Episode Mining
Autonomous work

Julien Blanchard

December 2020

Sacha Chua [1] logs and publishes all her daily activities on the internet. The Lifelog dataset on the extradoc contains Sacha's life events over 2019. The project goal is to extract frequent episodes from this sequence and assess them. To generate episodes, you will use the Python notebook `tutorial.ipynb`, directly adapted from the code proposed in the tutorial *Analyzing Sequential User Behavior on the Web* [2] from the conference World Wide Web Conference 2016.

The Algorithm PrefixSpan implemented in the notebook is an algorithm for mining sequential patterns from a set of sequences. In order to mine episodes from a single sequence S, you first have to transfom S into a set of subsequences $\{S_i\}_{i\in[1;k]}$. The usual approach is to apply a sliding window of fixed size on the sequence S, which is called windowization. In the end, applying PrefixSpan on $\{S_i\}_{i\in[1;k]}$ comes down to applying Winepi on S.

— Part A.

1. Write a function performing the *windowization* of the Lifelog sequence. The window size is the parameter max_{span}, that you will set to 1 or 2 hours to begin your project in order to limit the number of patterns (you may change it later).
2. Mine the episodes by applying `prefixSpan()` on the set of windows. The frequency in the output is expressed as a number of windows. Use the frequency threshold min_{freq} of your choice (begin with high values !).
3. What are the 5 most frequent patterns of length 2 ? of length 3 ?
4. Write a function that turns the set of episodes into a set of temporal rules, by isolating the last event of each episode. For a rule $X \to Y$, get the cardinalities n_X, n_Y and $n_{X,Y}$ from the `prefixSpan()` output, and compute the measures frequency (in %), confidence, recall, lift and j-measure. Store them in a dictionary or a pandas dataframe.
5. According to the lift, what are the best rules of length 2 ? of length 3 ?
6. Repeat the steps 1 to 5 in order to discover rules that you find interesting. You can supervise your search by using the parameters min_{freq} and max_{span}. You may also focus the search on specific time periods or specific event types that you are interested in.

— Part B.

1. Produce a large set of patterns from the dataset (at least 2000 patterns).
2. Draw the scatterplot matrix of the five measures over the whole set of rules extracted with the function of the question A.4 above. Compute the correlations between measures and comment.

1. http ://quantifiedawesome.com/
2. http ://sequenceanalysis.github.io/

3. Apply a reduction dimension technique over the set of rules described by the measures, then visualize the rules in the 2D embedded space. Comment.

Your work has to be uploaded on the Extradoc before January 15th. Your notebook must include :

— your answers to the questions,

— your comments,

— the source code you used to answer the questions.