Национальный исследовательский университет ИТМО
Факультет программной инженерии и компьютерной техники
Кафедра вычислительной техники

Отчет о прохождении дисциплины
“Интеллектуальный анализ данных”

Работу выполнил:
студент группы P42111
Губарев В.Ю.
Преподаватель:
Бессмертный И.А.

2019, г. Санкт-Петербург

Содержание

1. Введение

В рамках настоящей работы необходимо выполнить следующие этапы:

1. Выбрать предметную область, в выбранной предметной области найти 8 документов размером ≥ 50 килобайт на различные темы.
2. Выполнить лемматизацию каждого документа.
3. Извлечь термины предметной области, используя частотный анализ.
4. Извлечь термины предметной области, используя меру tf-idf.
5. Построить HAL-матрицу для каждого документа, определить наиболее связанные слова.

2. Выбор предметной области

Для выполнения работы интеллектуальный анализ данных был выбран как предметная область. Книга "Morgan Kaufmann et al. Data Mining: Concepts and Techniques, Third Edition." была преобразована из формата *pdf* в простой текстовый формат. Текстовый файл в ручном режиме разделен на отдельные документы, соответствующие главам книги (см. Листинг 1).

Листинг 1: Список файлов после разделения книги на главы

```
vladimirg@sirius:~/wrk/6grade/analysis/data/preprocessed$ ls -lh
total 1.6M
-rw-r--r-- 1 izoomko izoomko 100K Oct 3 22:40 chapter1
-rw-r--r-- 1 izoomko izoomko 137K Oct 3 22:42 chapter10
-rw-r--r-- 1 izoomko izoomko 107K Oct 3 22:42 chapter11
-rw-r--r-- 1 izoomko izoomko 105K Oct 3 22:43 chapter12
-rw-r--r-- 1 izoomko izoomko 116K Oct 3 22:45 chapter13
-rw-r--r-- 1 izoomko izoomko 100K Oct 3 22:40 chapter2
-rw-r--r-- 1 izoomko izoomko 109K Oct 3 22:40 chapter3
-rw-r--r-- 1 izoomko izoomko 160K Oct 3 22:40 chapter4
-rw-r--r-- 1 izoomko izoomko 148K Oct 3 22:41 chapter5
-rw-r--r-- 1 izoomko izoomko 95K Oct 3 22:41 chapter6
-rw-r--r-- 1 izoomko izoomko 129K Oct 3 22:41 chapter7
-rw-r--r-- 1 izoomko izoomko 169K Oct 3 22:42 chapter8
-rw-r--r-- 1 izoomko izoomko 131K Oct 3 22:42 chapter9
```

3. Лемматизация

Для выполнения первого этапа был написан скрипт (см. Листинг 2) на языке *Python*. Для работы скрипта необходим пакет `nltk`. При первом запуске Python предложит установить оставшиеся зависимости из пакета `nltk`: tokenize, corpus, stem.

```
pip3 install -U nltk
```

Листинг 2: Исходный код лемматизатора

```
#!/usr/bin/python
```

```
import math
import operator
import os
import sys
import string

from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()

import collections

printable = set(list("abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWXYZ␣"))

if len(sys.argv) < 2 or not os.path.isdir(sys.argv[1]):
    print ("./lemmatize.py␣<directory>")
    sys.exit()

texts_tokens = []
tf_texts = []

os.chdir(sys.argv[1])

corpus = []

for file in os.listdir('.'):
    if not os.path.isfile(file):
        continue

    f=open(file, "r")

    text=filter(lambda x: x in printable, f.read())

    text = text.lower()

    tokens = word_tokenize(text)
    texts_tokens.append(tokens)

    stop_words = set(stopwords.words('english'))
    filtered_tokens = [w for w in tokens if not w in stop_words]
    filtered_tokens = [w for w in filtered_tokens if len(w) > 2]
    filtered_tokens = [lemmatizer.lemmatize(w) for w in filtered_tokens]

    corpus.append(filtered_tokens)

for tokens in corpus:
    for word in tokens:
```

```
        print word
```

Лемматизация – это преобразование слова в лемму, то есть в его словарную форму. Для этого выполнить фильтрацию текста (убрать все, что не является словами) и токенизацию по терминальным символам.

В Листинге 3 приведены последние 10 строк вывода лемматизатора 2.

