CS474 Text Mining

# Term Project Description

Fall 2021

## 1. Project Overview

The goal of the term project is to give the students an opportunity to have a hands-on experience on addressing key text mining problems by exploring and applying various text mining concepts and techniques appropriate for solutions. Each team of three students needs to design and implement a system for the tasks in the problem description below, preferably by developing new ideas beyond simply applying existing tools. We hope that the project is a good instance of problem-based learning: you will get to not only reinforce your understanding of the related skills and knowledge but also acquire and attempt to utilize new techniques in a real context.

We will attempt to provide appropriate resources and pointers about what you should be doing, but an important part of this project is that you should explore various tools and models. If there are things that you feel you need to know more about, however, feel free to ask. We can hopefully at least point you in useful directions.

The first step is to form a team consisting of three students. Each team should perform every task in the description below.

## 2. Tasks

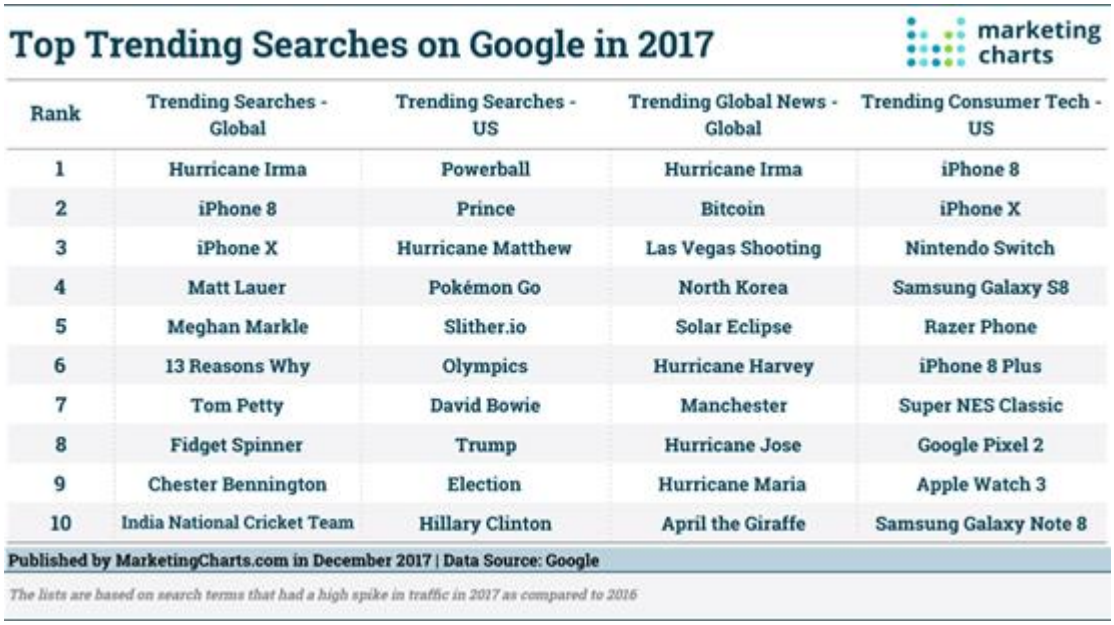
The tasks are similar to the one in Homework 1, but this time, you will be given a dataset constructed by a set of crawled news articles from Korea Herald. Given the dataset, you are going to conduct two different analyses tasks as follows: **issue trend analysis** and **issue tracking analysis**.

### I. Issue Trend Analysis

This task is to find the top ten most significant issues for each year and rank them, from the news articles over the period of three years. The criteria for selecting and ranking

the issues should be clearly defined by each team. (ex. the number of related news articles, the length of period for the coverage, …). **Figure 1** below is the Google Trending list in 2017. A topic modeling technique was applied to automatically extract the issues from a collection of news articles and present the list. **The data that will be provided for this project is a collection of Korean newspaper articles.**

The trending list for 2015-2017 is published on some sources[1][2][3]. Compare your result against these lists as a way to verify whether your result is on the right track. The lists are not a gold answer by any means (it is just an example), but you can use them as a reference to ensure that your result is reasonable. The lists were in Korean, but the translated version is shown in **Figure 2** below for international students.



