

텍스트 마이닝: part 1

Web Scarping

Jong-June Jeon

University of Seoul

September 5, 2017

들어가며

텍스트 마이닝

- 의미있는 정보를 추출하기 위해서 문자로 표시된 대용량 자료원을 분석하는 과정
- 텍스트 마이닝의 목적은 텍스트 데이터를 분석에 가용한 형태로 변형함으로써 텍스트와 관계있는 정보를 추출하는 것이다.

텍스트 마이닝 알고리즘

- 단어의 빈도 분석: 텍스트 내에서 출현하는 단어의 빈도를 시간에 따라 분석하는 방법
- 단어의 연관성 분석: 특정 단어와 연관이 있는 단어 집합을 찾고, 그것들의 연관성을 찾음으로써 텍스트가 가진 정보를 추출하는 방법
- 주제 모형화 (topic modeling): 텍스트가 가지고 있는 주제 분석
- 텍스트 데이터로부터 정보의 추출: 의미론적 정보(semantic information)를 탐색함으로써 텍스트가 가진 정보를 끌어내는 방법
- 텍스트 요약: 긴 텍스트를 짧은 텍스트로 요약하는 방법

텍스트 마이닝의 절차


- 텍스트 수집(ex: web scraping)
- 형태소 분석(Morphological segmentation) 혹은 단어의 특징 추출(word embedding)
- 빈도 분석(Word count)
- 네트워크 분석 (Word Network analysis)
- 통계량 산출 및 시각화(Produce summarized statistics and visualization)

Web scraping: 웹 문서를 자동으로 수집하는 기술

- 대부분의 웹 문서는 HTML로 되어 있음
- <http://ranking.uos.ac.kr> 을 보자

Journey Into Data Science Home Teaching >

Statistical Learning



Deep Ranking and Model Selection Criteria

Work Information

Work	Assistant Professor, Department of Statistics, University of Seoul 서울시립대학교 통계학과 조교수
Office	Musee Building 7th floor 8712 (151세관 7층 712호, 전화 02-4680-2637)
email	jjjeon@uos.ac.kr jjjeon@gmail.com

Work Information
Education
Publication
Students

Web scraping: 웹 문서를 자동으로 수집하는 기술

- 웹브라우저인 크롬에서 페이지 소스 보기를 선택한다.

```
<!DOCTYPE html>
<html lang="ko">

<head>
  <meta charset="utf-8">
  <meta http-equiv="X-UA-Compatible" content="IE=edge">
  <meta name="viewport" content="width=device-width, initial-scale=1">
  <meta name="description" content="">
  <meta name="author" content="Jong-June Jeon">

  <title>Jong-June Jeon's Homepage</title>

  <link rel="stylesheet" href="http://maxcdn.bootstrapcdn.com/bootstrap/3.2.0/css/bootstrap.min.css">

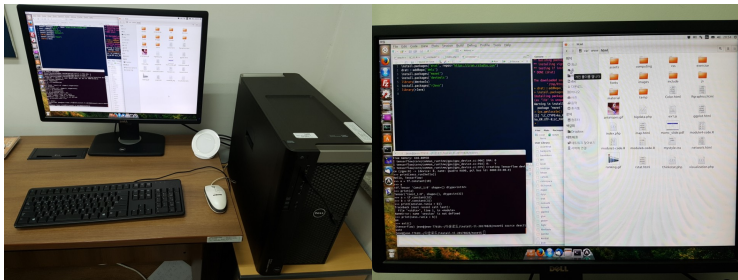
  <link rel="stylesheet" type="text/css" href="mystyle.css">

</head>
<!-- ===== Body start ===== -->
<body>

  <nav class="navbar navbar-default navbar-fixed-top" role="banner">
    <div class="container">
      <div class="navbar-header">
        <button class="navbar-toggle" type="button" data-toggle="collapse" data-target=".navbar-collapse">
          <span class="sr-only">Toggle navigation</span>
          <span class="icon-bar"></span>
          <span class="icon-bar"></span>
        </button>
        <a class="navbar-brand" href="http://ranking.uos.ac.kr/index.php">Journey Into Data Science</a>
      </div>
```

Web scraping: 웹 문서를 자동으로 수집하는 기술

- <http://ranking.uos.ac.kr>에서 다운로드 받은 페이지 소스를 웹브라우저가 해석한 화면을 보고 있는 것이다.
- <http://ranking.uos.ac.kr>는 어디에 있나?



