Advanced Big Data Analytics

A11. Automatic Storytelling on Public Event

Xinluan Tian (xt2233)
Jiayao Wang (jw3514)
Cognitive Machine
02/13/2020

Abstract:

Social media is a new platform people can learn about news including public events. However, information on social media can be overwhelmed and fragmented. It’s very important to have a platform that can aggregate all information on one public event and automatically generate stories that describe the event in human style. We plan to implement a platform that can stream all the twitter text on a particular public event and use NLP techniques to extract key information about the event. An automatic storytelling algorithm will be implemented to generate human-write style stories from storyline extracted from key information.

1. Background (Review of Related Literature):

In the past, information was collected, curated by journalists and published. What we learned from journals or newspapers are filtered, well-written, sometimes even biased stories on certain topics. With the rise of social media, the web has become a vibrant and lively Social Media realm in which billions of individuals all around the globe interact, share, post, and conduct numerous daily activities. Collecting information on social media gives us a chance to get a huge amount of raw, on time and comprehensive information. The combination of social media and big data has created a new area of research, namely social media mining, which is similar to data mining but limited to the worlds of Twitter, Facebook, Instagram, etc. Social media mining is "the process of representing, analyzing, and extracting operational patterns from social media data."[1] In simple terms, social media mining occurs when a company or organization collects data about social media users and analyzes it for rendering. In the process. Conclusions about these user groups. Results are often used for targeted marketing activities in specific market segments.

Automatic storytelling is a process that involves using artificial intelligence (AI) to create written stories. Given a topic and a storyline, machines should generate a story in human written style which is easy to read by humans. Although automatic storytelling is still far away from building truly creative and insightful novels, it has been steadily improving while working on basic applications. Automated storytelling tools combine AI, machine learning, and big data to create content. In general, automatic storytelling can be used to write headlines, financial reports, and weather updates, as well as anything from screenwriters or short stories. Today, the practical use of automated storytelling is to use the process to "write" more technical headlines or reports and allow human writers to focus their time on more creative stories that may be less structured. Automated narratives begin by collecting large amounts of data into a database. This data may include information such as hundreds or thousands of different stories or titles. Tools such as Natural Language Processing (NLP) will then scan the data and parse it into structured data. Templates are created by humans, so AI can replace the information in the template with its own template. Templates can be lower-level, meaning they can be simple points that replace data values with AI (e.g., avid reports), or higher-level ones that are intended for more complex and meaningful writing Template. Natural language generation (NLG) is used to automatically generate text-based summaries from a database.

There are some successful automatic storytelling algorithms. Yao et al present a plan-and-write hierarchical generation framework that first plans a storyline, and then generates a story based on the storyline[2]. Ammanabrolu et al propose a Neural network-based approach to automated story plot generation attempts to learn how to generate novel plots from a corpus of natural language plot summaries[3]. Guan et al propose to utilize commonsense knowledge from external knowledge bases to generate reasonable stories to avoid repetition, logic conflicts, and lack of long-range coherence in generated stories[4].

2. Introduction to the Project:

In our project, we will perform social media data mining on Twitter data and generate stories from key information summarized from social media data.

Spark streaming together with Twitter developer API will be used to retrieve tweets on specific topics. Tweets will be preprocessed using standard text preprocessing procedures such as text lowercase, remove punctuation and stop words. There are also some steps for twitter data, such as remove #hashtag, URL, @username, typos, etc. Since there are many people posting tweets that are irrelevant to certain hashtags, it is essential to filter spam tweets.

In this day and age, the status of Twitter users is updated frequently and information is spread more quickly. The base of the number of people using Twitter is large, and hundreds of millions of tweets are generated every day. Thus, a clear classification of text could give us a good shot to analyze data more efficiently. What's more, detecting users' emotions is another big deal for us to quantify different attitudes towards the same public event.