Листинг 3: Первые строки вывода лемматизатора

```
vladimirg@sirius:~/wrk/6grade/analysis$ python ./src/lemmatize.py data/
    ↪ preprocessed | tail
data
section
data
mining
concept
techniquesc
elsevier
inc
right
reserved
```

4. Частотный анализ

Попробуем извлечь термины предметной области на основе частоты встречаемости слов в тексте (см. Листинг 4).

Листинг 4: Исходный код частотного анализатора

```
#!/usr/bin/python

import math
import operator
import os
import sys
import string

from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()

import collections

printable = set(list("abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWXYZ␣"))

if len(sys.argv) < 2 or not os.path.isdir(sys.argv[1]):
    print ("./freq.py␣<directory>")
    sys.exit()

texts_tokens = []
```

```
tf_texts = []

os.chdir(sys.argv[1])

corpus = {}

for file in os.listdir('.'):
    if not os.path.isfile(file):
        continue

    f=open(file, "r")

    text=filter(lambda x: x in printable, f.read())

    text = text.lower()

    tokens = word_tokenize(text)
    texts_tokens.append(tokens)

    stop_words = set(stopwords.words('english'))
    filtered_tokens = [w for w in tokens if not w in stop_words]
    filtered_tokens = [w for w in filtered_tokens if len(w) > 2]
    filtered_tokens = [lemmatizer.lemmatize(w) for w in filtered_tokens]

    corpus[file] = filtered_tokens

map = {}

for file in corpus.keys():
    for word in corpus[file]:
        if file not in map.keys():
            map[file] = {}

        if word not in map[file].keys():
            map[file][word] = 1
        else:
            map[file][word] = map[file][word] + 1

for file in corpus.keys():
    print ("==================")
    print (file)
    print ("==================")
    print
    print (sorted(map[file].items(), key=operator.itemgetter(1), reverse=True)
        ↪ [:30])
```

