YONSEI UNIVERSITY

Course recommendation system using LDA topic modeling, based on textual data

Sukhyun Hong
Creative Technology Management
Yonsei University
sukhyun96@yonsei.ac.kr

Keeheon Lee
Creative Technology Management
Yonsei University
keeheon@yonsei.ac.kr

Abstract

Information available to students prior to course registration is limited - students need to make decision only based on textual information from course description. It is difficult to decide if the course would be relevant to the future career they would like to pursue. Therefore, the study tries to suggest a course recommendation model for students, based on their career or industry interest. Due to small data of students and cold start problem, it is difficult to apply traditional methodologies as collaborative filtering, so this study suggest LDA topic modeling using textual data. By modeling LDA topic-document distribution for courses and job description data, the study suggest a model which could generate course recommendation which is relevant to career or industry keyword.

INTRODUCTION

Approaches commonly proposed on recommendation system are collaborative filtering, content-based filtering, and hybrid filtering, and these approaches generate recommendation based on user data of previous rating or reaction of users against each items. However, these methodologies bears Cold-Start (CS) problem - It is difficult to give appropriate recommendation to new users, who does not provide any historical record or data regarding usage pattern or rating. This project tries to suggest a course recommendation system which can assist decision making of UIC students, but due to lim it of user data it is difficult to apply traditional recommender methods. Therefore, To generate recommendation for college courses and help student make decision, this study proposes a recommender system based on textual data. Suggested model uses topic modeling of course description, and job description of which students are interested in.

LITERATURE REVIEW

Keyword Extraction

- TF-IDF: for document d and term t in the document,
  $tf(d,t)$ d is Term frequency of t in document d
  $idf(t) = \log(\frac{n}{1+df(t)})$ then, tf$idf(t) = tf(d,t) * idf(t)$
- LDA topic modeling:
  When variables are defined as the following
  $\boldsymbol{\theta_d}$: Per document topic proportions
  $\boldsymbol{Z_{d,n}}$: Per term topic assignment
  $\boldsymbol{W_{d,n}}$: Observed term in document d
  $\boldsymbol{\beta_k}$: Topics
  $\boldsymbol{\eta}$: Topic hyperparameter
  $\boldsymbol{\alpha}$: Dirichlet parameter of θ_d
  when $\boldsymbol{p(\theta_d|\alpha)}$ $\boldsymbol{and}$ $\boldsymbol{p(\beta_k|\eta)}$, follows **Dirichlet Distribution:**
  $\boldsymbol{p(\theta, z, w\,|\alpha, \beta)} =$
  $\prod_{k=1}^{K} \boldsymbol{p(\beta_k|\eta)} \prod_{d=1}^{D} \boldsymbol{p(\theta_d|\alpha)} \prod_{n=1}^{N} \boldsymbol{p(z_{d,n}|\theta_d) p(w_{d,n} \mid z_{d,n}, \beta_k)}$

Document Similarity Computation

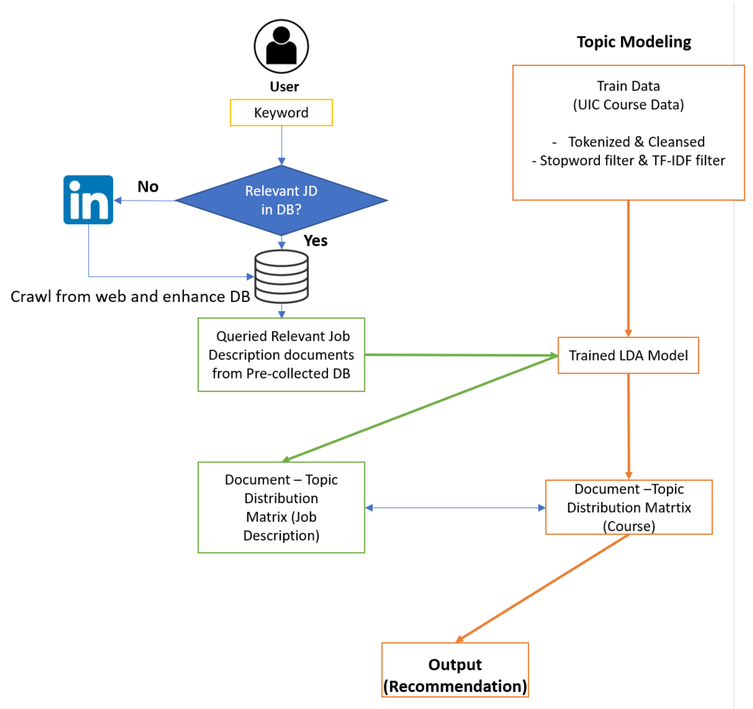
- KL Divergence: when P, G are two different probability distribution:

$$KL(P \,||\, G_\theta) = \sum_{k=y} P(Y) \log(\frac{P(Y)}{G_{\theta(Y)}})$$

- Jensen Shannan Divergence: when p, q are from two different probability distribution, and D_{KL} denotes KL divergence:

$$D_{js}(p,q) = \frac{1}{2}\,[D_{KL}(p, \frac{p+q}{2}) + D_{KL}(q, \frac{p+q}{2})]$$

