Data Mining: Assignment 2

Lucia-Eve Berger (lucia.berger@mail.mcgill.ca)
Giorgi Gabunia (ggabunia@unibz.it)
European Masters in Software Engineering

November 9, 2017

Question (1). how did you construct the attributes.

Solution. The attributes were constructed after cleaning the data. The csv was passed through a pandas (python) data-frame. In the data-frame, the text-content was isolated (tweet-column).

- When isolated,
  - text was transformed to lowercase
  - punctuation was removed
  - stop words were removed
  - emoji characters, links and usernames were removed
  - stemming was removed (loves vs love)

```
all_tweets[’tweet’] = all_tweets[’tweet’].str.lower().str.replace(’[^\w\s]’, ’’).str.split()
df[’tweet’] = df[’tweet’].apply(lambda x: [item for item in x if item not in get_stop_words()])
```

□

Question (2) the set of class labels and the strategy you followed to extract them,.

Solution. After processing the data, a count was made on each individual word. A collection of the most frequent words was made following the top-k frequent attributes. The highest frequency words were then transformed into attributes and a *yes no* matrix was made in the ARFF file format.

```
totals = collections.Counter(i for i in list(itertools.chain.from_iterable(df[’tweet’])))
```

100 top attributes were selected. The ARFF file was processed by the FPGrowth Algorithm in Weka at 0.05 min support value. Experimentation was made with lower min support bound (0.001), however, it resulted in more frequent, but weaker rules. The selected value resulted in 203 association rules. From the association rules, exta-domain knowledge was applied to remove some rules. For example, standalone brands and joint names were removed if not accompanied by another word.

Based on the weka generated association rules, a data clean was performed on the associations. The associations were lists of two elements and duplicates were removed.

For example : $['time','warner']wascountedthesameas['warner','time']$

```
{”className”: ”night_at_the_museum”,
”includedWords”:
[”museum”, ”night”]
},
{”className”: ”google_io”,
”includedWords”:
[”google”, ”io”]
```

□

Question (3) the assessment of the quality of the classes..

Solution. The quality of the classes were determined initially by length of the the includedWords array. Each had to be at least size 2. They were then processed by an occurrence counter in python returning there occurrence together and the returned tweetIDS. An example of the nightAtMusuem class is shown below. Then finally a comparison was made between the words contained and the tweet tag as a sanity check.

□