Рассмотрим 30 наиболее распространенных слов в каждом документе (см. Листинг 4).

```
==================
chapter9
==================
```

```
[('classification', 175), ('data', 167), (u'network', 132), ('rule', 123), ('
   ↪ set', 118), ('class', 110), ('training', 101), ('learning', 99), ('
   ↪ tuples', 88), (u'value', 81), ('classifier', 78), (u'unit', 75), ('input
   ↪ ', 73), ('frequent', 72), ('tuple', 71), ('used', 69), (u'pattern', 69),
   ↪  (u'method', 62), ('given', 61), (u'algorithm', 59), ('one', 56), ('
   ↪ output', 55), ('support', 54), (u'attribute', 54), (u'weight', 54), ('
   ↪ may', 53), ('number', 52), ('label', 51), ('layer', 46), ('vector', 44)]
==================
chapter8
==================

[('tuples', 272), ('class', 261), (u'tree', 236), (u'rule', 201), ('data',
   ↪ 185), (u'attribute', 182), ('classifier', 176), ('decision', 171), ('set
   ↪ ', 153), ('classification', 139), ('training', 119), ('tuple', 102), ('
   ↪ accuracy', 98), (u'value', 94), ('model', 87), ('test', 86), (u'positive
   ↪ ', 85), ('measure', 82), ('given', 80), ('may', 73), ('node', 71), ('
   ↪ number', 70), ('used', 67), ('section', 66), ('computer', 66), ('one',
   ↪ 65), (u'example', 63), (u'method', 60), ('probability', 59), ('negative
   ↪ ', 56)]
==================
chapter5
==================

[('cube', 479), ('data', 349), ('cell', 211), (u'dimension', 186), ('cuboid',
   ↪ 182), ('computation', 143), ('query', 129), (u'value', 123), ('iceberg',
   ↪  94), ('example', 74), ('measure', 72), (u'aggregate', 67), (u'fragment
   ↪ ', 66), ('count', 66), ('compute', 64), (u'method', 57), ('base', 56), (
   ↪ u'chunk', 56), ('multidimensional', 54), ('space', 52), ('sample', 50),
   ↪ ('set', 50), ('computed', 50), ('used', 48), ('one', 47), ('full', 46),
   ↪ ('shell', 46), ('ranking', 43), (u'algorithm', 42), ('partition', 41)]
==================
chapter4
==================

[('data', 804), (u'warehouse', 213), ('cube', 161), (u'dimension', 140), ('
   ↪ attribute', 127), ('olap', 112), ('table', 100), ('cuboid', 98), ('
   ↪ mining', 95), ('may', 86), (u'query', 85), (u'system', 83), (u'database
   ↪ ', 78), ('item', 76), (u'sale', 75), ('concept', 71), (u'level', 67), ('
   ↪ generalization', 66), ('set', 66), ('processing', 65), ('value', 65), ('
   ↪ multidimensional', 64), (u'example', 61), ('relation', 61), ('online',
   ↪ 59), ('schema', 59), (u'operation', 58), (u'hierarchy', 57), ('
   ↪ generalized', 57), ('fact', 57)]
==================
chapter7
==================

[('pattern', 619), ('mining', 229), ('data', 199), ('frequent', 154), (u'rule
   ↪ ', 152), ('set', 121), (u'constraint', 112), ('association', 90), (u'
   ↪ item', 84), ('support', 83), (u'method', 68), ('may', 54), ('transaction
```

```
   ↪ ', 54), ('used', 49), (u'section', 46), ('example', 44), ('space', 43),
   ↪ (u'level', 41), (u'measure', 41), ('number', 40), ('context', 40), ('
   ↪ also', 39), (u'attribute', 38), ('cluster', 38), ('information', 37), ('
   ↪ itemset', 36), ('colossal', 36), ('itemsets', 35), ('quantitative', 35),
   ↪  (u'concept', 35)]
==================
chapter6
==================

[('frequent', 245), ('itemsets', 152), ('mining', 133), ('itemset', 127), (u'
   ↪ pattern', 127), (u'set', 114), ('support', 101), (u'association', 93), (
   ↪ u'item', 91), ('transaction', 91), ('data', 89), (u'rule', 84), (u'
   ↪ measure', 66), ('candidate', 54), (u'database', 52), ('minimum', 51), ('
   ↪ count', 50), ('algorithm', 49), ('confidence', 49), ('apriori', 48), ('
   ↪ example', 43), ('table', 43), (u'subset', 42), (u'method', 40), ('two',
   ↪ 40), ('closed', 39), ('number', 33), (u'correlation', 30), ('one', 30),
   ↪ ('using', 28)]
==================
chapter1
==================

[('data', 621), ('mining', 283), (u'pattern', 104), ('database', 94), ('
   ↪ knowledge', 79), (u'system', 69), ('information', 67), ('analysis', 61),
   ↪  ('may', 57), ('example', 53), ('user', 51), (u'model', 50), (u'set',
