SSEM X UNIST health care center Project Progress presentation

Improving UNIST Healthcare Center counseling Service using text mining

20176008 Junmo Nam
20176019 Jingyu Jang

Overview

UNIST student survey data (undergraduate /graduate)

Survey 1

No.	Question	Disagree				agree
1	Q1	1	2	3	4	5
2	Q2	1	2	3	4	5
...	...	...	...	...	...	...

Survey 1

No.	Question	Disagree				agree
1	Q1	1	2	3	4	5
2	Q2	1	2	3	4	5
...	...	...	...	...	...	...

...

Survey 6

No.	Question	Disagree				agree
1	Q1	1	2	3	4	5
2	Q2	1	2	3	4	5
...	...	...	...	...	...	...

Survey Results example

Score	Document
~69	Doc 1
70~92	Doc 2
93~115	Doc 3
116~	Doc 4

The purpose of the client's visit dataset (UNIST Healthcare center)

No	Topic1	Topic2	Topic3
Student 1	W1	w2	w3
...	...	...	...
...	...	...	...

Naver Q&A dataset

Question	Answer
Q1	A1
...	...
...	...

Goal

- Identify similarities between Q & A data sets and survey data
- Healthcare service center self-diagnosis service improvement

Data set

- Naver Q&A data
- UNIST healthcare survey data
- UNIST healthcare visit schedule data

Assumptions

1. Text generated from survey will represent survey questions' reply
2. Q and A from Naver data will be related with each other
3. Answer and Survey guideline will be similar in terms of Topic generated from Doc2vec model

Overview

Student 1

UNIST student survey data (undergraduate /graduate)

Question	Disagree				Agree
Survey1 Q1	1	2	3	4	**5**
Survey1 Q2	**1**	2	3	4	5
Survey1 Q3	1	2	**3**	4	5
…	…	…	…	…	…

0%   **~**   **100%**

Generate text based on answer

Question	Survey 1 text	Survey 1 results	…
Student 1	Q1,Q3…	Doc 1	…
Student 2	Q1,Q5…	Doc 2	…
…	…	…	…

Question	Survey text	Survey results
Student 1	Q1,Q3…	Results doc1
Student 2	Q1,Q5…	Results doc2
…	…	…

- Probability of text generating will be calculated based on the score of each question. (0~100%)
- By probability calculated above, each student will get text from question sentences in survey
- Final input for each student will be merged text for all survey questions

Student 1 example
Survey text example
Q1 I try not to tell my feelings to anyone.
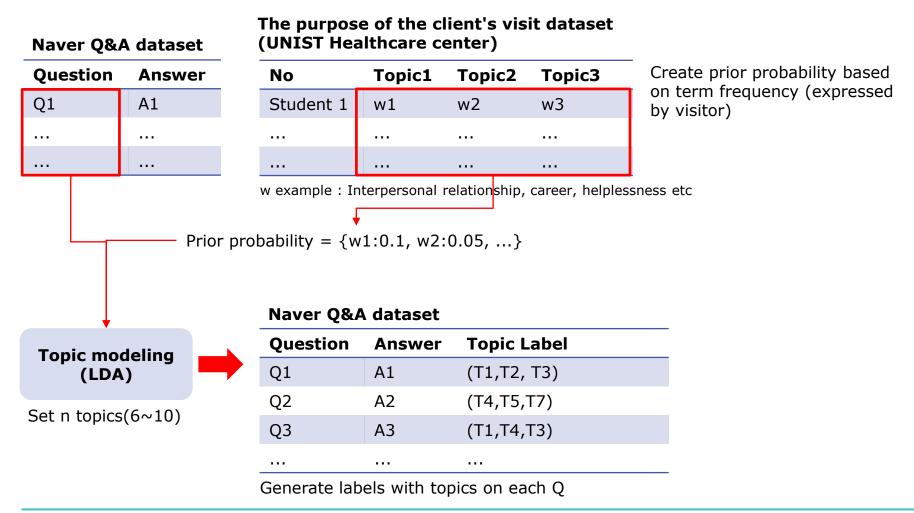Q2 You change your mood by eating something, smoking a cigarette, or taking medication.
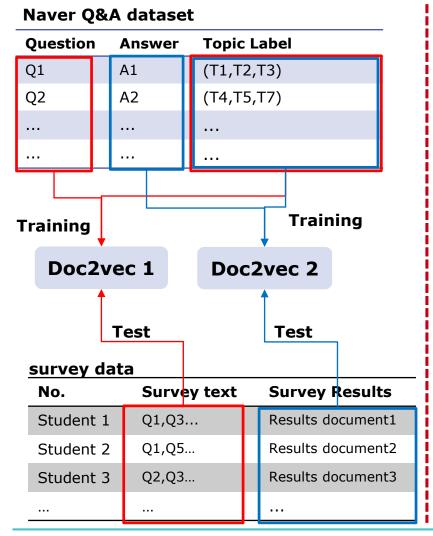…
Survey results example
Survey 1 results: It is possible that your life is stable because you are perceived by yourself. I feel that changes in my life are predictable and calm. Identify which factors create and maintain subjective comfort, and then use it as a way to help when unexpected stress comes. ….

Overview

Naver Q&A dataset

Question	Answer
Q1	A1
...	...
...	...

The purpose of the client's visit dataset (UNIST Healthcare center)

No	Topic1	Topic2	Topic3
Student 1	w1	w2	w3
...	...	...	...
...	...	...	...

w example : Interpersonal relationship, career, helplessness etc

Create prior probability based on term frequency (expressed by visitor)

Prior probability = {w1:0.1, w2:0.05, ...}

Topic modeling (LDA)

Set n topics(6~10)

Naver Q&A dataset

Question	Answer	Topic Label
Q1	A1	(T1,T2, T3)
Q2	A2	(T4,T5,T7)
Q3	A3	(T1,T4,T3)
...	...	...

Generate labels with topics on each Q

Overview

Naver Q&A dataset

Question	Answer	Topic Label
Q1	A1	(T1,T2,T3)
Q2	A2	(T4,T5,T7)
...	...	...
...	...	...

Training → **Doc2vec 1**

Training → **Doc2vec 2**

Test

Test

survey data

No.	Survey text	Survey Results
Student 1	Q1,Q3...	Results document1
Student 2	Q1,Q5...	Results document2
Student 3	Q2,Q3...	Results document3
...	...	...

Model results

No.	Survey text	Model1 results	Model 2 results	Similarity?
Student1	Text1	(T1,T2,T3) ...	(T1,T3,T5)...	0.66
Student2	text2	...	...	...
...	...	...	...	...
...	...	...	...	...

- Embedding two types of test data at each models to obtain similar document labels.

- Model 1 – based on question data(survey questions, questions in Q&A)
- Model 2 – based on answer data (survey results, doctor's answer in Q&A)

- *Model performance can be measured by k/n*
  (k = number of topics shared in both model,
  n = total number of topics in Model 2)