LDA 알고리즘을 활용한 인터넷 기사 트렌드 분석 웹 사이트

4조

2015154023 양현용
2015154040 한승훈
2017152022 안윤빈

- 1. LDA란?
- 2. LDA 예시
- 3. 설계 개요
- 4. 설계 목표
- 5. 설계 방향
- 6. 구현 과정
- 7. 구현 결과 (영상)
- 8. 역할 분담

1. LDA란?

모든 글에는 **주제(Topic)**가 있다고 가정할 수 있다!

해당 문서에 주제를 분석하는 알고리즘이 있다?

1. 실제 글을 작성하려면 '**주제**'를 정하고 나서, 어떤 **단어**를 사용할지 결정
2. 즉, **반대로 생각**하면 '**단어**'를 통해 여러 **'주제'를 파악**할 수 있고
3. 여러 주제 중의 **출현 빈도 확률이 가장 높은 주제**가 '문서의 주제' !

1. LDA란?

- **토픽 모델링 기법** : 말뭉치로 부터 토픽을 추출하는 기법

- 즉, 기계 학습 및 자연언어 처리 분야에서 **토픽 모델링(Topic modeling)**이란 문서 집합의 추상적인 '주제'를 발견하기 위한 통계적인 기법

- **LDA 알고리즘** : 주어진 **문서**에 대하여 각 문서에 **어떤 주제들이 존재**하는지에 대한 **확률모형**

  - **깁스 샘플링 기법**을 통해 '단어'들을 잠재적으로 토픽으로 할당

  - LDA 알고리즘을 통해 출력된 토픽 분포를 통해, 사용자가 직접 토픽을 결정할 수 있게 도와준다.

2. LDA 예시

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

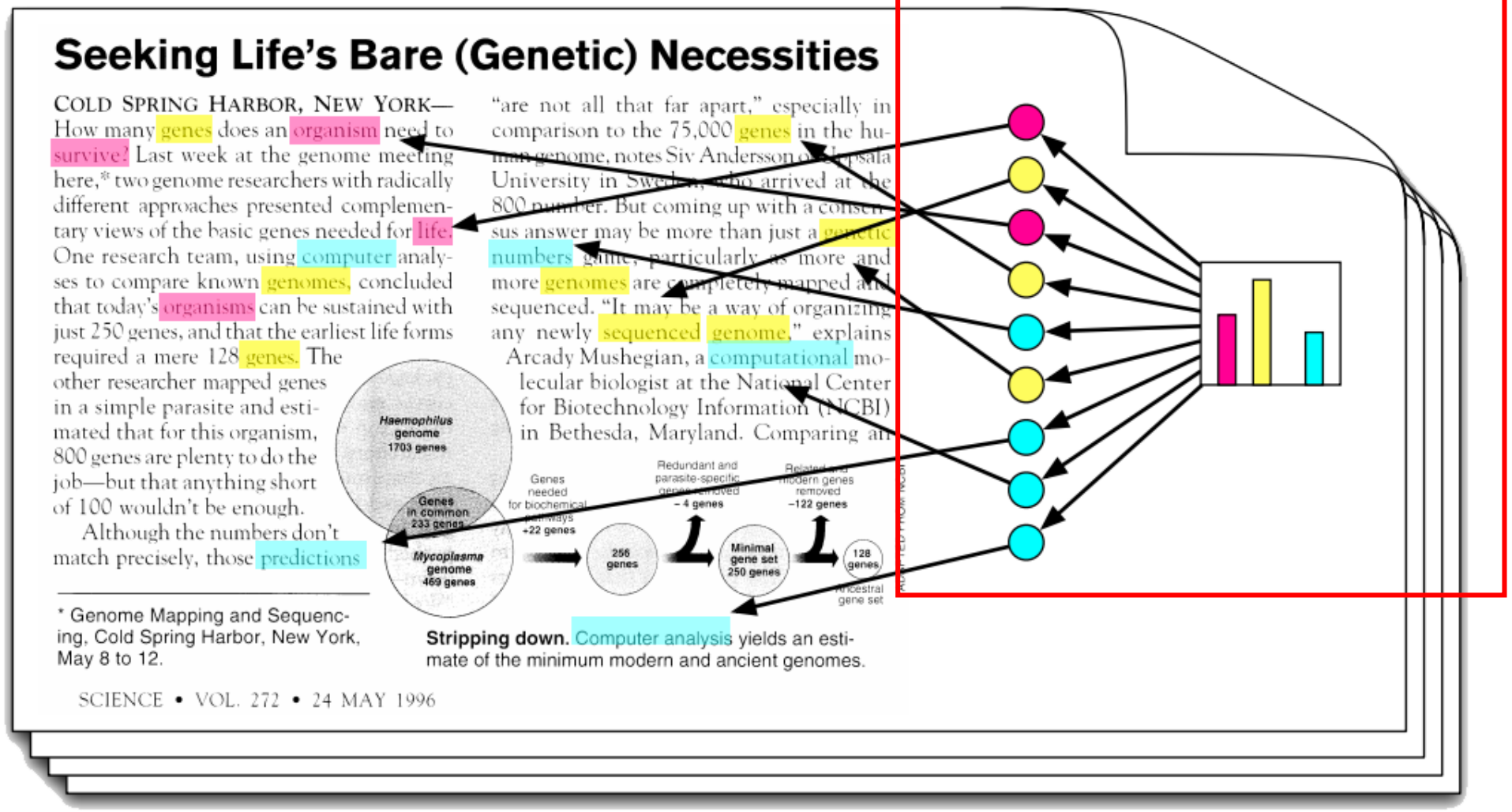
life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Topic proportions & assignments