   ↪ 46), ('customer', 42), ('learning', 40), (u'query', 40), ('search', 39),
   ↪  (u'kind', 37), (u'class', 37), ('many', 36), ('discovery', 35), ('web',
   ↪  35), ('used', 34), ('conference', 33), ('interesting', 33), ('process',
   ↪  33), (u'transaction', 33), ('classification', 32), (u'application', 32)
   ↪ , ('warehouse', 31)]
==================
chapter3
==================

[('data', 562), (u'attribute', 234), ('value', 173), ('may', 98), ('hierarchy
   ↪ ', 62), ('example', 61), (u'method', 60), ('concept', 58), ('section',
   ↪ 58), ('set', 58), ('also', 51), ('number', 49), ('analysis', 46), ('
   ↪ reduction', 45), ('mining', 44), (u'technique', 42), ('given', 40), ('
   ↪ use', 38), ('used', 37), ('transformation', 35), ('normalization', 34),
   ↪ ('discretization', 34), (u'bin', 33), (u'level', 32), ('missing', 32),
   ↪ ('cleaning', 31), ('two', 31), (u'cluster', 31), ('database', 30), (u'
   ↪ histogram', 30)]
==================
chapter2
==================

[('data', 359), (u'attribute', 246), (u'value', 172), (u'object', 114), ('set
   ↪ ', 71), ('two', 69), (u'measure', 68), ('dissimilarity', 56), (u'plot',
   ↪ 51), ('visualization', 50), ('numeric', 50), (u'example', 49), (u'number
   ↪ ', 47), ('distance', 45), (u'mean', 44), ('median', 44), ('may', 43), ('
   ↪ binary', 39), ('similarity', 38), ('section', 38), ('also', 38), ('used
```

```
    ↪ ', 36), ('figure', 34), (u'distribution', 33), (u'technique', 32), ('
    ↪ nominal', 32), ('dimension', 32), ('matrix', 29), (u'type', 29), ('
    ↪ ordinal', 28)]
==================
chapter11
==================

[('cluster', 297), ('clustering', 178), ('data', 112), ('graph', 104), ('set',
    ↪  100), ('object', 96), (u'constraint', 96), ('method', 83), (u'vertex',
    ↪ 81), ('two', 73), (u'customer', 62), ('example', 59), ('similarity', 57)
    ↪ , ('algorithm', 57), ('may', 56), ('distance', 51), ('using', 49), ('
    ↪ gene', 48), ('one', 46), ('analysis', 45), ('section', 43), ('matrix',
    ↪ 43), ('product', 40), ('fuzzy', 39), ('cut', 37), ('network', 36), ('
    ↪ search', 36), ('bicluster', 35), ('probabilistic', 35), ('
    ↪ highdimensional', 34)]
==================
chapter10
==================

[('cluster', 447), ('clustering', 349), (u'object', 313), ('data', 246), (u'
    ↪ method', 195), (u'set', 139), ('algorithm', 95), ('hierarchical', 93),
    ↪ ('may', 82), ('number', 80), ('analysis', 65), ('distance', 57), (u'
    ↪ point', 55), ('two', 54), ('partitioning', 54), ('quality', 52), (u'cell
    ↪ ', 51), ('density', 51), (u'example', 48), ('kmeans', 46), ('given', 43)
    ↪ , ('one', 42), ('using', 42), (u'value', 41), ('used', 40), (u'measure',
    ↪  40), ('space', 39), (u'group', 38), ('find', 38), ('neighborhood', 35)]
==================
chapter13
==================

[('data', 582), ('mining', 370), (u'network', 103), ('system', 96), (u'pattern
    ↪ ', 95), ('analysis', 92), ('information', 83), (u'sequence', 81), ('
    ↪ customer', 73), (u'method', 70), ('web', 58), ('may', 57), ('many', 57),
    ↪  ('research', 49), ('user', 48), (u'example', 46), (u'trend', 46), ('set
    ↪ ', 45), ('model', 42), ('used', 39), ('search', 39), (u'application',
    ↪ 37), (u'graph', 34), (u'technique', 32), ('biological', 31), ('
    ↪ classification', 30), (u'help', 30), (u'item', 30), ('also', 30), ('
    ↪ include', 30)]
==================
chapter12
==================

[('outlier', 549), (u'object', 290), ('data', 245), ('detection', 207), (u'
    ↪ method', 158), (u'cluster', 99), ('normal', 98), ('set', 84), ('model',
    ↪ 81), ('may', 79), ('point', 65), ('contextual', 57), ('example', 56), (u
    ↪ 'attribute', 47), ('distance', 44), ('using', 42), ('used', 41), ('
    ↪ distribution', 40), ('density', 39), ('two', 38), ('many', 36), ('
    ↪ collective', 36), ('detect', 35), ('small', 35), ('number', 35), ('one',
    ↪  34), ('context', 34), ('statistical', 33), ('application', 33), ('cell
    ↪ ', 31)]
```

