**Note:**

Default: means it follow the original code implementation without lowercasing terms for calculate metrics like accuracy. 833 terms at most get involved in evaluation.

Revised: lowercasing all terms, 916 terms at most.

This is because the code try to interest the raw data(the csv file containing uppercase letter) with the terms in corpus.vocabulary which is composed of terms from core ontology and lowercased csv file. I guess the author aims to eliminate the terms in core ontology but not present in csv file. It is weird to do so because we can directly use terms from lowercased raw csv file.

# default

## original

## revised

The overall result from revised code slips a bit. But the highest f1 score is very close though still less than the alleged result.

# SVM classifier

**For any classification method, we use the pseudo labeled terms which is terms from core ontology as the training dataset, the training dataset size is 38 terms. Besides, we only consider the performance of the first iteration.**

## original

The result very close to the result in the Article

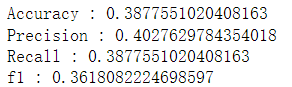
## revised

the precision even more than that of our pipeline, but get a less f1 score. For Classification the result is common. As our pipeline is clustering-based, the performance is more acceptable.

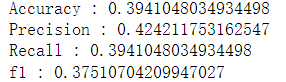
### ANN classifier



## original

The same to the result in article.

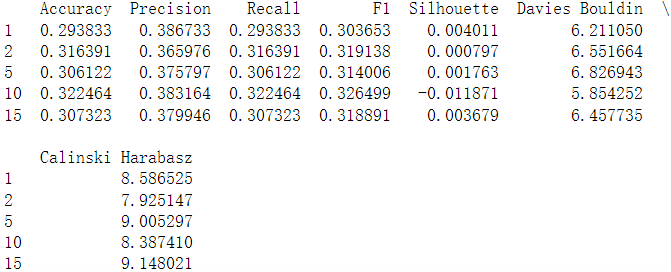
## revised



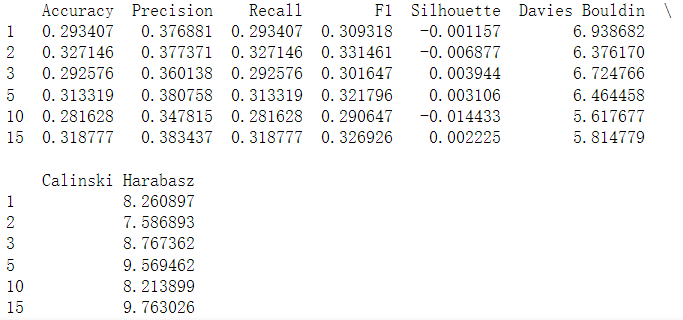
# PCA + CDBSCAN



## original

The result of **5th** iteration is quite close to that in the article

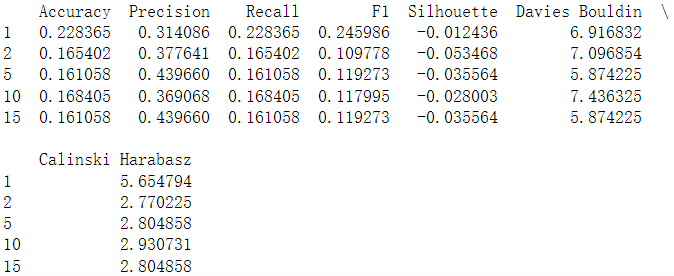
## revised



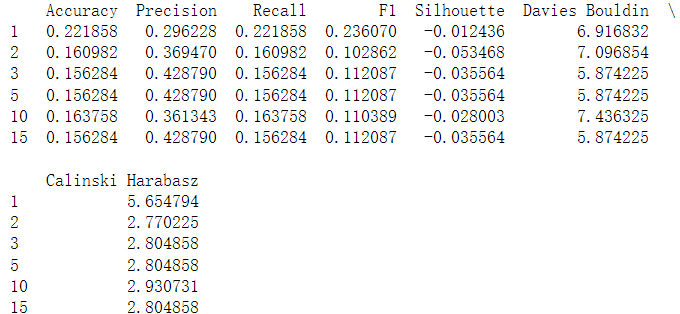
# None reduction + CDBSCAN



## original

the 5th iteration is the result in the article.

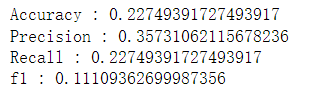
## revised



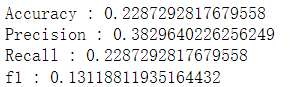
# UMAP +DBSCAN

Actually, for UMAP, the measurement of accuracy, recall, precision and f1 is **meaningless.** Because the index of cluster in DBSCAN does not correspond to any of a single core concept. It is just an index, a number. Whereas in CDBSCAN each cluster index refers to a core concept index.