**Top Trending Searches on Google in 2017**

marketing charts

| Rank | Trending Searches - Global | Trending Searches - US | Trending Global News - Global | Trending Consumer Tech - US |
|---|---|---|---|---|
| 1 | Hurricane Irma | Powerball | Hurricane Irma | iPhone 8 |
| 2 | iPhone 8 | Prince | Bitcoin | iPhone X |
| 3 | iPhone X | Hurricane Matthew | Las Vegas Shooting | Nintendo Switch |
| 4 | Matt Lauer | Pokémon Go | North Korea | Samsung Galaxy S8 |
| 5 | Meghan Markle | Slither.io | Solar Eclipse | Razer Phone |
| 6 | 13 Reasons Why | Olympics | Hurricane Harvey | iPhone 8 Plus |
| 7 | Tom Petty | David Bowie | Manchester | Super NES Classic |
| 8 | Fidget Spinner | Trump | Hurricane Jose | Google Pixel 2 |
| 9 | Chester Bennington | Election | Hurricane Maria | Apple Watch 3 |
| 10 | India National Cricket Team | Hillary Clinton | April the Giraffe | Samsung Galaxy Note 8 |

Published by MarketingCharts.com in December 2017 | Data Source: Google

*The lists are based on search terms that had a high spike in traffic in 2017 as compared to 2016*

Figure 1. Top trending list in 2017

| 2015 |
| --- |
| Sewol ferry |
| MERS |
| History textbook nationalization |
| National Intelligence Service |
| IS |
| aversion |
| Daycare center |
| Megalia |
| Silver spoon / Plastic spoon |
| Hell-Chosun |

| 2016 |
| --- |
| President Park |
| Choi Soon-sil national affair scandal |
| Sewol |
| Gangnam Station Murder |
| United States presidential election |
| Filibuster |
| Japanese military sexual slavery agreement between Korea and Japan |
| AlphaGo vs. Lee Se-dol |
| Oxy humidifier sterilizer |
| Brexit |

| 2017 |
| --- |
| president impeachment |
| Moon Jae-in government launched |
| North Korea nuclear test |
| China's THAAD Revival |
| Pohang earthquake/Korea SAT delay |
| brutal crime (lee-young hak, elementary ) |
| Japanese military sexual slavery agreement between Korea and Japan |
| AlphaGo vs. Lee Se-dol |
| Oxy humidifier sterilizer |
| Brexit |

Figure 2. Trending lists for 2015-2017 (Most Koreans should be able to recognize the actual events. If you want to understand each keyword for better analysis, ask Korean colleagues. Our TA will be also happy to answer your questions.)


**Example (Issue trend analysis)**

(* Make sure to print the output of your program as in this example! *)

**Output**

$ >>

2015 : Sewol ferry, MERS, ...
2016 : President Park, Choi Soon-sil national affair scandal, ...
2017 : President impeachment, ...

*The issues should be in order by the ranking criteria of your choice! (e.g., length, importance, etc.)

# II. Issue Tracking

Your task is to track down the events related to two of the issues identified already for the first task. The first step is to choose two issues most suitable for detailed analyses. The next step is to automatically identify/extract the events related to each of the issues from the news articles. The number of events should be at least five for each issue. Finally, you will have to extract detailed facts such as people, organizations, and places, for each of the two events and the pertinent events that come along with the main issue. This issue tracking should be done in two different ways: "On-issue Event Tracking" and "Related-issue Event Tracking" as described below.

## 1) On-issue Event Tracking

The result should describe a sequence of the events specifically tied to the issue on a temporal line. Find detailed information for the detected events, including Person, Organization, and Place. Extracting additional information is optional. If there are multiple occurrences of an entity like a place associated with the event, they all must be included.

