviews summarization

Our Dataset : The Reddit Universe!

- Comments

  149.6 GB : 1,659,361,605(~ 1.66 billion) entries

- Submissions

  39.7 GB : 196,531,736(~ 1.96 million) entries

Comment

- Comment - a statement of fact or opinion, especially a remark that expresses a personal reaction or attitude.

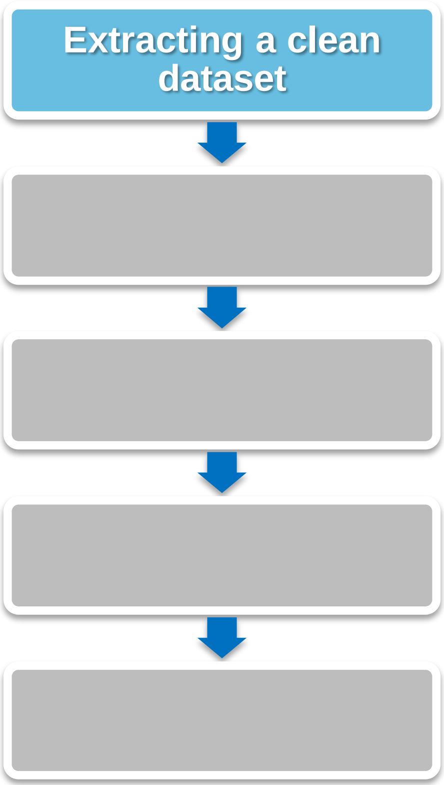
```
{"archived":true,"author":"jaquehamr","body":"Thanks for proving the point of the
quote.\n\nTL;DR: WOOSH", "controversiality":0, "created_utc":"1239192802",
"downs":0,"edited":"false", "gilded":0, "id":"c08q8en", "link_id":"t3_8auok",
"name":"t1_c08q8en","parent_id":"t1_c08q4sz","retrieved_on":1425950159,"score":3,"score_hidden":false,"subreddit":"atheism","subreddit_id":"t5_2qh2p","ups":3}
```

Submission

- Submission - a statement of fact or opinion posted by a registered user with the intention to be elaborated by other users.

{"archived":true,"author":"[deleted]","created":1297290547,"created_utc":"1297290547","domain":"self.WeAreTheFilmMakers","downs":0,"edited":"false","gilded":0,"hide_score":false,"id":"fibse","is_self":true,"media_embed":{},"name":"t3_fibse","num_comments":2,"over_18":false,"permalink":"/r/WeAreTheFilmMakers/comments/fibse/question_about_resumes/","quarantine":false,"retrieved_on":1442846972,"saved":false,"score":2,"secure_media_embed":{}, "selftext":"I'm currently a film student at the University of Cincinnati and I'm going to start applying for internships soon so I was wondering what I should put on my resume when applying.\n\ntl;dr I'm going to be sending out my resume soon and I'm looking for help on what I should include on it”, ”stickied":false,"subreddit":"WeAreTheFilmMakers", "subreddit_id":"t5_2qngr","thumbnail":"default","title":"Question about resumes", "ups":2, "url":"http://www.reddit.com/r/WeAreTheFilmMakers/comments/fibse/question_about_resumes/"}

Process of Summarization



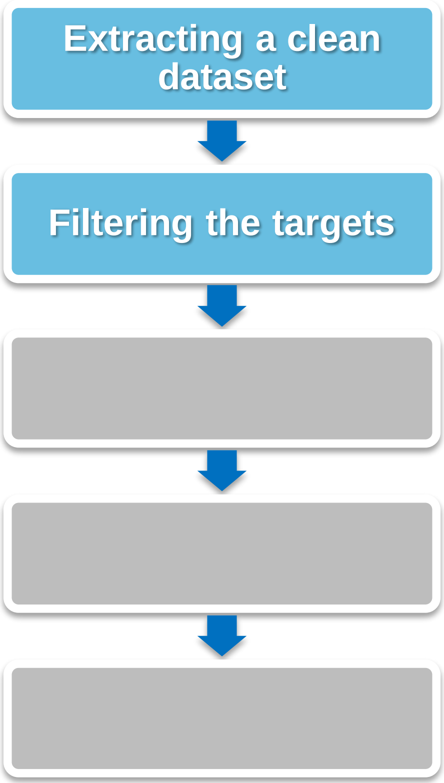
Extracting a clean dataset

"The most important tasks with regard to understanding the information available in comments are filtering, ranking and summarizing the comments." - *(Potthast et al. 2012)*

