Praktikum Data Mining

Merkmalsextraktion mit der Nicht-Negativen Matrixfaktorisierung

Oliver Fesseler Maria Florusß Stefan Seibert

Daniel Grießhaber

3. Juli 2014

Durchführung des Versuchs

RSS Nachrichten Feeds einbinden und parsen

Was für eine Datenstruktur liefert die Funktion `feedparser.parse()` zurück?

Die Funktion `feedparser.parse()` liefert ein FeedParserDict zurück, welches ein verschteltes Python Dictionary ist. Eingabe ist ein XML-, bzw. HTML-Dokument, das in ein Python Dictionary gewandelt wird, welches das eingegebene XML-/HTML Dokument strukturiert abbildet. Innerhalb der FeedParserDict Klasse wird ein Dictionary an keys angelegt, welche vom zurückgelieferten FeedParserDict Objekt aus aufgerufen werden können. Ruft man beispielsweise auf ein solches Objekt `description` auf, so erhält man den die Werte von `subtitle` und `summary`. Man kann demnach sehr einfach auf verschiedene Elemente des ursprünglichen XML-/HTML-Dokuments zugreifen.

Hier ein Ausschnitt aus dem resultierenden FeedParserDict:

```
{'feed': {
  'lastbuilddate': u'Mon, 30 Jun 2014 12:19:07 GMT',
  'subtitle': u'The latest stories from the Europe section of the BBC
   News web site.',
  'language': u'en-gb',
  'links': [
    {'href': u'http://www.bbc.co.uk/news/world/europe/#sa-ns_mchannel=
   rss&ns_source=PublicRSS20-sa', 'type': 'text/html', 'rel': '
   alternate'},
    {'type': 'text/html', 'rel': 'alternate'},
    {'href': u'http://feeds.bbci.co.uk/news/world/europe/rss.xml', '
   type': u'application/rss+xml', 'rel': u'self'}
  ],
  'title': u'BBC News - Europe',
  'image': {
    'height': 60,
    'width': 120,
    'href': u'http://news.bbcimg.co.uk/nol/shared/img/bbc_news_120x60.
   gif',
    'link': u'http://www.bbc.co.uk/news/world/europe/#sa-ns_mchannel=
```

```
    rss&ns_source=PublicRSS20-sa',
16   'title': u'BBC News - Europe'
17  },
18  ...
```

Wie kann auf den Titel und die Beschreibung des RSS-Feeds zugegriffen werden?

Über den Key `entries` erhält man von einem geparsten Feed dessen einzelne Artikel. Jeder Artikel ist wiederum ein Dictionary in dem über die Keys `title` und `description` auf Titel und Beschreibung zugegriffen werden.

Lassen Sie sich die Titel und Inhalte der aktuellen Artikel der von Ihnen ausgewählten Nachrichten-Feeds anzeigen.

Der Code-Ausschnitt `stripHTML(article.title + ' ++++ ' + article.description)` liefert so zum Beispiel folgendes Ergebnis:

> As caliphate declared, Iraqi troops battle for Tikrit ++++ BAGHDAD (Reuters) - Iraqi troops battled to dislodge an al Qaeda splinter group from the city of Tikrit on Monday after its leader was declared caliph of a new Islamic state in lands seized this month across a swathe of Iraq and Syria.

Sammeln und speichern aller Worte der aktuellen Artikel aller eingebundenen Feeds

Erklären Sie den Ablauf und die Rückgabewerte der Funktionen `stripHtml(h)` und `separatewords(text)` und nehmen Sie diese in das File `newsfeatures.py` auf.

`stripHtml(h)`:

Die Funktion `stripHtml(h)` iteriert über alle Character eines eingegebenen Strings und prüft ob es sich um einen validen Character an einer validen Position handelt. Es werden

alle Character aus dem String eliminiert, die sich innerhalb von spitzen Klammern befinden, inklusive der spitzen Klammern selbst. Dieser bereinigte String wird letztendlich zurückgegeben.

