|  |
| --- |
| Nanyang Technological University |
| Assignment Report |
| CZ4034 Information Retrieval |

|  |  |
| --- | --- |
| **Assignment Group** | 25 |
| **Group Member** | **Matriculation Number** |
| Soh Teck Seng | U1222654G |
| Tan Boon Keat, Winston | U1222265E |
| Tan Chan Wei | U1222128B |
| Tan Chao Jun | U1222824F |

|  |
| --- |
| This assignment involves (1) crawling a text corpus of interest, (2) building a search engine to query over the corpus, and (3) performing text classification and clustering. |

5 Submission

You should submit the following items via the course site, where instructions will be given in due course:

\_ A video presentation up to 5 minutes, uploaded to YouTube. In the video, introduce your group members and their roles, explain the applications of your works and their impact, and highlight, if any, creative parts of your works.

\_ A document that contains your answers for all the questions above. Note that you do not have to give all the answers in the video presentation.

\_ A zip (or gzip) file with crawled text data, queries and their results, manual classifications, automatic classification results, and any other data for Questions 3 and 5.

\_ A zip (or gzip) file with all your source codes and libraries, with a readme file that explains how to compile and run the source codes. If the file is too big to be uploaded to the course site, you can upload it to any other site (e.g., dropbox) and share the link with us.

In the case of multiple submissions, only the latest submission will be graded and time-stamped.

Motivations and goals are too general.  
Please list and better explain specific motivations and goals of your project, e.g., why are you doing what you are doing? what is a possible marketable application for it? how is your proposed system different from or better than available COTS systems? etc.  
  
No implementation details, no UI, no preprocessing, no classification, no evaluation, no innovations.  
  
Please take a more scientific approach to the project you are developing, i.e., (1) motivate every choice you make (e.g., why did you choose that specific keyword? why did you go for that classification method instead of another? etc.) and (2) prove that you were right in making such choices (e.g., calculate F-measure and compare obtained results with other possible choices and/or baseline methods).

# CZ4034 Information Retrieval

Course Assignment

# Objective

This assignment covers three main areas:

1. Crawling a text corpus of interest
2. Building a search engine to query over the corpus
3. Performing text classification and clustering

# Group

Assignment Group 25 consists of four members:

|  |  |
| --- | --- |
| **Assignment Group** | 25 |
| **Group Member** | **Matriculation Number** |
| Soh Teck Seng | U1222654G |
| Tan Boon Keat, Winston | U1222265E |
| Tan Chan Wei | U1222128B |
| Tan Chao Jun | U1222824F |

The source our group has chosen for crawling is **Twitter**, using the keywords “**European Union**”.

# Motivation

The recent years have been tumultuous for the European Union. Various happenings and events are occurring throughout the European Union including the Eurozone debt crisis, political tensions, fightings and war, etc. With the region being a hotbed of news and uncertainty, our group has decided to focus on the European Union as our area of interest. Hence, we have decided on the keywords “**European Union**”.

For our source, we have turned to **Twitter**. Being one of the major social networks on the Internet, Twitter boasts 288 million monthly active users with 500 million Tweets being sent per day.[[1]](#footnote-1) This reflects how real-time and updated Tweets are with respect to happenings across the world, including the European Union region. With this huge volume of Tweets as a basis, it suggests how much information Twitter and Tweets can present to us. This relates to the recent trend of Big Data as well. Therefore, Twitter presents itself to us as a treasure trove of information which we can tap from.

On the technical end, Twitter provides an in-house Application Programming Interface (API) for developers to crawl their tweets.[[2]](#footnote-2) Hence, this can help facilitate and ease the retrieval of information from Twitter’s database of information.

With **Twitter** as our source and “**European Union**” as our keywords, our group wishes to provide users with a fast, interactive and reliable **search engine** to find our latest happenings in the European Union. In order to do so, our group is exploring ways to enhance the indexing and ranking of relevant information in the index at the back-end of our search engine. Further **text classification and clustering** of the crawled information will be performed as well to present a logical categorisation of happenings in the European Union to users.

Our search engine uniquely differentiates itself from other search engines with its specific focus on the European Union. Information retrieved from our search engine is localised and specific to the European Union, unlike general search engines which retrieves broad-ranging information from various sources.

# Goal

There are three goals and milestones to be achieved.

## 1 Crawling a text corpus of interest

The first goal is to create a program that is able to crawl **Twitter** for relevant information regarding the **European Union** and to process the information for indexing.

## 2 Building a search engine to query over the corpus

The second goal is to create a web search engine based on the information stored in the index. The web search engine provides a front-end user interface for users to find out the latest happenings in the European Union. Also, our group will explore innovative methods in enhancing the speed of search queries and ranking to suit the user’s needs.

## 3 Performing text classification and clustering

The third goal is to perform classification on the collected information to identify interesting patterns which might provide initially unseen trends of information. This presents a logical categorisation of the happenings in the European Union to the users.

# Current Progress

~~Currently, we have created a standalone program that crawls twitter and collected 10,000 tweets (records) with 291416 words. For each tweet, the author, creation date, content, favourite counts and tags were collected and stored in a JSON format. A separate program was created to index the information into a standalone web server with Solr.~~

~~We are in the midst of creating a web application which will allow users to query information based on the data indexed.~~

# 1 Crawling

## Question 1

### Q1.1 How you crawled the corpus (e.g. source, keywords, API, library) and stored them (e.g. whether a record corresponds to a file or a line, meta information like publication date, author name, record ID)

[+]

First we created a Java program using eclipse and use the libraries from twitter4j.