- Extract only the items which contain tl;dr

- **Challenge**
  - "body":"It's pretty sad that someone can sum up ten years of your life with a tl;dr"

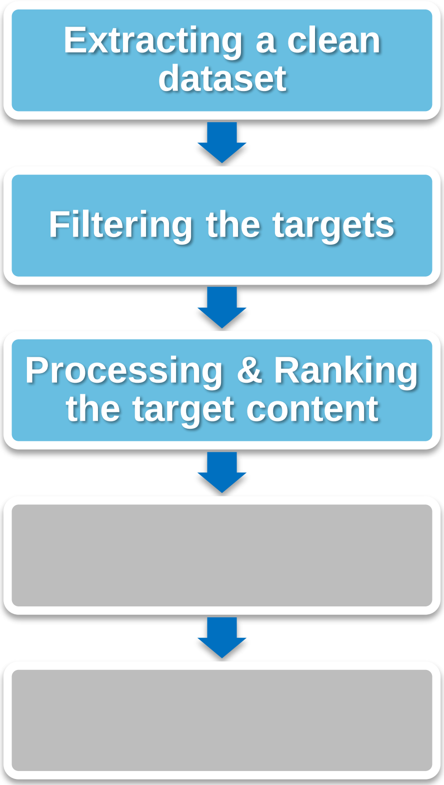
Process of Summarization



- Filter out comments/submissions with content length < 50 chars (our approach)
- E.g "body":"Thanks for proving the point of the quote.\n\nTL;DR: WOOSH“ – invalid

- Filter out tl;dr’s with content length < 5 chars (our approach)
- E.g. "body":"It's pretty sad that someone can sum up ten years of your life with a tl;dr“ - invalid

Process of Summarization


Extracting a clean dataset
Filtering the targets
Processing & Ranking the target content

- TF-IDF – ranking models

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

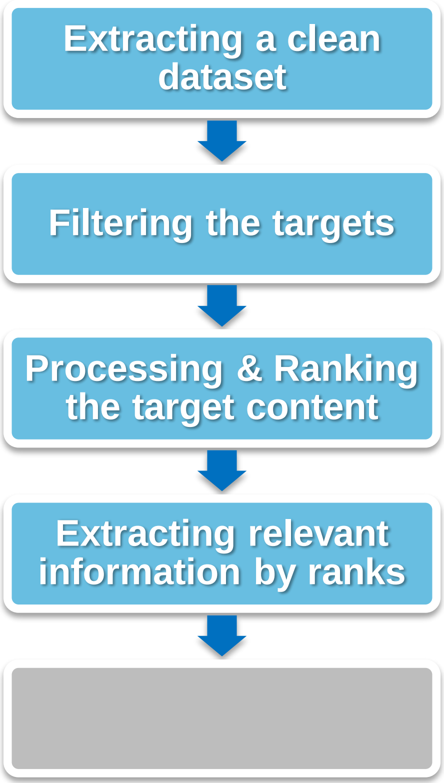
$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