**Example (On-issue Event Tracking):**

(* Make sure to print the output of your program as in this example! *)

**Output**

$ >>

[ Issue ]

North Korea Nuclear Test

[ On-Issue Events ]

North Korea 1st Nuclear Test -> Trump tweets -> North Korea 2nd Nuclear Test -> South-North Summit -> North Korea-United States summit -> …

[ Detailed Information (per event) ]

Event: North Korea 1st Nuclear Test

- Person: Kim Jung Un, …
- Organization: North Korea Government, the Labor Party, …
- Place: North Korea nuclear launch site, ...

Event: Trump tweets

- Person: Donald Trump, …
- Organization: The United States Government, the White House

## 2) Related-issue Event Tracking

Extract and describe related events that are not directly tied to the particular issue. The core of this task is to identify the events that are topically related but not directly linked to the current issue, in terms of time, place, and participants. The types of information to be extracted for each event are the same as in "On-issue Tracking".

**Example (Related-issue Event Tracking):**
(* Make sure to print the output of your program as in this example! *)

**Output**

$ >>

[ Issue ]

North Korea Nuclear Test

[ Related-Issue Events ]

Iran Nuclear Test, DMZ Crossfire, Kim Jung Un Health Issues Reemerge, ...

[ Detailed Information (per event) ]

Event: Iran Nuclear Test

- Person: Hassan Rouhani, ...
- Organization: Iranian Government, ...
- Place: Iran Nuclear Testing Facility, Tehran

Event: DMZ Crossfire

- Person: Moon Jae In, …
- Organization: The Blue House, Korean Government
- Place: South Korea, North Korea, 38th parallel

# 3. Submission & Grading

Your submission should include:

- **A program package** of your system that contains all of your source code, data, executable file and other libraries. For a programming language that is hard to make an executable file with, you should make a requirement file (**requirements.txt**) which contains the libraries necessary to execute the program.

- **Provide an executable shell script**. For those using Python, make sure your program can be run by executing a shell script (**run.sh**). We will NOT run your program with **python main.py** (because we do not know what kind of extra arguments we need to use to reproduce your results). If you want to include extra arguments (e.g., batch size, smoothing, other hyperparameters), then include it in **run.sh** (make sure to create them as "executables" - use **chmod 755 [file_name]**).

- **You are free to use any model (or methodology).** Make sure to specify the reason as to why you decided to employ such a model (or take such an approach).

- **Output of your program.** As specified in the examples in Section I and II, format your output accordingly. Please make sure your **outputs match the ones in the examples** above (for the sake of readability).

- **API documentation** about your program that explains the package you used and how the function works.

- **A report for the tasks**. It should include an explanation of your project in detail within 5~6 pages in the ACM word template (http://www.acm.org/publications/proceedings-template).

A guideline for your report and presentation:

- You can choose how you will show the result of the tasks with different presentation methods, such as a table, a graph, a mind-map, …

- You should set certain criteria for your result evaluation.

The criteria must be reasonable and explained clearly.

- Both quantitative and qualitative analyses should be included. A qualitative analysis usually shows some example cases to demonstrate the system either works as intended and/or reveal some problems or limitations. If you can find some gold labels, compute accuracy.

We will grade your project submissions using the following criteria:

- **Documentation**: We will read your documentation to understand the techniques that you used and test the system.

- **Report**: All your team tried and accomplished must be well documented and explained clearly in the report. This is a very important source of information for us to grade the project. As we chose ACM word template, you need to make the quality of your report at a publishable level in terms of presentation and content. Your report should address the problem statement, relevant work, your solution to the problem, a deep analysis of your solution from various perspectives. It should also include the reasoning behind your design decisions.

- **Class Presentation:** Your team presentation is another chance to demonstrate the quality of your work for the project. Your presentation should contain the key content of the report. We will make an arrangement for scheduling later.