Die fiktive Eingabe

```
'Dies it ein <strong>fiktiver</strong> Testtext'
```

würde zu folgendem Rückgabewert führen:

```
'Dies ist ein  fiktiver  Testtext.'
```

`separatewords(text)`:

Die Funktion `separatewords(text)` extrahiert aus dem Eingabetext zunächst alle Wörter als einzelne Strings. Diese werden daraufhin auf ihre Länge geprüft - Wörter mit einer Länge kleiner 4 werden ignoriert - und mit einer Liste von 'stopwords' abgeglichen. Diese Liste enthält Worte wie 'mine', 'yours', 'there', 'their' und 'and', also Worte, die für unsere Zwecke als irrelevant angesehen werden. Für englischsprachige Feeds wird hier selbstverständlich eine Liste an englischen stopwords verwendet. Zurückgegeben wird eine Liste an Worten, die länger als 4 Zeichen und nicht in der Liste der 'stopwords' enthalten sind.

Die fiktive Eingabe

```
'This is a fictitious testtext, with a few important words and many words contained in
 the stopwords list.'
```

liefert als Rückgabe:

```
['fictitious', 'testtext', 'important', 'words', 'words', 'contained', 'stopwords']
```

Anzeige der Merkmale und der Gewichte

Geben Sie mindestens 3 Merkmale an und zu jedem Merkmal mindestens 3 Artikel, die das jeweilige Merkmal behandeln.

Mit den Parametern `m = 5` und `it = 10` ergeben sich folgende Merkmale:

1. {'obama', 'former', 'mcdonald', 'secretary', 'president', 'chief'} unter anderem mit den Artikeln
    - Obama to nominate former P&G CEO Bob McDonald as veterans secretary

- From Big House to White House: Ex-Convicts To Be Honored By Obama Administration
- Biden, Clinton, you're rich. Own it
- UK leaving EU 'would be bad news'
- What Obama discovered about Iraq

2. {'european', 'including', 'tikrit', 'iraqi', 'inflation', 'across'} unter anderem mit den Artikeln
   - ... as Iraq's Shiites answer call to arms
   - Iraqi army presses Tikrit assault as lawmakers scramble to fill posts
   - ISIS declares Islamic State amid battle for Tikrit ...
   - Group: ISIS 'crucifies' men in Syria
   - What are Namibia's 'fairy circles'?
3. {'korea', 'south', 'monday', 'hostile', 'american', 'military'} unter anderem mit den Artikeln
   - North Korea says to try two detained U.S. citizens
   - Boeing, Airbus enter bids for $1.38 billion South Korean refueling plane order: sources
   - N. Korea Preparing to Indict 2 American Tourists
   - Can North Korea take this Hollywood joke?
   - VIDEO: Italy finds 30 bodies in migrant boat
4. {'ukraine', 'ukrainian', 'separatists', 'russian', 'president', 'poroshenko'} unter anderem mit den Artikeln
   - Ukraine's Poroshenko urges Putin to tighten borders after violence
   - Israel bombs Gaza after rocket attacks, Hamas gunman killed
   - Russia protests over shooting of cameraman in Ukraine
   - Separatists attack Ukraine base

- By Numbers: Where Are the Hardest Places to Live in the U.S.?

5. {'family', 'honor', 'police', 'reports', 'young', 'saturday'} unter anderem mit den Artikeln
   - Syria fighters hail declaration of Islamic 'caliphate'
   - Islamist rebels in Somalia kill three as Ramadan starts
   - Two police officers killed in bomb blasts near Cairo palace
   - Is a Vast Marine Sanctuary Any Use if You Can't Police It?
   - Gay Pride Parades From Around the World
6. {'first', 'india', 'economy', 'years', 'decades', 'minister'} unter anderem mit den Artikeln
   - Exclusive: Facebook scores record 1 billion interactions for World Cup
   - India's Modi eyes first labor overhaul in decades to create jobs
   - Israel cranks up media campaign ahead of Iran nuclear deal deadline
   - Eyes on defense deals, Western powers rush to court India's Modi
   - Stocks set for quarterly gains, yen at five-week high

Implementierung eines primitiven Clusterings

Um eine Gruppierung nach den gefundenen Merkmalen und deren Gewichtung zu testen, wurde ein primitiver Gruppierungsalgorithmus implementiert. Dieser Iteriert über alle Artikel und gruppiert diese nach ihren höchst bewerteten Merkmalen. Im Versuch wurden so alle Artikel in nur 5 Cluster gegliedert.

Ungeklärte Fragen

Auffällig hohe Kosten nach $N = 10$ Iterationen Bei der Nacharbeitung der Aufgaben fiel uns auf, dass die Kosten nach $N = 10$ Iterationen zu einem Wert deutlich über der

Abbruchbedingung von $k = 5$ konvergieren. Das bedeutet dass die Abbruchbedingung nie auch nur annähernd ausgeführt wird.