- Reduces the influence of more common words

Process of Summarization



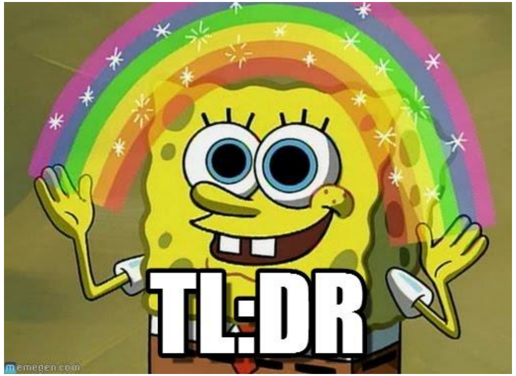
Extracting a clean dataset

Filtering the targets

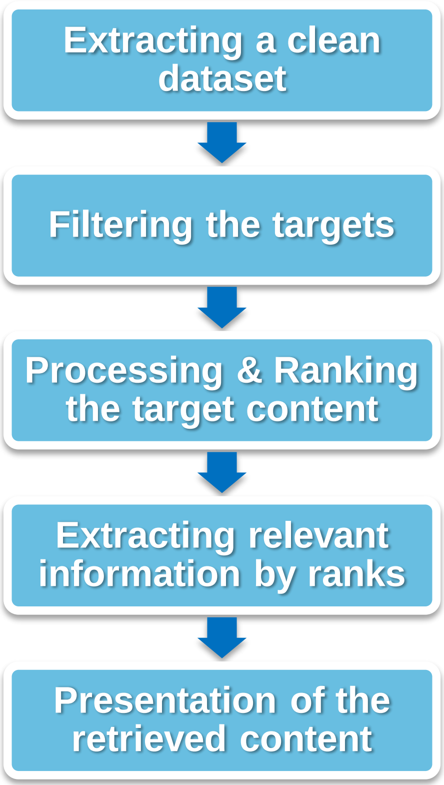
Processing & Ranking the target content

Extracting relevant information by ranks

- Highest rank terms form the summarization (tl;dr)



Process of Summarization



Extracting a clean dataset

Filtering the targets

Processing & Ranking the target content

Extracting relevant information by ranks

Presentation of the retrieved content

(12 sentences)

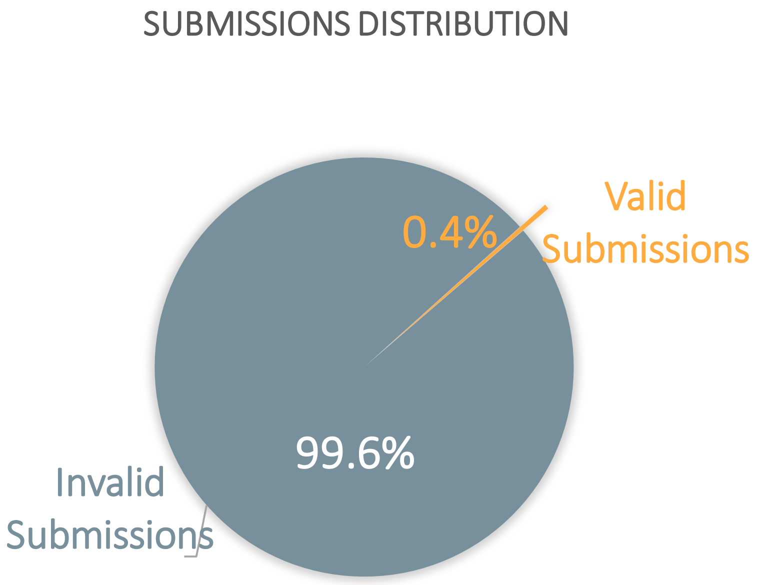
1. there used to be several **channels** related to technology and geek culture.
2. then it merged with g4tv, a shitty comcast **channel** of little note that wanted techtv's audience and cancelled all the decent reasons to ever tune into techtv.
3. there is nothing decent that comes on cable television that you can't watch for free (and legally) on either hulu or the comedy **channel's** website.

