Выделение тем из коллекции текстов с помощью сингулярного разложения матрицы "документ-слово"

Герасименко Н. А.

В ходе работы составлена матрица «документ-слово» и произведено ее разложение методом SVD на матрицы «документ-тема» и «тема-слово». Результат проинтерпретирован.

1 Полученная матрица «документ-слово»

В качестве коллекции документов был взят набор из произведений А. Барто «Мячик», «Бычок», «Зайка», первые две строки каждого произведения. Поскольку в этих документах нет совпадающих слов, предполагается, что темы будут успешно выделены и будут иметь очевидную интерепретацию принадлежности к этим произведениям. В качетстве метрики частоты вхождения слова в документ использована tf-idf. Получена следующая матрица «документ-слово»:

```
[[ 0.000  0.354  0.354  0.354  0.000  0.000  0.354  0.000  0.354  0.000
   0.000  0.354  0.354  0.000  0.000  0.000  0.354  0.000  0.000  0.000]
 [ 0.408  0.000  0.000  0.000  0.408  0.000  0.000  0.408  0.000  0.000
   0.408  0.000  0.000  0.408  0.408  0.000  0.000  0.000  0.000  0.000]
 [ 0.000  0.000  0.000  0.000  0.000  0.333  0.000  0.000  0.000  0.333
   0.000  0.000  0.000  0.000  0.000  0.667  0.000  0.333  0.333  0.333]]
```

Действительно, каждое слово встречается только одном из текстов.

2 Полученные матрицы «документ-тема» и «тема-слово»

Полученная матрица «документ-тема», как мы видим, подтверждает наше предположение относительно распределения по темам:

```
[[ 0.000  1.000  0.000]
 [ 0.000  0.000  1.000]
 [ 1.000  0.000  0.000]]
```

Каждый документ однозначно отнесен к одной из тем. Теми же соображениями объясняется то, что каждое слово отнесено только к одной теме, это видно по матрице «тема-слово»:

```
[[ 0.000  0.000  0.000  0.000  0.000  0.333  0.000  0.000  0.000  0.333
   0.000  0.000  0.000  0.000  0.000  0.667  0.000  0.333  0.333  0.333]
 [ 0.000  0.354  0.354  0.354  0.000  0.000  0.354  0.000  0.354  0.000
   0.000  0.354  0.354  0.000  0.000  0.000  0.354  0.000  0.000  0.000]
 [ 0.408  0.000  0.000  0.000  0.408  0.000  0.000  0.408  0.000  0.000
   0.408  0.000  0.000  0.408  0.408  0.000  0.000  0.000  0.000  0.000]]
```