Web scraping: 웹 문서를 자동으로 수집하는 기술

- 클라이언트 컴퓨터에서 웹 브라우저를 통해 `http://ranking.uos.ac.kr`로 연결하는 순간 기다리고 있던 `http://ranking.uos.ac.kr`의 서버컴퓨터가 `'/www/html/index.php'` 파일을 클라이언트 컴퓨터로 보내줌.
- Web scraping 은 서버(server) 컴퓨터에 웹페이지를 요청하여, 다운로드 받은 html 문서를 수집하는 기술

Web scraping: HTML 문서의 구조 (태그들로 이루어진 트리구조)

```
<!DOCTYPE html>
<html>
  <head>
    <meta charset="utf-8">
    <title>Hello</title>
  </head>
  <body>
    <p>Hello HTML5!</p>
  </body>
</html>
```

- Document Type Definition
- html
- head
- body

Web scraping: HTML 문서의 구조 (태그들로 이루어진 트리구조)

```
<p>  
  <strong>HTML</strong>은  
  <em>Hypertext Markup Language</em>  
  의 약자입니다.  
  <a href="http://www.w3.org/People/Berners-Lee/">  
  Tim Berners-Lee</a>가 최초로 고안하였습니다.  
</p>
```

HTML은 Hypertext Markup Language의 약자입니다.
[Tim Berners-Lee](http://www.w3.org/People/Berners-Lee/)가 최초로 고안하였습니다.

HTML은 Hypertext Markup Language의 약자입니다. Tim Berners-Lee가 최초로 고안하였습니다.

중요한 태그의 예

- ` ... ` 하이퍼 링크
- `
` 줄바꾸기
- `<hr>` 가로줄
- `<center>...</center>` ...을 가운데 정렬
- `...` ...의 폰트를 바꿈
- `......` ...을 순서없는 목록으로 만들
(기본: 까만동그라미)
- `......` ...을 순서있는 목록으로 만들
(기본: 숫자)
- `<table></table>` 표만들기
- `<tr></tr>` 행(`<table>...</table>...에 넣는다`)
- `<td></td>` 열(`<tr>...</tr> ...에 넣는다`)

Web scarping 예

Web scarping tool in R

- Package: httr
 - 'get' 방식으로 읽기: 요청 자료의 형태를 URL 주소 형식으로 전달
 - 'post' 방식으로 읽기: 서버에서 요청하는 form의 형태로 자료를 요청
- Package: rvest
 - html,xml 형식의 파일을 효과적으로 읽음
 - 태그를 이용하여 원하는 텍스트를 추출할 수 있는 함수를 내장하고 있음 'html_nodes', 'xml_nodes'
 - 문서 내에 링크들을 다 저장하고 싶으면 < a > 를 찾아 텍스트를 저장하면 된다.
 - 문서 내에 테이블을 저장하고 싶으면 < table >를 찾아 필요한 텍스트를 저장하면 된다.

형태소 분석

- 단어를 형태소 별로 정리, 분할, 원형 추출의 작업을 거침
- 분리된 단어, 동사 원형, 형용사등으로 변형된
- R에서 한글의 형태소 분석은 'KoNLP' 패키지를 사용함.

Bags of words model

- $D1 = [\text{the cat chased the mouse}]$
- $D2 = [\text{the dog chased the cat}]$
- $W = [\text{the, chased, dog, cat, mouse}]$ ($n = 5$)
- $V1 = [2, 1, 0, 1, 1]$
- $V2 = [2, 1, 1, 1, 0]$

N-gram model

- $W = [\text{the cat, cat chased, chased the, the mouse, the dog, dog chased}]$
2-gram model
- $V1 = [1, 1, 1, 1, 0, 0]$
- $V2 = [1, 0, 1, 0, 1, 1]$

The bags of words model 은 N-gram model 의 특별한 형태임.

word vector 의 정규화

- Unusual words like elephant determine the topic much more than common words such as 'the' or 'have'.
- weight each term frequency by its inverse document frequency

$$idf_i = \log(N/n_i)$$

(N is the total number of term frequency and n_i is that of the i th term frequency.

- $w_i = tf_i \times idf_i$ (weighted term frequency).

How to measure of the similarity of sentence

- Cosine measure
- Euclidean distance

연관성 분석

- 신뢰도(Confidence) and 특이도(lift)와 같은 측도로 단어들의 연관성을 분석하는 도구
- 시각화 도구로 Graph, River plot 등이 있다.

군집분석

- 유사한 단어 혹은 유사한 문서들의 그룹을 만들어 주는 분석 방법
- 대표적으로 $K - means$ 클러스터링 방법이 있다.