After obtaining the authentication consumer key and access token, we proceed to crawl the twitter by searching 10000 records of the keyword “European Union”. Then we store the results in corpus.txt with json format and how one tweet is considered as a record.

To validate whether the results obtained is the expected data, we print the results according to the meta information required.

### Q1.2 What kinds of information users might like to retrieve from your crawled corpus (i.e. applications), with example queries

[+]

* Authors of the tweets
  + e.g. tweets by XXX
* Creation date of the tweets
  + e.g. tweets on DD/MM/YYYY
* Tags of the tweets
  + e.g. #ABC
* Content of the tweets
  + e.g. debt crisis
* Favourite counts the tweet has
  + e.g. trending tweets

### Q1.3 The numbers of records, words, and types (i.e. unique words) in the corpus

[+]

* Number of records: = 17021
* Number of words: >= 340754 words 2749611 characters
* Types of words: e.g. unique words, common words, stop words[[3]](#footnote-3)
  + Unique words: 853 words 6262 characters
  + Common words:
    - 13,834 european
    - 11,884 union
    - 11,055 the
    - 5,559 to
    - 4,129 rt
    - 4,085 eu
    - 2,505 a
    - 2,043 http
    - 1,454 t.co
    - 1,395 with
  + Stop words: 105 words, 447 characters

# 2 Indexing and Querying

## Question 2

### Q2.1 Build a simple Web interface for the search engine (e.g. Google)

[+]

* Search bar
* Search button
* Query
  + e.g. Author:XXX Date:DD/MM/YYYY
* Results
  + Pagination

### Q2.2 A simple UI for crawling and incremental indexing of new data would be a bonus (but not compulsory)

[+]

### Q2.3 Write five queries, get their results, and measure the speed of the querying

[+]

* tweets by XXX
  + results + speed
* tweets on DD/MM/YYYY
  + results + speed
* hashtag #ABC
  + results + speed
* content debt crisis
  + results + speed
* favourite tweets
  + results + speed

## Question 3: Explore some innovations for enhancing the indexing and ranking. Explain why they are important to solve specific problems, illustrated with examples. Possible innovations include (but are not limited to) the following:

* Interactive search (e.g. refine search results based on users’ relevance feedback)
* Improve search results by integrating machine learning or data mining techniques (e.g. classification or cluster techniques)
* Go beyond text-based search (e.g. implement image retrieval or multimedia retrieval)
* Exploit geo-spatial data (i.e. map information) to refine query results/improve presentation/visualization
* Others (Brainstorm with your group members)
* Interactive search (e.g. refine search results based on users’ relevance feedback)
* Personalized vs. General search results
* e.g. Google account search results vs. General Google search results
* Improve search results by inegrating machine learning or data mining techniques (e.g. classification or cluster techniques)
* Classification by topic
* e.g. debt, war
* Go beyond text-based search (e.g. implement image retrieval or multimedia retrieval)
* Exploit geo-spatial data (i.e. map information) to refine query results/improve presentation/visualization
* Rank tweets with same geo-location as query location higher
* e.g. localized trending topics
* Others
* Qn 3: (Brainstorm~)
* Ranking: Time, Location, Fav count
* Indexing: ???
* Machine Learning: ???
* Interactive Search: ???
* Others: ???
* Innovations
* Include chronological order with latest on top
* Include location with higher ranking for nearest location
* Search image based on hashtags
* Categorization
* Popularity based on favourite counts and retweet counts

# 3 Classification

Define a set of categories (minimum three) the collected data could belong to and perform automatic classification on them (e.g. auto-categorization into specific topics, sentiment analysis):

* Knowledge based e.g. semantic networks and ontologies
* Rule based e.g. linguistic patterns and POS tagging
* Machine learning based e.g. SVM and ANN
* Hybrid
* Categories
* Economic
* Technology
* Politics
* Social
* Auto-categorization Techniques
* SVM
* Naive Bayes
* Weka

## Question 4

### Q4.1 Motivate the choice of your classification approach in relation with the state of the art

[+]

* Reason for choice
* Refer to remarks from Prof

### Q4.2 Discuss whether you had to preprocess data and why

[+]

* Preprocess
* Tweet remove RT
* Tweet remove URL

### Q4.3 Build an evaluation dataset by manually labeling 10% of the collected data (at least 1,000 records) with an inter-annotator agreement of at least 80%

[+]

* Precision
* Recall
* F-measure
* Refer to remarks from Prof

### Q4.4 Provide evaluation metrics such as precision, recall, and F-measure and discuss results

[+]

* Metrics

### Q4.5 Discuss performance metrics e.g. records classified per second and scalability of the system

[+]

### Q4.6 A simple UI for visualizing classified data would be a bonus (but not compulsory)

[+]

## Question 5: Explore some innovations for enhancing classification. Explain why they are important to solve specific problems, illustrated with examples:

* Ensemble classification (e.g. leverage on multiple classification approaches)
* Cognitive classification (e.g. use brain-inspired algorithms)
* Multi-faceted classification (e.g. take into account multiple aspects of data)

# Future Plans

As Twitter allows images to be posted, we are looking into capturing images and storing them into the index to allow search results to include images. We may also be looking into storing data according to location to allow users to search for location-based information.

As for classification, we can classify information according to location or hashtags/favourite tweets or seriousness of the information based on key words found in them. More topics for classification can be explored and implemented.

1. <https://about.twitter.com/company> [↑](#footnote-ref-1)
2. <https://dev.twitter.com/overview/api> [↑](#footnote-ref-2)
3. <http://design215.com/toolbox/wordlist.php> and <http://www.wordcounter.net/> [↑](#footnote-ref-3)