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

3. 설계 개요

NEWSWiRE Korea Press Release Network 뉴스와이어는 보도자료 전문 통신사입니다 서비스 안내»

· 로그인 · 회원가입 · 고객센터 · English

키워드 검색

뉴스홈 | 뉴스등록 | 마이뉴스 | 상장사뉴스 | 기업프로필 | 사진 | 동영상 | 행사 | 전시회 | 전문가검색 | 서비스안내

보도자료 수신 언론인 회원 현재 9,337명 오늘 04월 27일(수)

관련 뉴스: 생활 복지 정부 지방자치단체 행사 서울

Hi Seoul SOUL OF ASIA

서울시, '2011 제8회 서울시 장애인IT챌린지' 열려

(서울=뉴스와이어) 2011년 04월 22일 -- 장애인의 정보활용 능력 향상 위한 장애인 정보화 축제의 한마당 '2011 제8회 서울시 장애인IT챌린지'가 오는 5월 21일(토), 고려대학교 공과대학(성북구 안암동)에서 열린다.

생활 분야 최근 보도자료

- 부천시민 화장장 이용불편, 5월부터 해소된다 [04-27]
- 전북도, 무주택 기초생활수급가구 230호 대상 '임대보증금' 무이자 지원 [04-27]
- 풍기사람들 '조광지처 홍삼진액', 가정의 달 특별할인 시판 [04-27]

더보기 »

서울시장애인재활협회 장애인재활지원센터 주관으로 2004년에 시작되어 올해로 8회째를 맞는 '서울시 장애인IT챌린지' 대회는 만14~24세의 서울지역 특수학교와 일반학교 특수학급에 재학 중인 장애학생이 참여하는 장애청소년부문과, 만25세 이상의 서울시 등록장애인과 특수학교 전공과 또는 대학교에 재학 중인 장애인이 참여하는 장애청·장년부문으로 나누어 진행되며, 특히 다문화장애가정 외국인부모도 20명 정도 참여하여 그 동안 갈고 닦은 실력을 발휘할 예정으로 있어, 장애인과 비장애인이 함께하는 어울림행사로 치러져 의미를 더하게 된다.

이번 대회는 필수종목으로 참여해야 하는 e-Life(정보검색)챌린지와 e-Tool(엑셀, 파워포인트)챌린지 및 e-Typing(한글타자)챌린지, e-Sports(카트라이더, 피파온라인2, 센스게임)챌린지 등 누구라도 쉽게 참여할 수 있는 종목들로 구성하여 총7개 종목, 총24개 분야로 나뉘어 펼쳐진다.