<u>Original tl;dr:</u>

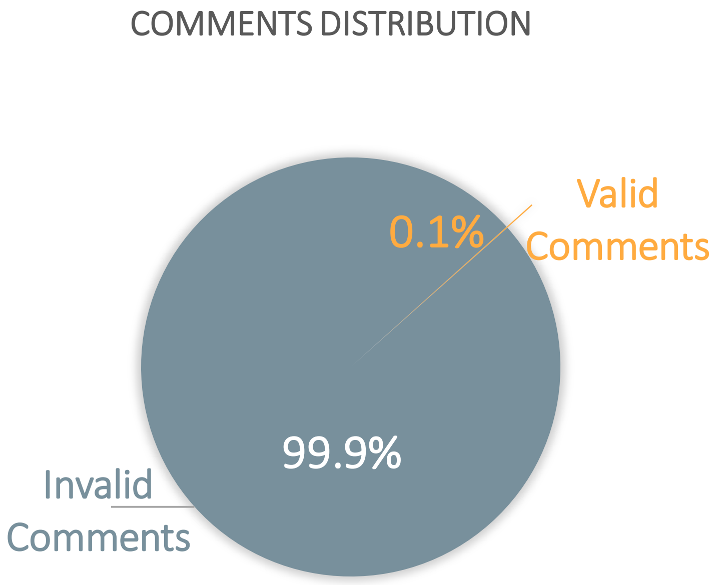
there was a decent one. comcast more or less bought it out and axed all it's programming to get viewers for its gaming **channel** but only succeeded in destroying the market and causing kevin rose to run off and create digg.

(2 sentences)

Statistics about the data set

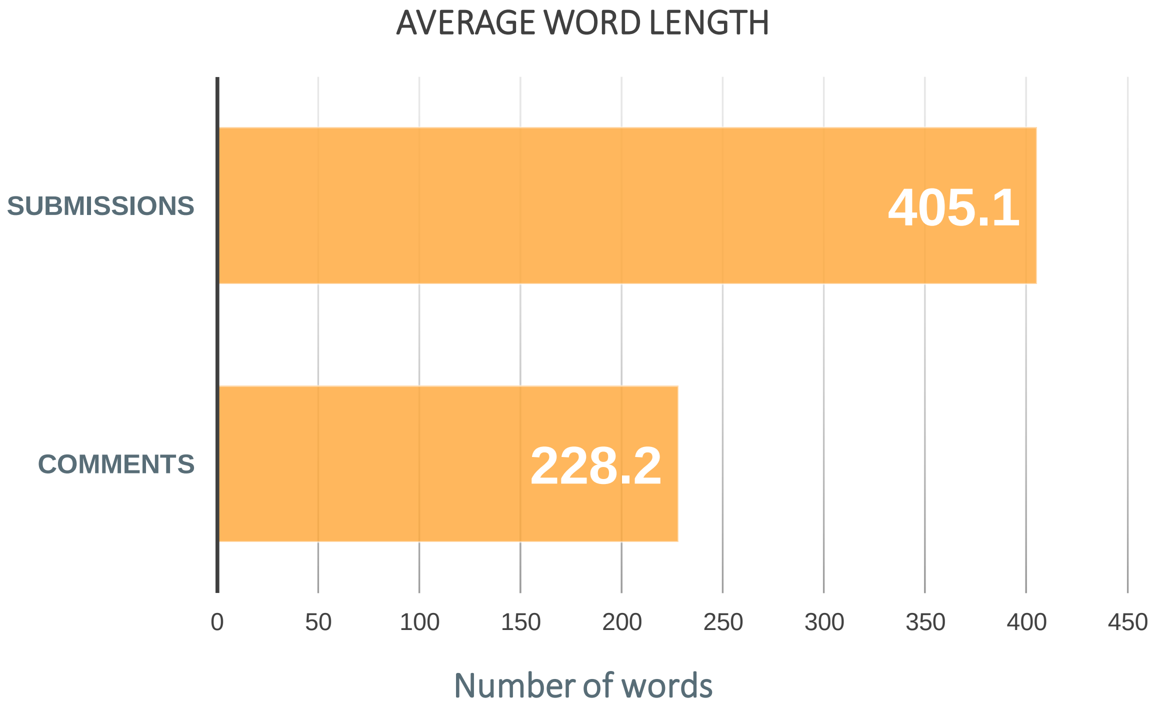


749376 (~ 0.75 million)

