Data Mining

Versuch 5 Merkmalextraktion mit Nicht-Negativer Matrixfaktorisierung

von

Armin Schwarz (as219)

Florian Tatzel (ft020)

Marc Walter (mw136)

Das git Repository mit den Quellcodes findet sich unter https://github.com/floriant/DataMining
Alle Dateien zu diesem Versuch befinden sich im Unterordner Versuch_5.

2.1 Binden Sie die Bibliothek feedparser ein. Übergeben Sie der Funktion feedparser.parse() die Elemente der angelegten Feed-Liste.

Was für eine Datenstruktur liefert die Funktion zurück?
Die Funktion parse() liefert ein Dictionary zurück. Hauptsächlich interessant ist dabei die Liste entries, die alle Einträge des RSS-Feeds enthält.

Wie kann auf den Titel und die Beschreibung des RSS-Feeds zugegrien werden?
Um auf den Titel und die Beschreibung eines Eintrags des RSS-Feeds zuzugreifen, kann entry.title und entry.description verwendet werden

Lassen Sie sich die Titel und Inhalte der aktuellen Artikel der von Ihnen ausgewählten Nachrichten-Feeds anzeigen.
Ist implementiert in der Funktion scrape_feedlist() in der Unterfunktion download_feeds()

2.2.1 Erklären Sie den Ablauf und die Rückgabewerte der Funktionen stripHtml(h) und separatewords(text)

stripHtml(h)
Die Funktion erhält eine Zeichenkette mit HTML-Tags und liefert eine um diese Tags bereinigte Kopie der Zeichenkette zurück.

```
def stripHTML(h):
  p = ''
  s = 0
  for c in h:
    if c == '<':
      s = 1
    elif c == '>':
      s = 0
      p += ' '
    elif s == 0:
      p += c
  return p
```

Die Funktion iteriert über jedes Zeichen c einer übergebenen Zeichenkette h.
Die Variable s wird dafür benutzt um festzustellen, ob der aktuelle Buchstabe zwischen ‘<’ und ‘>’ liegt, oder außerhalb. Wenn das Zeichen ‘<’ eingelesen wird, wird s auf 1 gesetzt und alle nun folgenden Zeichen werden ignoriert, bis ein ‘>’ gefunden wird. Dann wird der Wert wieder auf 0 gesetzt und alle folgenden Zeichen (bis zum nächsten ‘<’) werden zur Zeichenkette p hinzugefügt, die dann zurückgegeben wird.

separatewords(text)
Dieser Funktion wird eine Zeichenkette text bestehend aus durch verschiedene Trennzeichen separierte Wörter übergeben. Zurückgegeben wird eine Liste aus Zeichenketten, wobei jede Zeichenkette ein Wort ist.

```
def separatewords(text) :
  splitter=re.compile('\\W*')
  return [ s.lower() for s in splitter.split(text) if len(s)>4 and s not in sw
]
```

Um dies zu Erreichen wird zuerst ein Regulärer Ausdruck kompiliert, der alle Nicht-alphanumerischen Zeichen (ausgenommen '_') erkennt. Dieser Ausdruck wird dazu verwendet, die Wörterkette zu trennen.
Durch die Verwendung des Regulären Ausrucks entfällt die Notwendigkeit, verschiedenste Satz- oder Trennzeichen gesondert zu betrachten oder durch ein bestimmtes Trennzeichen, zum Beispiel ein Leerzeichen, zu ersetzen.

Die Rückgabeliste selbst wird durch eine List Comprehension erzeugt:
Hierbei wird jedes Wort s aus der Liste aller Wörter, die durch Aufteilen der Übergebenen Zeichenkette an allen Trennzeichen gebildet wird, geprüft: Wenn dieses Wort mehr als vier Zeichen enthält und nicht in der Liste der Stopfworte existiert, wird es der Rückgabe-Liste hinzugefügt.
Die Liste der Stopfworte entstammt der Bibliothek nltk.corpus.

2.2.2 Implementation der Funktion makematrix

Die Funktion makematrix erhält als Input zwei Python Dictionarys, allwords und articlewords. Um eine spätere Rekonstruktion des Kontextes aus der sich die Article/Word Matrix bildet herstellen zu können wird das Ergebniss stattdessen jeweils als Dictionary in eine Globale Variable geschrieben. Diese muss dann entsprechend in eine Liste oder Matrix für die weitere Verwendung konvertiert werden.

Die Textdatei mit der Article/Word Matrix findet sich unter Versuch_5/res/awMatrix.txt .

Der Programmcode ist entsprechend Dokumentiert.

2.3 In welchen Artikeln sind welche Merkmale (Feature) stark vertreten?

Feature 0:
Most important words:
media, record, would, return, spent, depressed
Most important articles:

Power workers strike despite ruling,
Engineering strike hits South Africa,
France country profile

Feature 5:
Most important words:
house, visits, diplomat, since, fellow, corruption
Most important articles:
Sarkozy 'shocked' by allegations,
Sarkozy questioned by police,
VIDEO: Sarkozy placed under investigation

Feature 8:
Most important words:
involved, china, southern, cancer, rockets, territory
Most important articles:
Japanese woman abducted by North Korea an icon, but South Korean husband forgotten,
China's Xi visits South Korea as ties strengthen, wary eye on the North,
North Korea's economy 'grows 1.1%'

Feature 10:
Most important words:
ireland, thursday, japan, first, cashless, close
Most important articles:
Brazil land investors lose millions,
Brazil 2014: The ones to watch,
Favelas show hidden side of Brazil

Feature 11:
Most important words:
retailer, investors, largest, founder, soccer, suspected
Most important articles:
India summons U.S. diplomat over report of NSA spying,
India summons U.S. diplomat,
India to U.S.: Let's talk NSA spying

Feature 12:
Most important words:
share, ceasefire, sanctions, france, court, nation
Most important articles:
China urges U.S. to be more objective ahead of key meeting,
China cites Japan wartime 'confessions' in propaganda push,
Top China aides ousted from Party

Feature 13:
Most important words:
early, record, would, customers, depressed, doctors
Most important articles:
Nagorno-Karabakh profile,
Karachay-Cherkessia profile,
Germany country profile

Merkmale können eine Thema erfassen und sind deshalb in Artikeln zum selben Thema stark vertreten.
Merkmale können aber auch einen Typ von Artikel beschreiben. Hier “Overview” und “Country profile” die sich ebenfalls stark ähneln.

Wie lassen sich die insgesamt m Merkmale beschreiben, so dass aus dieser Merkmalsbeschreibung klar wird, welches Thema den Artikeln, in denen das Merkmal stark vertreten ist, behandelt wird?

Ein Merkmal m lässt sich durch eine Menge der wichtigsten Wörter für das jeweilige Merkmal beschreiben.
Hierfür müssen nur die Worte nach ihrer Gewichtung in H sortiert werden. Zur Übersichtlichkeit wählt man eine Obergrenze, hier N=6 Worte, die das Merkmal beschreiben.