시상은 장애청소년부와 장애청·장년부 각각 종목별 점수를 합산하여 부문별 최고 득점을 한 대상 수상자에게 서울특별시장상이 수여되며, 끝까지 최선을 다한 중증장애인 참가자 1명에게는 특별상을 수여하고 우수한 성적을 거둔 기관에 기관상도 수여할 계획이다.

"손의 근육에 강직이 심하고 누운 상태로 힘들게 컴퓨터를 사용해야 하지만, IT세상에서는 자유롭게 소통할 수 있어서 기쁩니다." '2010년 제7회 서울시 장애인IT챌린지'에서 특별상을 수상한 이○○군(지체장애1급, 19세)의 말이다.

또한, 장애청소년부의 수상자와 성적 우수자 28명은 서울시대표선수단으로 선발되어 오는 6월 "2011 전국 장애청소년 IT챌린지(주최 : SK텔레콤, 주관 : 한국장애인재활협회)"에 출전한다.

또한, 참가자 중 장애청소년 3명을 서울시대표로 선발하여 오는 9월 베트남에서 개최예정인 국제대회 "글로벌 IT 올림피아드(주최 : LG전자, 주관 : 한국장애인재활협회)"에 출전한다.

이번 대회는 장애인재활지원센터와 함께 LG전자 MC사업본부의 정보요원단이 함께 진행하게 되며, 참여를 희망하는 장애인은 4.18(월)부터 4.30(토)까지 선착순 접수로, 신청방법은 서울시장애인재활협회 홈페이지(www.ssrpd.or.kr)에서 참가신청서를 다운받아 신청하면 된다.

서울시는 "이번 대회가 중증장애인과 가족의 정보활용 능력을 향상시키고 우수 인력을 발굴함으로써 장애인 정보격차 해소에 기여할 것으로 보이며, 앞으로도 장애인의 정보화능력 개발을 위해 지속적이고 체계적인 지원을 해 나갈 계획"이라고 말했다.

출처: 서울특별시청

홈페이지: http://www.seoul.go.kr

- 뉴스 기사 접근성 : 신문 < 인터넷

- 기사는 읽기 귀찮지만, 올해 관심있는 분야의 트렌드를 알고 싶다!

- 기사가 너무 많아 다 읽는 것은 거부감이 든다!

- 시간 없으니 빠르게 트렌드만 알고 싶다!

4. 설계 목표

WHO

- 난독증, 긴 글이나 여러 글을 읽는 것에 대한 거부감을 가진 사람
- 시간이 없지만 기사는 읽고 싶은 바쁜 직장인
- 면접을 준비는 하는데 기사들을 읽기 싫은 취업 준비생
- 트렌드에 민감한 지식인

WHAT

- 각 기사들의 토픽을 기반으로 트렌드를 분석하여 출력하는 프로그램

WHY

- 모든 사람이 기사와 트렌드를 손쉽게 접하게 하기 위해!

5. 설계 방향

Python, 데이터 크롤링

- 2017, 2018, 2019년도 별로 인터넷 뉴스 기사 본문을 텍스트 파일 형태로 크롤링하여 가져온다.
- 모든 분야를 포함하는 것이 어려워 IT 주제의 기사만 가져왔습니다.

R Studio

- LDA 알고리즘을 구현하여 확률모형 결과물을 구현한다.