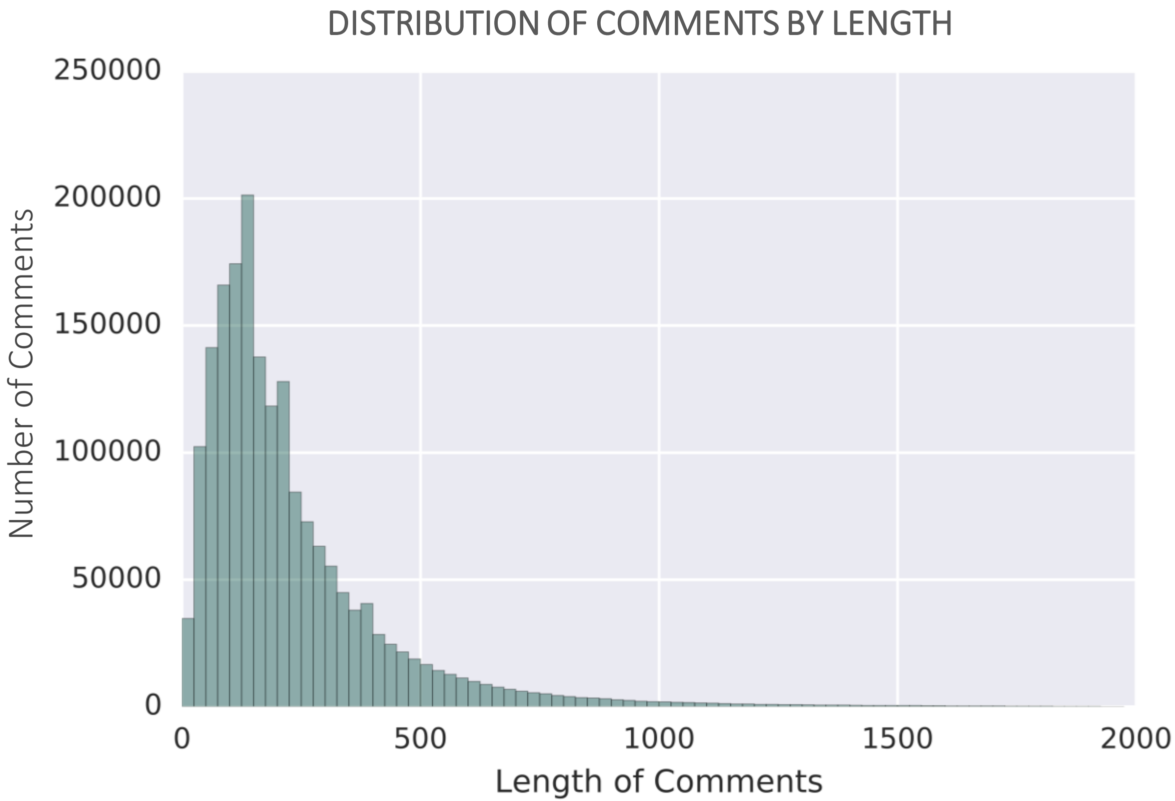


1,850,031 (~ 1.85 million)