Данный метод анализа дает некоторое представление о содержании документа и используемой терминологии. Например, в 12 главе наиболее частые слова это “outlier“, “object”, “data”, “detection“. Можно предположить, что глава 12 книги о поиске аномалий (или выбросов, англ. *outliers*) в данных.

В то же время, заметны термины вроде “data mining“, как в Главе 1. Или “distribution density” в Главе 12.

В список наиболее часто употребляемых слов так же попали и общие слова вроде “example”, “using”.

5. TF-IDF анализ

Мера TF-IDF считается как произведение важности слова в документе против частоты его встречаемости в домене. То есть, позволяет оценить уникальность слова в документе относительно всего домена.

TF(*TermFrequency*) – отношение числа вхождений некоторого слова к общему числу слов документа. Таким образом, оценивается важность слова в пределах отдельного документа.

IDF(*Inversedocumentfrequency*) – инверсия частоты, с которой некоторое слово встречается в документах коллекции. Она измеряет непосредственно важность термина. Подсчет TF-IDF выполняется скриптом в Листинге 5.

Большой вес в TF-IDF получают слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах. Поэтому рассмотрим по 30 слов с самыми большими весами из каждого документа.

Листинг 5: Исходный код tf-idf

```
==================
chapter9
==================

[('classification', 0.06549010994764948), ('data', 0.06195895019812775), (u'
    ↪ network', 0.049859943553163685), ('rule', 0.046030191563205064), ('set',
    ↪  0.043779377984305826), ('class', 0.041165211967093954), ('training',
    ↪ 0.03944331884440897), ('learning', 0.03704869077038456), ('tuples',
    ↪ 0.033949812004398075), ('classifier', 0.030461176929345544), (u'value',
    ↪ 0.030051945904481122), (u'unit', 0.028329513382479365), ('tuple',
    ↪ 0.0281085837700573), ('input', 0.02757405969227992), ('frequent',
    ↪ 0.027196332847180194), ('used', 0.02559980577048392), (u'pattern',
    ↪ 0.02559980577048392), (u'method', 0.023002724025652215), ('given',
    ↪ 0.022631712347819116), (u'algorithm', 0.021889688992152913), (u'weight',
    ↪  0.02108850710493153), ('one', 0.020776653958653617), ('support',
    ↪ 0.020604136436435137), ('output', 0.020405642280820514), (u'attribute',
    ↪ 0.020208376783846124), ('layer', 0.01977792337886112), ('may',
    ↪ 0.019663618925154315), ('label', 0.01945946218996652), ('number',
    ↪ 0.019292607247321212), ('neural', 0.01719567823414889)]
==================
chapter8
==================

[('tuples', 0.08077849624281876), ('class', 0.07518831231929143), (u'tree',
    ↪ 0.06862178531755989), (u'rule', 0.05790364282060375), ('classifier',
```

```
   ↪ 0.05290989538594747), ('data', 0.052836186744006475), (u'attribute',
   ↪ 0.052430164146019316), ('decision', 0.049261308071259904), ('set',
   ↪ 0.04369695444234049), ('classification', 0.04004281767195981), ('
   ↪ training', 0.035774304266634936), ('tuple', 0.03108514848018114), ('
   ↪ accuracy', 0.02849548712339351), (u'value', 0.02684649488614383), ('
   ↪ model', 0.025062770773097147), ('test', 0.02500624380216165), (u'
   ↪ positive', 0.02471547352539233), ('measure', 0.023419282773019086), ('
   ↪ given', 0.02284808075416496), ('may', 0.020848873688175527), ('node',
   ↪ 0.020453525573447096), ('number', 0.01999207065989434), ('used',
   ↪ 0.019135267631613155), ('computer', 0.019013136448556452), ('section',
   ↪ 0.018849666622186092), ('one', 0.01856406561275903), (u'example',
   ↪ 0.017992863593904906), (u'method', 0.01713606056562372), ('probability',
   ↪  0.01699659167370956), (u'split', 0.016761599670685908)]
==================
chapter5
==================

[('cube', 0.16047781296323027), ('data', 0.11484567345790793), ('cell',
   ↪ 0.07220026460448283), ('cuboid', 0.06753907086022162), (u'dimension',
   ↪ 0.061737956069588144), ('computation', 0.047057109754959414), ('query',
   ↪ 0.04365681011660941), (u'value', 0.04047569580321684), ('iceberg',
   ↪ 0.03584696107874748), (u'fragment', 0.024492190531728717), ('example',
   ↪ 0.02435123162144753), ('measure', 0.023693090226273267), (u'chunk',