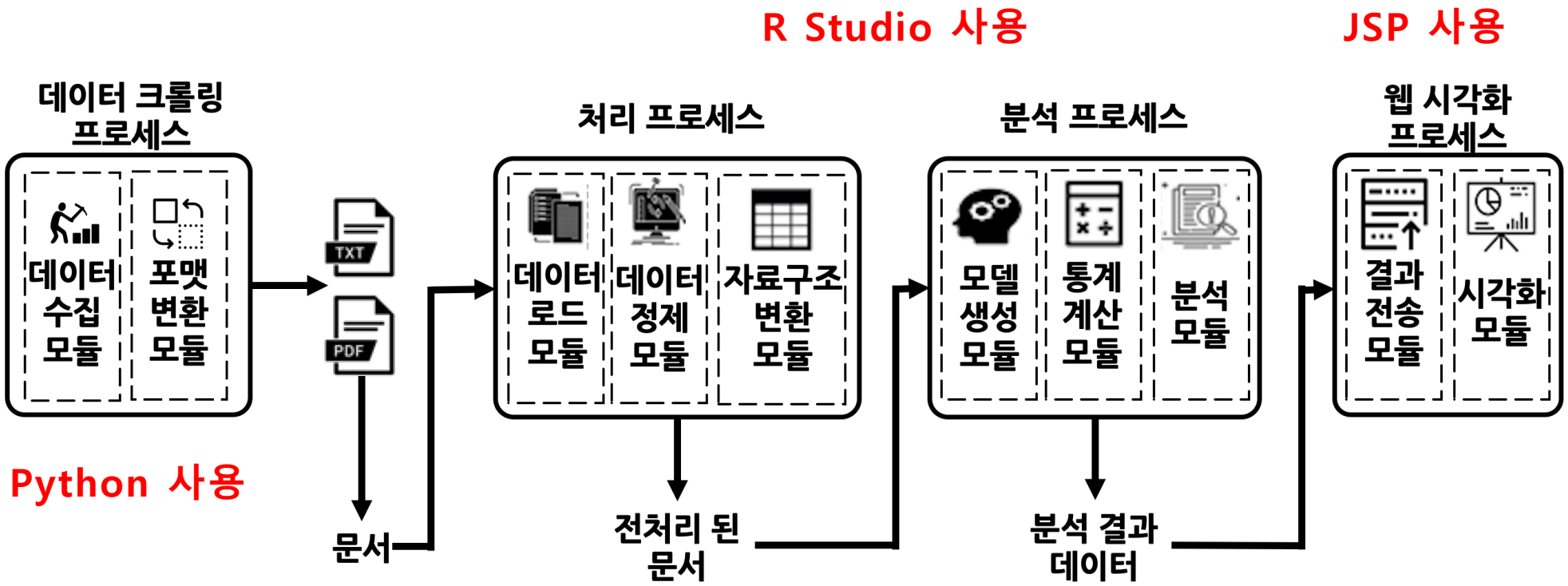
JSP

- 웹 프로그램을 구현하여 LDA 알고리즘 결과물을 출력한다.