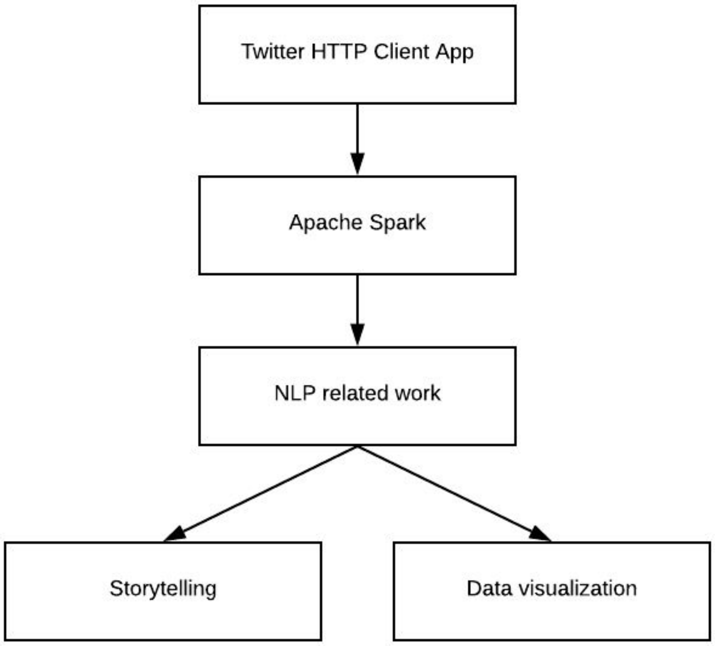

Figure 1: flow chart of our project.

Automatically storytelling will be at two levels: 1. key information extracted from huge amounts of twitter data on the public event and 2. machine-generated human written style summary for an event of interest. For level 2, It will be formulated as a conditional generation problem. We will adopt Yao et al. 's seq2seq model, as shown in Figure2, topics and the planned storyline will be encoded into a low-dimensional vector. Then the seq2seq model will be trained by minimizing the negative probability of the stories in the training data.