   ↪ 0.023549117239838806), (u'aggregate', 0.02292615037203957), ('count',
   ↪ 0.021907016669853858), ('compute', 0.021243167679858285), ('base',
   ↪ 0.01916215553483904), (u'method', 0.018757029762466338), ('shell',
   ↪ 0.01820711617287389), ('multidimensional', 0.01792392272988043), ('space
   ↪ ', 0.017111676274530696), ('computed', 0.016921244231243953), ('sample',
   ↪  0.016596224749889286), ('set', 0.016453534879356437), ('used',
   ↪ 0.01579539348418218), ('full', 0.015740342046474926), ('one',
   ↪ 0.01546632278659505), ('ranking', 0.015099087564474182), ('partition',
   ↪ 0.013875420269620043), (u'table', 0.013875420269620043)]
==================
chapter4
==================

[('data', 0.2366893195150035), (u'warehouse', 0.06691065049611467), ('cube',
   ↪ 0.04825460446148986), (u'dimension', 0.041571982836334016), ('attribute'
   ↪ , 0.03771172728724586), ('olap', 0.03573374369521041), ('cuboid',
   ↪ 0.0325344275684631), ('table', 0.03027580434500332), ('mining',
   ↪ 0.02796702158448424), (u'query', 0.02573443369325282), ('may',
   ↪ 0.025317514276480473), (u'system', 0.02443434517381255), (u'database',
   ↪ 0.023161533294528948), ('item', 0.022567647825438464), (u'sale',
   ↪ 0.0224788530100108), ('concept', 0.020901668763140854), ('generalization
   ↪ ', 0.019982030867702193), (u'level', 0.019724109959583624), ('processing
   ↪ ', 0.019679272824252157), ('schema', 0.01958705333203391), ('set',
   ↪ 0.019429720258694317), ('value', 0.019135330557805006), ('
   ↪ multidimensional', 0.01900433501089555), ('relation',
   ↪ 0.018673235283545534), ('online', 0.018060998061134207), (u'example',
   ↪ 0.017957771754247776), (u'hierarchy', 0.017662898160954125), ('
```

```
    ↪ generalized', 0.017662898160954125), ('warehousing',
    ↪ 0.01754782056461225), (u'operation', 0.017383646327741687)]
==================
chapter7
==================

[('pattern', 0.23451851545975017), ('mining', 0.08676048471774278), ('data',
    ↪ 0.075394482352973), ('frequent', 0.05940150931062796), (u'rule',
    ↪ 0.05808716309851988), ('set', 0.04584287620457152), (u'constraint',
    ↪ 0.043201097680456706), ('association', 0.03589165575955895), ('support',
    ↪  0.03233982439774094), (u'item', 0.03210080065970835), (u'method',
    ↪ 0.025762938693478208), ('transaction', 0.020636228995526796), ('may',
    ↪ 0.020458804256585635), ('used', 0.018564470529124), (u'section',
    ↪ 0.017427870292647023), ('example', 0.016670136801662367), ('colossal',
    ↪ 0.01640525962781287), ('space', 0.016291270056170044), (u'level',
    ↪ 0.01553353656518539), (u'measure', 0.01553353656518539), ('context',
    ↪ 0.015428963457305966), ('quantitative', 0.015367025450979172), ('number'
    ↪ , 0.015154669819693064), ('cluster', 0.014806184664025973), ('itemset',
    ↪ 0.014781784659012505), ('also', 0.014775803074200736), (u'attribute',
    ↪ 0.01452179077462997), ('itemsets', 0.014149710943860949), ('annotation',
    ↪  0.014126751346172193), ('information', 0.014018069583216084)]
==================
chapter6
==================

[('frequent', 0.13590259677070407), ('itemsets', 0.08837064467845752), ('
    ↪ itemset', 0.07499166619443592), ('mining', 0.07246412291043529), (u'
    ↪ pattern', 0.06919506473402467), (u'set', 0.06211210535180168), ('support
    ↪ ', 0.05659340838401262), (u'association', 0.05333580379692876), (u'item'
    ↪ , 0.050010694154504626), ('transaction', 0.050010694154504626), ('data',
    ↪  0.04849102961675745), (u'rule', 0.0461637176810812), (u'measure',
    ↪ 0.03595963994051676), ('candidate', 0.029954041737216403), ('apriori',
    ↪ 0.029492196460216), (u'database', 0.028577539516859787), ('confidence',
    ↪ 0.028101660065048488), ('minimum', 0.027786994499490222), ('count',
    ↪ 0.02747840338159595), ('algorithm', 0.026697308440686688), ('table',
    ↪ 0.024094223371411316), ('example', 0.02342825026427607), (u'subset',
    ↪ 0.0230818588405406), ('closed', 0.022674046989867396), (u'method',
    ↪ 0.021793721176070764), ('two', 0.021793721176070764), ('number',
    ↪ 0.01797981997025838), ('fptree', 0.017038794566092624), (u'correlation',
    ↪  0.01664113429845356), ('one', 0.016345290882053075)]