-> LDA 알고리즘 결과물에 대한 토픽을 테이블 형태로 출력하는 프로그램

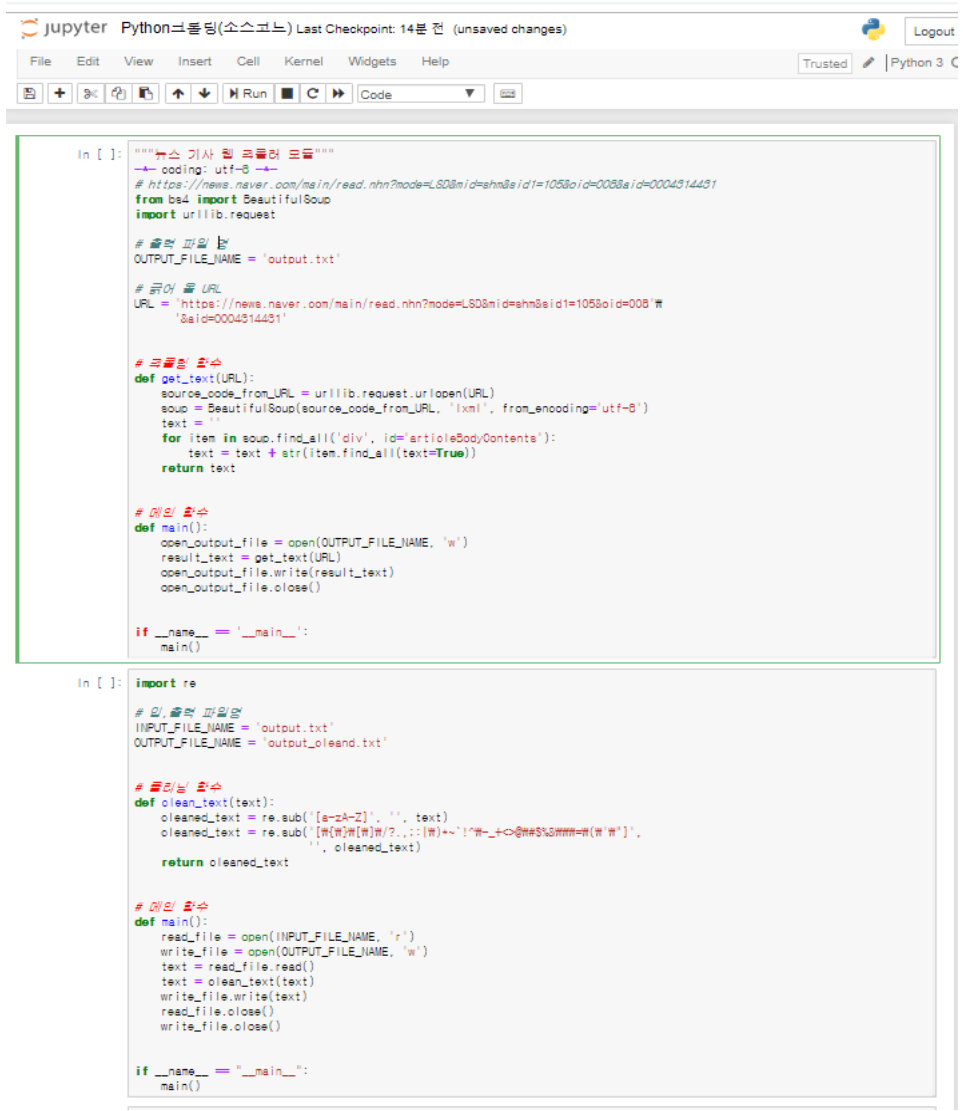
5. 설계 방향

<설계 구조도>



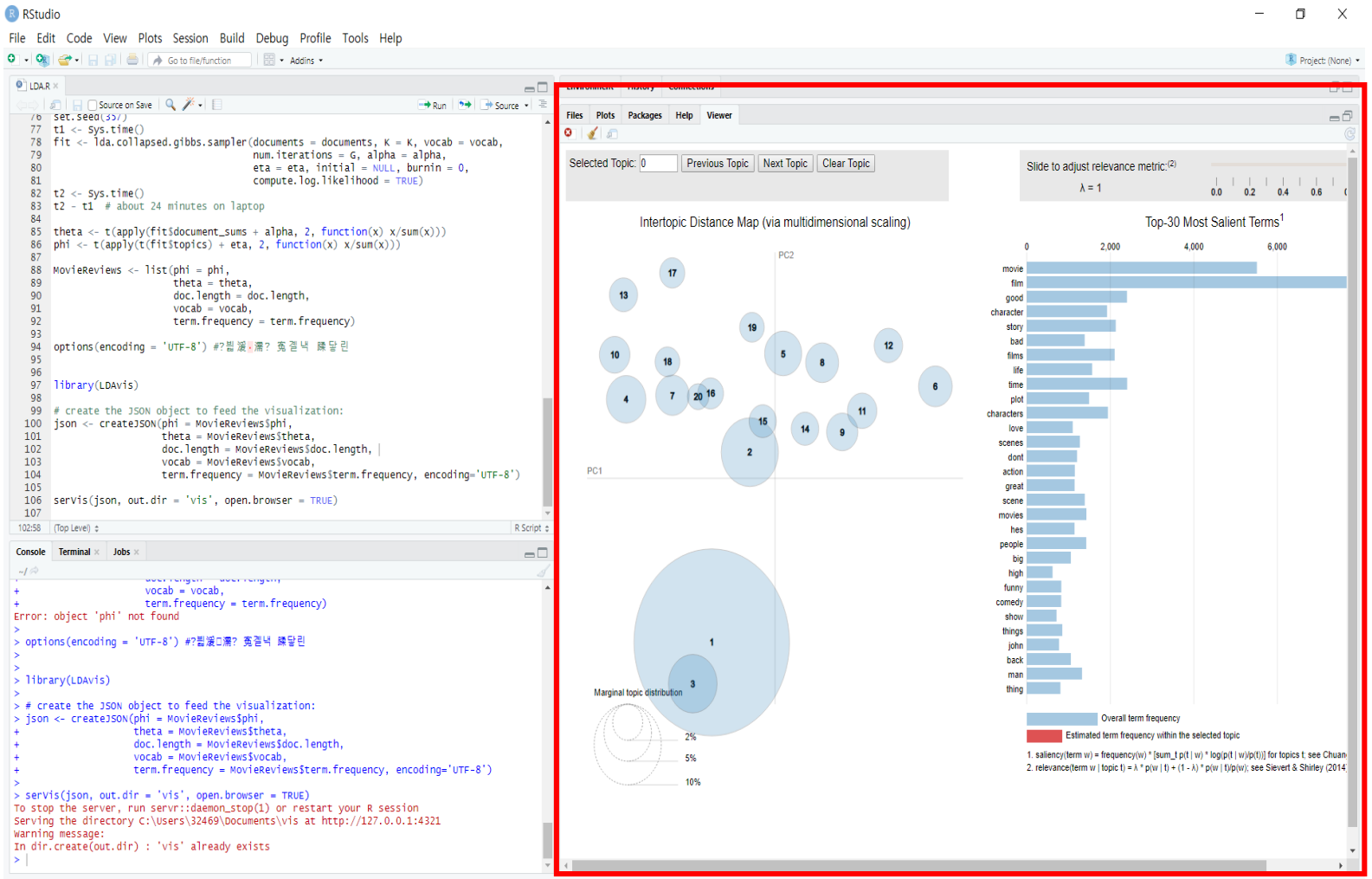
-> LDA 알고리즘 결과물에 대한 토픽을 테이블 형태로 출력하는 프로그램

6. 구현 과정



- Python3 + BeautifulSoup 라이브러리 사용
- 뉴스 기사 웹 크롤러 모듈 구현
- LDA 알고리즘을 돌리기 위한 문서 형태로 정제하기 위한 모듈 구현

