AIT 664: Information: Representation, Processing, and Visualization
Individual Class Project, Fall 2018

NOTES:

1. ***Goal***: Hands-on experience to process data, to extract information, and discover patterns or knowledge using data mining method
2. ***Scenario***: You are an information professional, and you analyze and monitor social media data streams. You want to identify the information and patterns that are helpful to understand the situation and facilitate decision making.
    a. Key task: You want to organize information stored in several documents, to identify patterns relevant to a better understanding of the data. So, use a data mining system of clustering, using Weka tool, to glean patterns of the similar information documents, followed by providing a visual analysis of your discovered patterns, using a tool of your choice (e.g., Tableau).
3. ***Policies***:
    a. Grading: Please follow the steps below for this project. Each milestone will be graded for your final project grade (25 points).
    b. Language and Tools: for data preprocessing, you can use any language or tools, such as Python. Similarly, tools for visualization can be based on your choice (e.g., Tableau), or you can also generate visualizations by writing codes in your choice of programming language. However, you will need to demonstrate the outputs of each step during the demo.
    c. Deliverable: A report with snapshots of each milestone output and demo (doable in class or by appointment) are due **by the Final Exam day**.

MILESTONES:

1. Data Acquisition
    a. Programmatically download the project data file by choosing one from the following to focus on:
        i. Hurricane Disaster Management.
        ii. Ebola Discussion.
    b. Each of them is a CSV format file, and in a table format.
    c. The format is something like:

MESSAGE	CREATED_AT
@Zuora wants to help @Network4Good with Hurricane Relief. Text SANDY to 80888 & donate $10 to @redcross @AmeriCares & @SalvationArmyUS #help	2012-10-30 22:15:41

        i. MESSAGE: message content from social media
            1. String
        ii. CREATED_AT: date and time of message arrival
            1. Date-Time
2. Data Preprocessing

    a. For each field, extract the substring that you want to preserve for analysis, e.g., I can discard time part in the DATETIME or DATE field to cluster documents by Dates; similarly, words that may not convey meaningful information and patterns in the MESSAGE or TEXT field; due to the noisiness of tweet data, optional operations could also be done if they can help improve your performance, such as stop word removal, tokenization, stemming, and name entity recognition. Likewise, impute the missing values in the fields using an assumption for function (e.g., average, median).
    b. Store the output in a CSV (or JSON) file
3. Mining Tool Preparation
    a. Load Weka GUI tool (see tutorial here: https://www.youtube.com/watch?v=TtBgfXmIDHQ),
    b. Install and go into 'Explorer' menu
    c. Open your output file of Milestone-2, and remove the fields/attributes that you think are not meaningful for pattern analysis, such as document's id.
    d. Choose a 'Filter' to apply StringToWordVector, for transforming MESSAGE string into a vector of words, which become part of attribute set (see example here, in 5th minute: https://www.youtube.com/watch?v=jSZ9jQy1sfE)
4. Clustering Analysis
    a. Choose 'Clustering' menu tab in Weka and apply a clustering algorithm of your choice, with minimum clusters between 3 to 5.
    b. Save the results of the clustering: cluster number is associated with each document/message in the result output.
5. Visualization
    a. Use the outputs of Clustering process, to visualize the patterns across document clusters, such as a word cloud of cluster words (e.g., check Tableau OR refer to http://www.wordle.net/create), and write your interpretations of observations in the DIKW framework, especially information and discovered knowledge.
    b. Find top-10 representative words for each cluster (e.g., by TF-IDF). What do you infer for a topic of the cluster?

OTHER RESOURCES could be used:

Python tutorials:

1. https://www.dataquest.io/mission/1/python-basics
2. https://www.python.org/about/gettingstarted/

NLP tools

3. NLTK (Python):
    a. • Natural language processing with Python. Bird, S. et al. 2009.
    b. • NLTK homepage: http://www.nltk.org/
4. Stanford NLP (Java): http://nlp.stanford.edu/software/
5. RM (R): https://cran.r-project.org/web/packages/tm/index.html

Geocoders for tweet messages

6. Carmen: https://github.com/mapbox/carmen

Text classifier for short messages