**Be sure to include who did what in each team.**

# 4. Important Dates

1. Project Design Paper Due: **November 3rd, 01:00 PM**

   Submit a project design paper (3~4 pages), which should include: a problem definition, a brief description of the overall ideas & approaches, related work, and possibly intermediate results. Based on the design paper, we will provide you with feedback (so, the earlier you submit the paper, the earlier you'll be receiving our feedback!).

2. Project Submission Due: **December 7th, 01:00 PM**

   We will open a board for submissions and make an announcement later. Because of the administration deadlines, we cannot accept any late submissions.

3. Project Presentation: **December 9th**

We will schedule a team presentation to be done **during the week prior to the final exam period**. Each team will be presenting for 20 minutes each. Your team's presentation should include demonstration of how your program works and explanation about the system using PowerPoint slides. This should be a team effort.

# 5. Supplementary Details

- **GPUs provided.** We will be providing **RTX2080TI-11G** (12x) to those who wish to run larger models in their project (e.g., BERT, RoBERTa). Each team will be assigned a GPU, and for those who require more compute power, please send an e-mail as early as possible because any additional remaining GPUs will be first come, first served.
- **Dataset.** The Korea Herald dataset will consist of 23,769 news articles (i.e., instances).

| | title | author | time | description | body | section |
|---|---|---|---|---|---|---|
| 0 | A snapshot of multiculturalism in South Korea | Lee Sun-young | 2018-01-01 17:07:00 | With birthrates persistently low and the senio... | With birthrates persistently low and the senio... | Social affairs |
| 1 | [Weekender] Korea's dynamic 2017 | Choi He-suk | 2018-01-01 13:22:00 | From North Korea's nuclear weapons program nea... | From North Korea's nuclear weapons program nea... | Social affairs |
| 2 | People's Party members support Ahn's push for ... | Yonhap | 2017-12-31 16:18:00 | The leader of the center-left People's Party g... | The leader of the center-left People's Party g... | Politics |
| 3 | [Newsmaker] Panamanian vessel probed over susp... | Yonhap | 2017-12-31 14:55:00 | PYEONGTAEK -- South Korea has seized and insp... | PYEONGTAEK -- South Korea has seized and insp... | North Korea |
| 4 | Hong Kong ship crew questioned in S. Korea for... | AFP | 2017-12-30 15:44:00 | The crew of a Hong Kong-registered ship have b... | The crew of a Hong Kong-registered ship have b... | North Korea |
| ... | ... | ... | ... | ... | ... | ... |
| 23764 | Korea Navy to build 1st dedicated training shi... | KH디지털2 | 2016-07-19 13:29:00 | Korea's Navy will build a dedicated training s... | Korea's Navy will build a dedicated training s... | Defense |
| 23765 | Korean man gets 4-year jail term for Yasukuni ... | KH디지털2 | 2016-07-19 13:16:00 | A Tokyo court handed down a four-year jail ter... | A Tokyo court handed down a four-year jail ter... | Social affairs |
| 23766 | N. Korean ferry to sail on 3-country tour rout... | KH디지털2 | 2016-07-19 13:13:00 | The North Korean passenger ferry Mangyongbong ... | The North Korean passenger ferry Mangyongbong ... | North Korea |
| 23767 | Presidential office denounces allegations invo... | KH디지털2 | 2016-07-19 13:09:00 | The presidential office Cheong Wa Dae on Tuesd... | The presidential office Cheong Wa Dae on Tuesd... | Politics |
| 23768 | PM says THAAD deployment does not require parl... | KH디지털2 | 2016-07-19 13:07:00 | Prime Minister Hwang Kyo-ahn on Tuesday said ... | Prime Minister Hwang Kyo-ahn on Tuesday said t... | Defense |

23769 rows × 6 columns

---

[1] https://www.huffingtonpost.kr/2015/12/13/story_n_8797824.html

[2] https://www.huffingtonpost.kr/2016/12/18/story_n_13713926.html

[3] https://www.bloter.net/newsView/blt201712140003