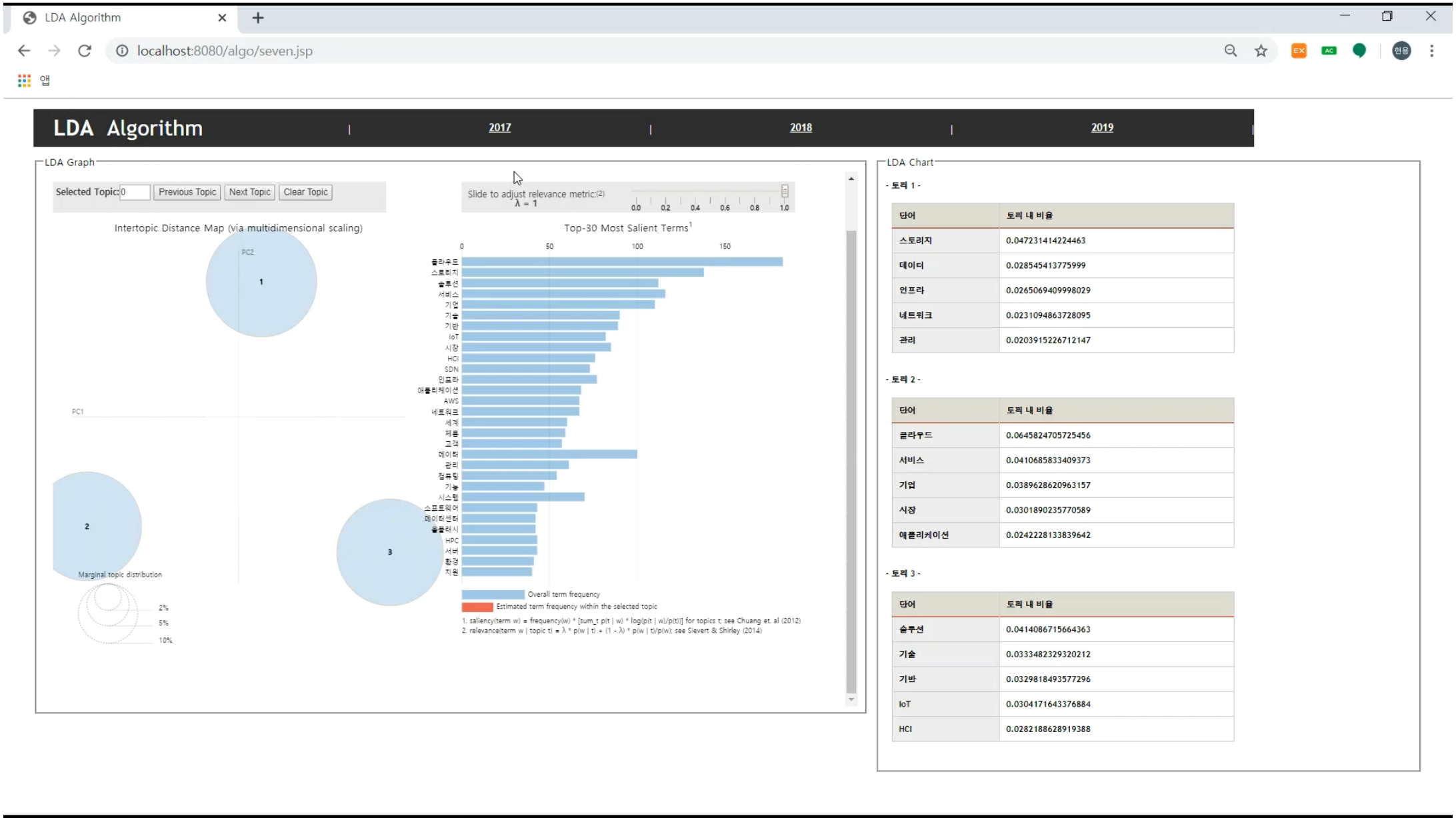
6. 구현 과정



- Rstudio

- 오른쪽은 LDA 알고리즘 실행 시 확률모형으로 나타나는 결과 화면

7. 구현 결과


LDA Algorithm
localhost:8080/algo/seven.jsp
앱
LDA Algorithm
2017
2018
2019
LDA Graph
Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
Marginal topic distribution
2%
5%
10%
Top-30 Most Salient Terms1
클라우드
스토리지
솔루션
서비스
기업
기술
기반
IoT
시장
HCI
SDN
인프라
애플리케이션
AWS
네트워크
세계
제품
고객
데이터
관리
컴퓨팅
기능
시스템
소프트웨어
데이터센터
플랫폼
HPC
서버
환경
지원
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)
LDA Chart
- 토픽 1 -
단어 토픽 내 비율
스토리지 0.047231414224463
데이터 0.028545413775999
인프라 0.0265069409998029
네트워크 0.0231094863728095
관리 0.0203915226712147
- 토픽 2 -
단어 토픽 내 비율
클라우드 0.0645824705725456
서비스 0.0410685833409373
기업 0.0389628620963157
시장 0.0301890235770589
애플리케이션 0.0242228133839642
- 토픽 3 -
단어 토픽 내 비율
솔루션 0.0414086715664363
기술 0.0333482329320212
기반 0.0329818493577296
IoT 0.0304171643376884
HCI 0.0282188628919388

8. 역할 분담

양현용
- LDA 알고리즘 결과 자료 정리
- JSP 웹 프로그램 View 구현 (**결과 화면**)

안윤빈
- 1차 알고리즘 발표
- Rstudio LDA 알고리즘 구현 (**문서 주제 분석 프로그램**)

한승훈
- 2차 구현 발표
- Python3 +BeautifulSoup 크롤러 구현 (**문서 크롤러 프로그램**)

감사합니다