==================
chapter1
==================

[('data', 0.28357515537693945), ('mining', 0.1292299017257228), (u'pattern',
    ↪ 0.047490847277297424), ('database', 0.04329667278598708), ('knowledge',
    ↪ 0.036074778220254775), (u'system', 0.031508350597437715), ('information'
    ↪ , 0.0305950650728743), ('analysis', 0.02785520849918407), ('may',
    ↪ 0.026028637450057245), ('example', 0.02420206640093042), ('user',
    ↪ 0.023288780876367005), (u'model', 0.0230301450989293), (u'set',
```

```
    ↪ 0.021005567064958475), ('customer', 0.019178996015831652), (u'query',
    ↪ 0.01878493287589725), ('learning', 0.01842411607914344), ('search',
    ↪ 0.017809067728986536), (u'kind', 0.01704230737320768), (u'class',
    ↪ 0.01704230737320768), ('conference', 0.01662926133563772), ('discovery',
    ↪  0.01661926240102204), ('web', 0.01661926240102204), ('many',
    ↪ 0.016439139442141416), ('used', 0.015525853917578004), (u'transaction',
    ↪ 0.015199895765293337), ('warehouse', 0.01510536795028973), ('interesting
    ↪ ', 0.015069211155296297), ('process', 0.015069211155296297), ('
    ↪ classification', 0.014739292863314752), (u'application',
    ↪ 0.01461256839301459)]
==================
chapter3
==================

[('data', 0.24719922243251993), (u'attribute', 0.10381897421430028), ('value',
    ↪  0.07609513430751949), ('may', 0.04310591423200526), ('hierarchy',
    ↪ 0.028705622629503998), ('example', 0.026831232328084903), (u'method',
    ↪ 0.026391376060411382), ('section', 0.025511663525064336), ('concept',
    ↪ 0.025511663525064336), ('set', 0.025511663525064336), ('also',
    ↪ 0.022432669651349674), ('number', 0.02155295711600263), ('analysis',
    ↪ 0.02023338831298206), ('reduction', 0.019793532045308534), ('mining',
    ↪ 0.019353675777635013), (u'technique', 0.018473963242287967), ('given',
    ↪ 0.01759425070694092), ('use', 0.016714538171593876), ('normalization',
    ↪ 0.016503351203720786), ('discretization', 0.016503351203720786), ('used'
    ↪ , 0.01627468190392035), (u'bin', 0.01601795852125841), ('transformation'
    ↪ , 0.01600832715419417), (u'histogram', 0.014880808279230846), ('missing'
    ↪ , 0.014815805228131097), ('cleaning', 0.014777819636364781), (u'level',
    ↪ 0.014075400565552737), (u'cluster', 0.014023149249213784), ('two',
    ↪ 0.013635544297879214), ('database', 0.013310124899269268)]
==================
chapter2
==================

[('data', 0.18012507824056384), (u'attribute', 0.12449873317032606), (u'value'
    ↪ , 0.0862994803826657), (u'object', 0.057694534883809635), ('set',
    ↪ 0.035623622716100374), ('two', 0.03462014038606938), (u'measure',
    ↪ 0.034118399221053884), ('dissimilarity', 0.032561397776214224), (u'plot'
    ↪ , 0.026934840426061975), ('visualization', 0.02640670630006076), ('
    ↪ numeric', 0.025800184751631317), (u'example', 0.024585317085759412), ('
    ↪ median', 0.02436210723148127), (u'number', 0.023581834755728413), ('
    ↪ distance', 0.02298701183509629), (u'mean', 0.02207661126068192), ('
    ↪ binary', 0.02159368595517658), ('may', 0.021574870095666426), ('
    ↪ similarity', 0.019411254438525753), ('section', 0.01906616427058893), ('
    ↪ also', 0.01906616427058893), ('used', 0.018062681940557934), ('figure',
    ↪ 0.017059199610526942), (u'distribution', 0.01655745844551144), ('nominal
    ↪ ', 0.01651211824104404), ('dimension', 0.016194957160367618), (u'
    ↪ technique', 0.016055717280495943), ('ordinal', 0.01584281121442675), ('
    ↪ matrix', 0.015315889654035242), (u'quartile', 0.01508738116144596)]
==================
chapter11
```

```
==================

[('cluster', 0.14918228284706334), ('clustering', 0.08693761542353712), ('data
    ↪ ', 0.05470231981705706), ('graph', 0.05223891385890433), ('set',
    ↪ 0.048841356979515235), (u'constraint', 0.047736351907872926), ('object',
    ↪  0.0472943265824943), (u'vertex', 0.04461359519816045), ('method',
    ↪ 0.04053832629299764), ('two', 0.035654190595046126), (u'customer',
    ↪ 0.030281641327299446), ('example', 0.028816400617913986), ('similarity',
    ↪  0.028343458945299545), ('algorithm', 0.027839573478323682), ('may',
