SSEM X UNIST health care center

Improving UNIST Healthcare Center counseling Service using text mining

20176008 Junmo Nam
20176019 Jingyu Jang

Overview

Student A

UNIST student survey data
(undergraduate /graduate)

Question	Disagree				Agree
Q_1	1	2	3	4	**5** 100%
Q_2	**1** 0%	2	3	4	5
Q_3	1	2	**3** 50%	4	5
…	…	…	…	…	…

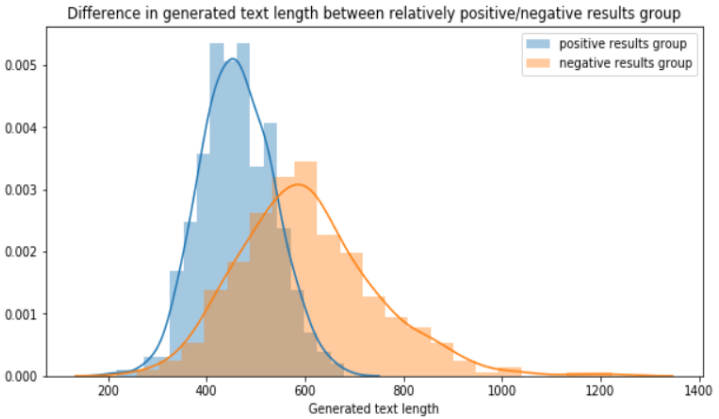
Question	Survey txt	Survey results
Student 1	Q_1, Q_3…	$Document_1$
Student 2	Q_1, Q_2…	$Document_2$
…	…	…

Student A
Survey text
I try not to tell my feelings to anyone. (from Q_1)

You change your mood by eating something, smoking a cigarette, or taking medication. (from Q_3)
…