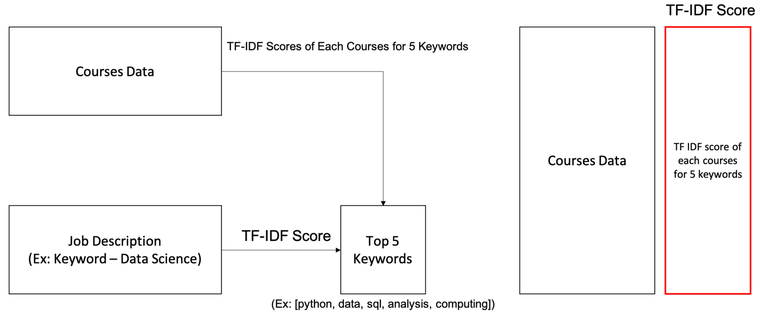
METHODOLOGY



Scoring Method Between Job Description and Courses

- $Score_{TFIDF}$: For each job description D, if $w_1 \ldots w_5$ are top 5 terms with high TF-IDF score, $Score_{TFIDF}$ of course document C is defined as:

$$Score_{TFIDF}(C) = \sum_{i=1}^{5} tfidf(C, w_i)$$



- $\boldsymbol{Score_{LDA}}$: For each job description D and course document D:
  $KL(D||C)_{scaled}$: Min-Max scaled KL divergence of D and C
  $D_{js}(D,C)_{scaled}$: : Min-Max scaled JS divergence of D and C

$$Score_{LDA}(C) = \frac{1}{avg(KL(D||C)_{scaled} + D_{js}(D,C)_{scaled}) + 0.01}$$

- Final scoring method: $Score_{TFIDF}(C) * weight + Score_{LDA}(C)$

The model set weight as 10, and recommend courses with lowest score

RESULT

Recommendation for sample keywords

Business Strategy

	Course_Name
0	BUSINESS MODELS IN CREATIVE TECHNOLOGY INDUSTRY
1	ENTREPRENEURIALLY MANAGING IN CREATIVE TECHNOL...
2	ATTRACTING INVESTMENT FOR AN ENTREPRENEURIAL V...
3	CREATIVE INDUSTRY INTERNSHIP
4	INTRODUCTION TO MANAGEMENT
5	DIGITAL BUSINESS STRATEGY
6	DESIGN BUSINESS
7	PROJECT MANAGEMENT IN CREATIVE INDUSTRY
8	INTRODUCTION TO CULTURE AND DESIGN BUSINESS
9	DESIGN PROJECT MANAGEMENT

Design

Main Recommendation:

	Course_Name
0	USER EXPERIENCE RESEARCH METHODS
1	USER EXPERIENCE PROTOTYPING
2	DESIGN RESEARCH
3	INTRODUCTION TO INFORMATION AND INTERACTION DE...
4	INDUSTRIAL DESIGN BASICS
5	VISUAL SYSTEM
6	STRATEGIC MARKETING
7	INTERACTION DESIGN
8	SOCIAL INNOVATION FOR SUSTAINABILITY
9	INTRODUCTION TO CULTURE AND DESIGN MANAGEMENT

Finance

	Course_Name	...
0	거시경제원론	...
1	금융과경제데이터	...
2	한국경제론	...
3	경기변동과경기예측	...
4	FINANCIAL DATA ANALYSIS	...
5	미시경제원론	...
6	FUNDAMENTAL ECONOMIC ANALYSIS	...
7	THEORY OF FINANCIAL ANALYSIS	...
8	기업금융론	...
9	부동산금융론	...

Data Scienced

	Course_Name
0	SOCIAL COMPUTING
1	INTRODUCTION TO DATA SCIENCE
2	CREATIVE INDUSTRY INTERNSHIP
3	DATA VISUALIZATION
4	BUSINESS STRATEGY AND DATA ANALYTICS
5	MACHINE LEARNING
6	SEMINAR ON SCIENCE, TECHNOLOGY, AND POLICY ISS...
7	AI PROJECT
8	CONTEXT MAPPING
9	CONTEXT MAPPING: INTERACTION MODEL FOR GAME DE...

CONCLUSION

- This study had built the baseline model to recommend courses, based on job / career keyword and relevant job descriptions
- When computing relevance between documents, utilizing TF-IDF score together with LDA provide better results in terms of human intuition.
- User test need to be proceeded as a future work, to improve and quantify result