    ↪ 0.02735115990852853), ('gene', 0.025870891767201466), ('distance',
    ↪ 0.02535993695105749), ('using', 0.023932264919962466), ('one',
    ↪ 0.022467024210577006), ('matrix', 0.02210653489738266), ('analysis',
    ↪ 0.021978610640781854), ('bicluster', 0.021845055828749585), ('fuzzy',
    ↪ 0.021020099560851185), ('section', 0.02100178350119155), ('cut',
    ↪ 0.019942145737217795), ('product', 0.019705969409372626), ('
    ↪ probabilistic', 0.018241006675526847), ('network', 0.017901131965452344)
    ↪ , ('search', 0.017582888512625485), ('highdimensional',
    ↪ 0.01707810645387257)]
==================
chapter10
==================

[('cluster', 0.16738050897158657), ('clustering', 0.12707195461811033), (u'
    ↪ object', 0.11495257713023825), ('data', 0.0895693433697855), (u'method',
    ↪  0.0710000892565373), (u'set', 0.05061032003414709), ('hierarchical',
    ↪ 0.03613268776482239), ('algorithm', 0.03458978707369765), ('may',
    ↪ 0.0298564477899285), ('number', 0.029128241746271707), ('analysis',
    ↪ 0.02366669641884576), ('distance', 0.02112950927096679), ('density',
    ↪ 0.020491648585635507), (u'point', 0.020025666200561797), ('two',
    ↪ 0.019661563178733403), ('partitioning', 0.019661563178733403), ('kmeans'
    ↪ , 0.01940963613272918), (u'cell', 0.01930907998174155), ('quality',
    ↪ 0.019276043545443388), (u'example', 0.017476945047763027), ('given',
    ↪ 0.01565642993862104), ('one', 0.015292326916792647), ('using',
    ↪ 0.015292326916792647), (u'value', 0.01492822389496425), ('used',
    ↪ 0.014564120873135853), (u'measure', 0.014564120873135853), ('space',
    ↪ 0.014200017851307457), ('neighborhood', 0.01406289608818123), (u'group',
    ↪  0.01383591482947906), ('find', 0.01383591482947906)]
==================
chapter13
==================

[('data', 0.2275328896297406), ('mining', 0.1446514934072234), (u'network',
    ↪ 0.04099668057633035), ('system', 0.03753119828944175), (u'pattern',
    ↪ 0.037140248307260065), ('analysis', 0.03596739836071501), (u'sequence',
    ↪ 0.03292860546418439), ('information', 0.03244884852107984), ('customer',
    ↪  0.028539348699262997), (u'method', 0.027366498752717943), ('web',
    ↪ 0.023578507616329566), ('may', 0.022284148984356038), ('many',
    ↪ 0.022284148984356038), (u'trend', 0.01984547368602976), ('research',
    ↪ 0.01915654912690256), ('user', 0.018765599144720875), (u'example',
    ↪ 0.017983699180357506), ('set', 0.01759274919817582), ('model',
```

```
   ↪ 0.016562297424154604), ('search', 0.01524704930508571), ('used',
   ↪ 0.01524704930508571), (u'application', 0.014465149340722341), (u'graph',
   ↪  0.01367014724148368), (u'intrusion', 0.012785381090446015), ('
   ↪ biological', 0.012756965718456546), (u'technique', 0.012510399429813916)
   ↪ , ('classification', 0.011830212445824719), (u'item',
   ↪ 0.011830212445824719), (u'help', 0.011728499465450546), ('also',
   ↪ 0.011728499465450546)]
==================
chapter12
==================

[('outlier', 0.26254896546044004), (u'object', 0.13602319713859953), ('data',
   ↪ 0.11392813120541295), ('detection', 0.1013210616302909), (u'method',
   ↪ 0.0734720193079806), ('normal', 0.048627730510549506), (u'cluster',
   ↪ 0.04734489541089902), ('set', 0.039061073556141586), ('model',
   ↪ 0.03799268609733297), ('may', 0.0367360096539903), ('point',
   ↪ 0.030225830727966703), ('contextual', 0.029890572054802587), ('example',
   ↪  0.026040715704094393), (u'attribute', 0.022045138846600615), ('distance
   ↪ ', 0.020830890570316337), ('density', 0.020012988062170035), ('using',
   ↪ 0.019530536778070793), ('used', 0.019065523997640537), ('collective',
   ↪ 0.018878256034612163), ('distribution', 0.018600511217210278), ('two',
   ↪ 0.017670485656349766), ('detect', 0.01692388461208486), ('many',
   ↪ 0.016740460095489254), ('small', 0.016275447315058995), ('number',
   ↪ 0.016275447315058995), ('context', 0.016096597258880806), ('one',
   ↪ 0.01581043453462874), ('statistical', 0.015345421754198481), ('
   ↪ application', 0.015345421754198481), ('cell', 0.014989726370703732)]
```

6. HAL-матрица