**Recommendations for Magellan**

Qing Li, Fan Ding

## 1. Installation

When we first tried to import Magellan on a Windows machine after installing both Anaconda and Magellan, we were facing an issue with some file missing. The issue description is as follows,

*Issue Description:*

ImportError: Building module magellan.cython.test_functions failed: ['Dist utilsPlatformError: Unable to find vcvarsall.bat\n']

Then we tried to install it on Windows 10 (A clean machine only installed the latest version of Anaconda as recommended in the user manual). We still got the same error. We thought there might be some dependent files need to be installed. (We did not test whether the latest version of Magellan will still have this issue.)

We haven't figured out which files should be installed. We tried to use Ubuntu instead of Windows in the end.

## 2. Blocking

### *2.1 Intersect Result Sets*

We think if it is possible, in the future, to provide the ability to intersect two blocking sets, not only to union them, would be very helpful.

### *2.2 Blocking based on Attribute Values*

Currently, if a blocking method is applied, this method will be applied to all tuples. However in our cases, we are looking for a Blocker which can be applied on some certain 'type' of tuples. For example, there are lots of tuples in our dataset are accessories. We only want a blocker to block only accessories with accessories, which is hard to implement under the current version of Magellan.

### *2.3 Provide More Information in Result Table*

Currently, in the result table, only similarity of two tuples is provided. The similarity score is very important, however, when to debug blockers, it is more helpful to provide some general statistic information, such as the distribution of similarity score over all tuples, instead of only similarity score on an individual level.

## 3. Matching

### *3.1 User Defined Features on Numerical Attributes*

We didn't find a good way to implement a user-defined similarity feature on numerical attributes. In our case, we want to define a feature to specify the price difference range of two matched tuples. However, it seems only strings can be the input of new features in current Magellan.

*3.2 Issues with SVM and Logistic Regression*

In our final result, either SVM or Logistic Regression is reporting a reasonable result. Most times, results only contain 0 in both precision and recall. We tried to figure out the reason. It seems when the number of input features is less than three, those two models will not work.