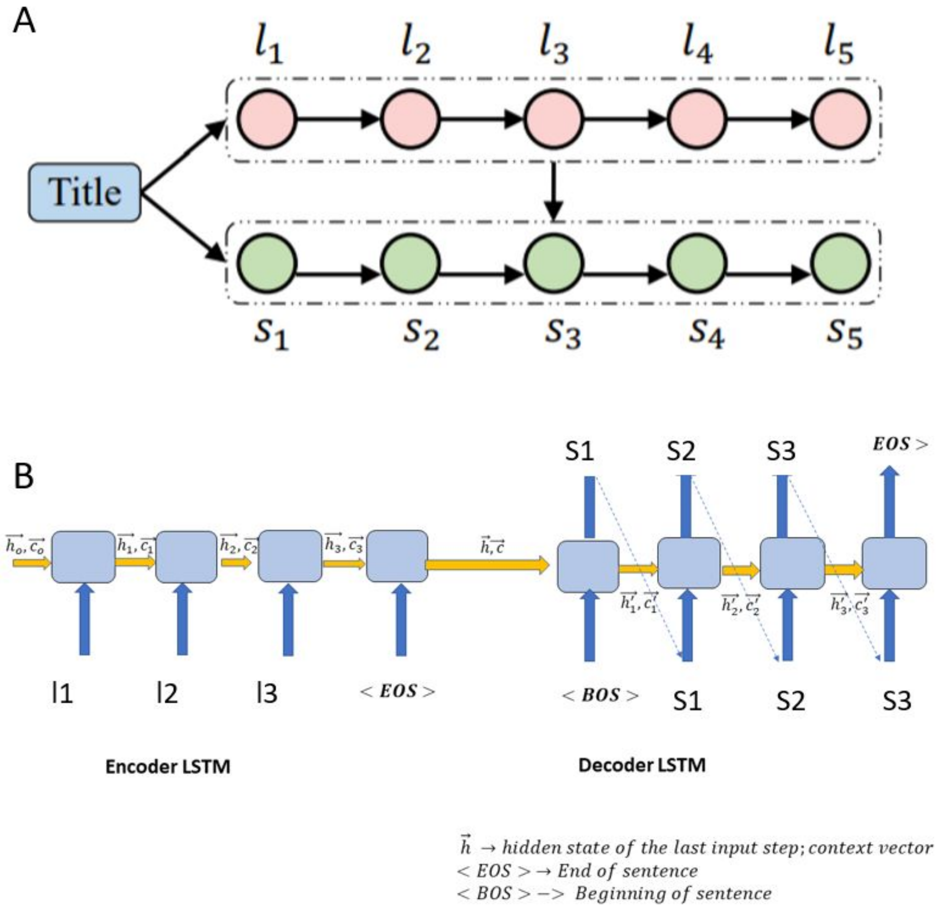

Figure 2: seq2seq model for generating story from storyline. A: story Si will be generated from planned storylines li. B: seq2seq model details, LSTM will be used to model content relationship between storylines and stories.

3. Introduction to the Dataset:

We will use the Twitter dataset as our primary dataset. Twitter developer has many APIs which allow us to search and retrieve tweets on certain topics or hashtags. Due to the enormous quantity of tweets, that information may be useful if we process them effectively. Here are several steps for us to get the Twitter dataset. Firstly, we register on Twitter Developer to create a new app to get secret keys and tokens for accessing online data streams. Secondly, we will get tweets of different hashtags by using Twitter HTTP client. Another important part is to preprocess the twitter data to remove non-useful terms and typos for optimal output.

For storytelling model training, we will use published news as training data. Dataset such as Millions of News Article URLs (2.3 million URLs for news articles from the front page of over 950 English-language news outlets in the six month period between October 2014 and April 2015)[5] and NYTimes Facebook Data (all the NYTimes Facebook posts). We will manipulate each article in the training dataset into the format of storyline with a similar process of summarizing key information from tweets, then use the storyline and original article to train the seq2seq model. The biggest challenge here could be, context or topics in our training data may be limited and may end up with a model that can only perform well on particular types of topics. We will try to find more training data or limit our application fields to maximize model performance.

4. Plan:

Milestone 1 (by Feb 21th):

- Data collection on social media: use Twitter API to read online stream and process the tweets using Apache Spark Streaming.
- Data cleaning: implement text lowercase, removing stop words, removing repeated characters in words, removing URL, removing punctuation, tokenization, etc.

Milestone 2 (by March 27th):

- Spam filter: filter irrelevant tweets on certain hashtags
- Text classification: classify context to different categories
- Topic modeling: use the Latent Dirichlet Allocation (LDA) model to analyze some documents to extract their topics, and then perform topic clustering or text classification based on the distribution.

Milestone 3 (by April 24th):

- Visualization: draw a timeline of each event to interpret the story and do sentiment analysis to quantify the attitudes of the authors hidden beneath their tweets.
- Storytelling: Train a seq2seq model with published news dataset to generate a story from planned storylines.

Reference:

[1] R Zafarani, MA Abbasi, H Liu. Social media mining: an introduction
[2] Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, Rui Yan. Plan-And-Write: Towards Better Automatic Storytelling. (2019)
[3] Ammanabrolu, Prithviraj, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara J. Martin and Mark O. Riedl. “Guided Neural Language Generation for Automated Storytelling.” (2019).
[4] Guan, Jian, Fei Huang, Zhihao Zhao, Xiaoyan Zhu and Minlie Huang. “A Knowledge-Enhanced Pretraining Model for Commonsense Story Generation.” ArXiv abs/2001.05139 (2020): n. pag.
[5] Jia, Sen et al. (2016), Data from: Women are seen more than heard in online newspapers, Dryad, Dataset, https://doi.org/10.5061/dryad.p8s0j