## original



## revised

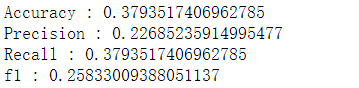


# UMAP +SVM(bagging)

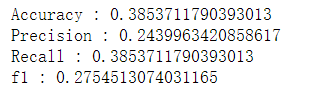


## Original

Same



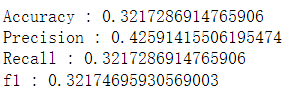
## revised



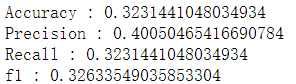
# UMAP +ANN



## original

 Close

## revised

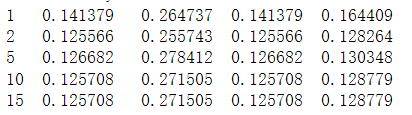


# Rebalance dataset



## original

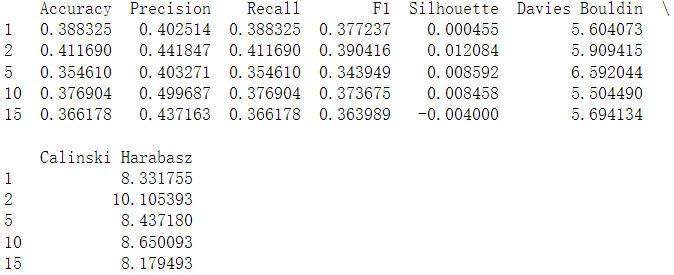
## revised

I deleted 22 terms with a sub-core concept prefix “Any” or “Anything” because they not in the core ontology.

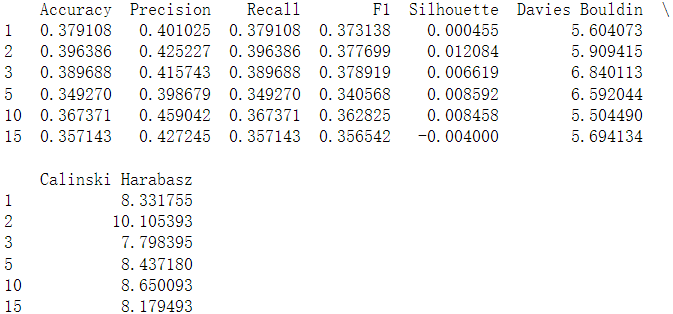
# Changing terms

## without Data (57 terms)

### original

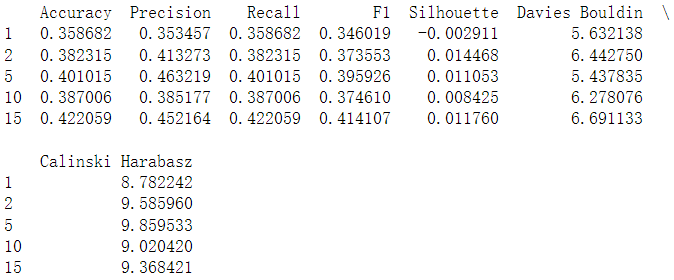


### revised

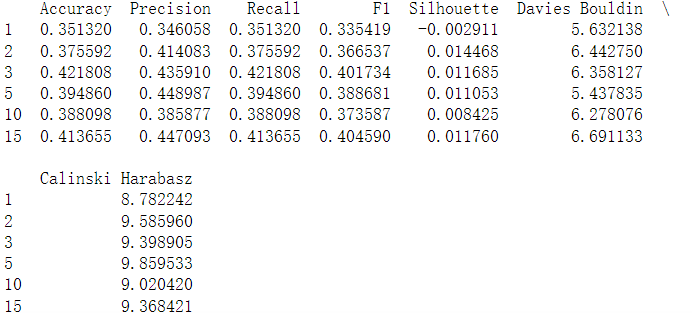


## Without Market(39 terms)

### Original

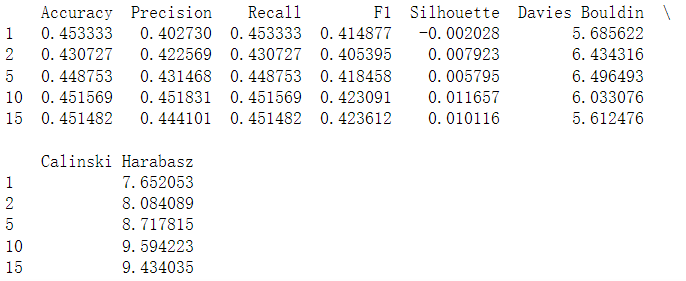


### revised

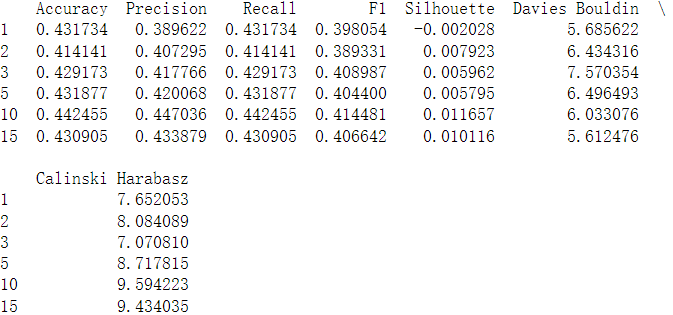


## Without Data and Market

### Original

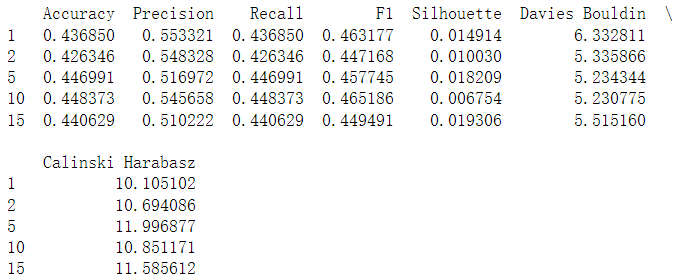


### revised



## Without Other(126 terms)

### Original



### revised

