CS474 Text Mining
Term Project Description
Fall 2019

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. The students need to design and implement a system for the tasks, preferably by developing new ideas beyond simply applying existing tools.

In this project, you will work on an application wherein one can examine some important concepts and techniques discussed throughout the course and explore new ones. 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 ask questions if there are things that you feel you need to know more about. 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

Tasks are based on the crime analysis of Homework 1 with a slight extension of the range of the analysis to be done. There are two tasks to carry out in this project: issue trend analysis and issue tracking.

I. Issue Trend Analysis

The task is to find 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, …). F**igure 1** below is an English translation of the Google Trending list in 2017 written in Korean. 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, 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.

[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] http://www.realmeter.net/2017/11/2017-%EC%98%AC%ED%95%B4%EC%9D%98-%EC%9D%B4%EC%8A%88-1%EC%9C%84-%E6%9C%B4-%EC%A0%84-%EB%8C%80%ED%86%B5%EB%A0%B9-%ED%83%84%ED%95%B5-2%EC%9C%84-%E6%96%87%EC%A0%95%EB%B6%80-%EC%B6%9C/

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	2016	2017
Sewol ferry	President Park	president impeachment
MERS	Choi Soon-sil national affair scandal	Moon Jae-in government launched
History textbook nationalization	Sewol	North Korea nuclear test
National Intelligence Service	Gangnam Station Murder	China's THAAD Revival
IS	United States presidential election	Pohang earthquake/Korea SAT delay
aversion	Filibuster	brutal crime (lee-young hak, elementary)
Daycare center	Japanese military sexual slavery agreement between Korea and Japan	Japanese military sexual slavery agreement between Korea and Japan
Megalia	AlphaGo vs. Lee Se-dol	AlphaGo vs. Lee Se-dol
Silver spoon / Plastic spoon	Oxy humidifier sterilizer	Oxy humidifier sterilizer
Hell-Chosun	Brexit	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.)

II. Issue Tracking

Your overall task is to track down the events related to an identified issue over time. The first thing to do is to choose two issues for detail analysis, among the issues found for the first task. The next step is to 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 information such as people, organizations, and places, for each event extracted already. 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

Describe the change of events related to the issue with a timeline. Detect events related to the issue and find detailed information including Person, Organization, and Place. Extracting additional information is optional. List up all the people, organizations, and places that are related to the issue. There can be multiple related entities (e.g. multiple organizations) or none.

Example:

Issue: North Korea Nuclear Test
Event: North Korea 1st Nuclear Test → Trump tweets → North Korea 2nd Nuclear Test → South-North Summit → North Korea-United States summit → …
Detailed Information:
North Korea 1st Nuclear Test
-Person (Kim jung eun, …)
- Organization (North Korea Government, …)
- Place (North Korea nuclear test site, …), …

Trump tweets
– Person (Donald Trump, …), …

2. Related-issue Event Tracking

Extract and describe related events that are not directly linked 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 are the same as in "On-issue Tracking".

Example:

Issue: Alpago vs. Lee Se-dol
Event: Online AI Go game program, Chess AI match, international artificial intelligence world-cup, …
Detail Information:
…
Chess AI match
– Person (Garry Kimovich Kasparov, …)
- Organization (IBM, …)
- Place (Commonwealth of Pennsylvania, …), …

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 executable file with , you should make a requirement file which contains the information about executing the program.
- **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 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 6th**
   Submit a project design paper (3~4pages), 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 a feedback. The exact schedule for advising sessions will be announced later.

2. Personal Consulting Session: **November 7th and 8th.**
   There will be a face-to-face consulting session. You can ask questions if any and get some feedback about your ideas, model designs, etc. You can apply for the consulting session on KLMS soon.
3. Project Submission Due: **December 8th**
   We will open a board for submissions and make an announcement later. Because of the administration deadlines, we cannot accept any late submissions.
4. Project Presentation: We will schedule a team presentation to be done **during the week prior to the final exam period**
5. Your team's presentation should include demonstration of how your program works and explanation about the system using PowerPoint slides. This should a team effort.