Statistics about the data set

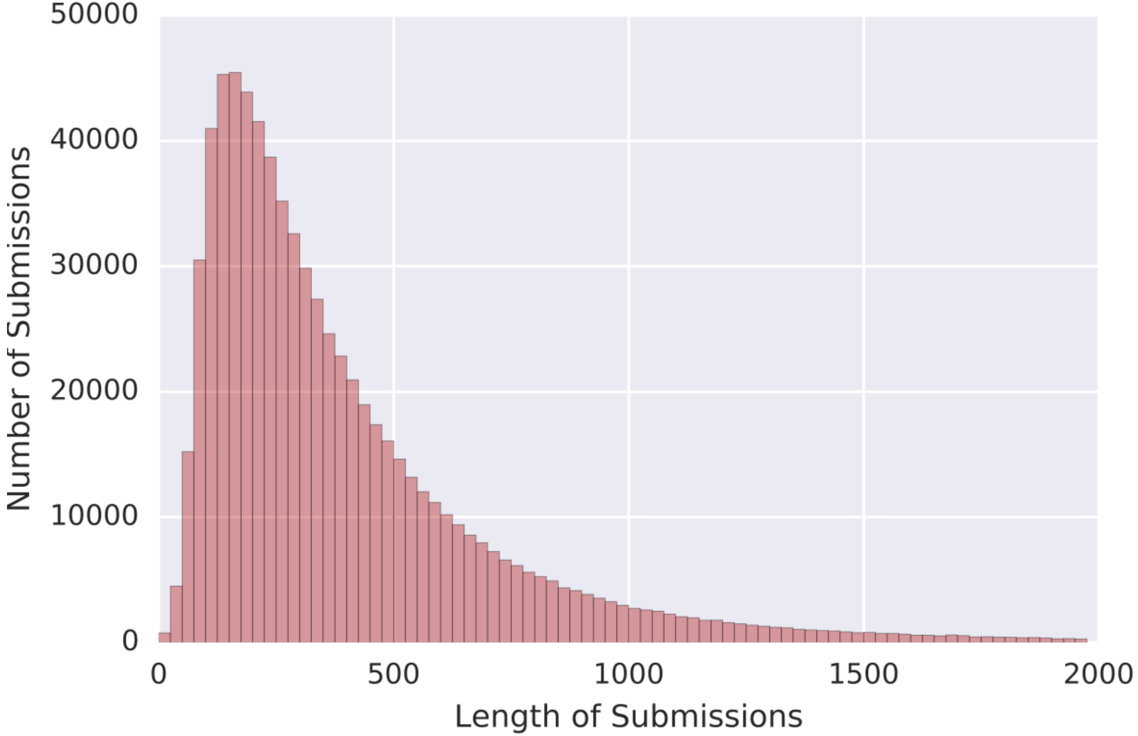


Statistics about the data set



Statistics about the data set

DISTRIBUTION OF SUBMISSIONS BY LENGTH



Further ideas

- Developing the automatic extractive summarizer on the valid comments & submissions (in progress)
- Dealing with tl;dr at the semantic level
  - "body":"It's pretty sad that someone can sum up ten years of your life with a tl;dr“
- Keyphrase extraction for summarization
- Form a proper representation of valid and invalid comments/submissions/tl;dr’s
- Dealing with encountered anomalies and faults in the detection process

Examples

- Valid Submission

```
"( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)
( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)( ͡° ͜ʖ ͡°)

"
" ( ͡° ͜ʖ ͡°)"
```

Examples

- Valid Comment

:omgwtfthatistotallyhitleri'llneverbeabletobuythatbrandoflotioneveragainthankyouforpointingthisouttomei'llsendanemailoutoeverylawyericanfindsothatthiscompanycanbebroughttojustice!)

- "**wordcount**":2

Conclusion

- Difficult to set up a good universal summarization tool (abstractive level)
- Our approach tends to generalize the idea of a comment/submission/tl;dr
- Yet the number of valid comments/submissions suggests of a good calibration
- The existing approach can be further improved

References

1. http://blog.mashape.com/list-of-30-summarizer-apis-libraries-and-software/
2. CS838-1 Advanced NLP:Automatic Summarization - Andrew Goldberg
3. Summarizing Newspaper Comments - Clare Llewellyn, Claire Grover and Jon Oberlander
4. Text Summarization using Singular Value Decomposition - Sharayu Rane
5. https://github.com/reddit/reddit/wiki/JSON
6. Automatic Summarization, Ani Nenkova and Kathleen McKeown
7. Automatic Summarization, Andrew Goldberg, 2007

Thank You!