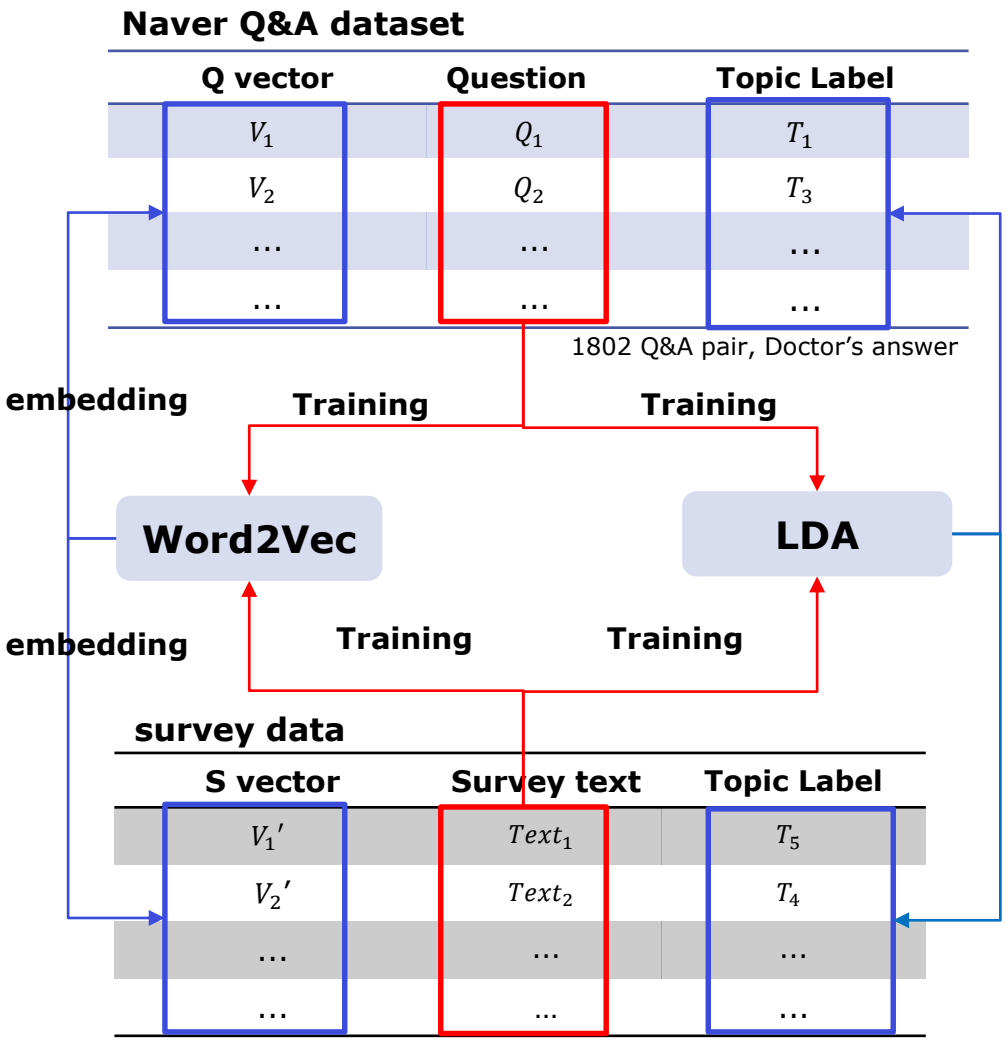
Generate text from survey question

- Probability of text generating will be calculated based on the score of each question.
- Each student will get text from sentences in every survey question
- Excluding 371 students who got normal results in all survey



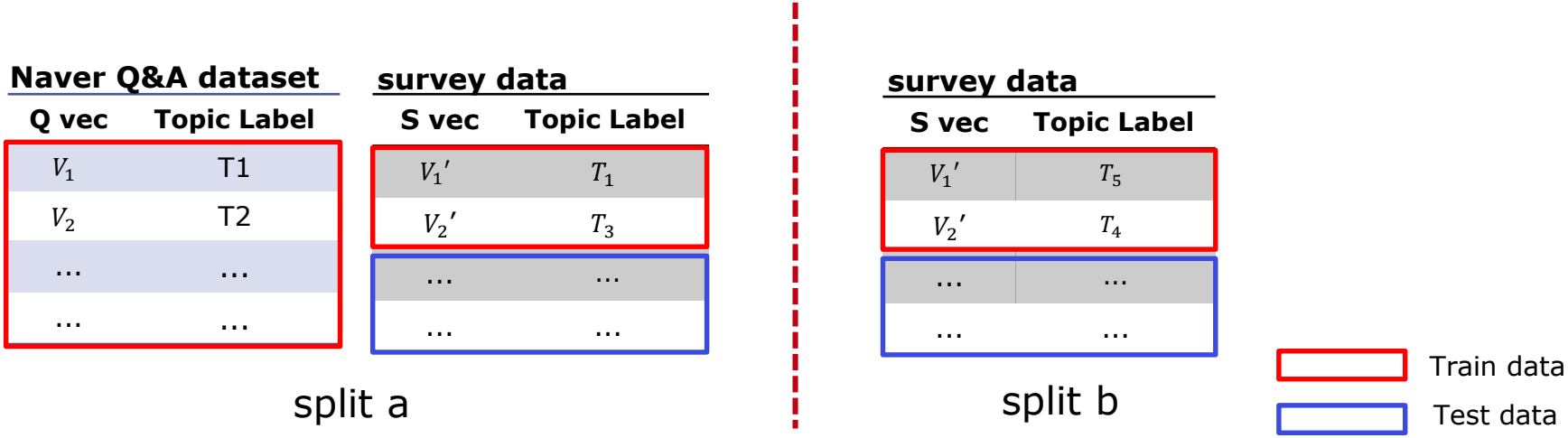
Used survey list
SRI(stress response inventory), PHQ-9(Patient Health Questionnaire-9), MINI+ (Mini-International Neuropsychiatric Interview-Plus)

Overview



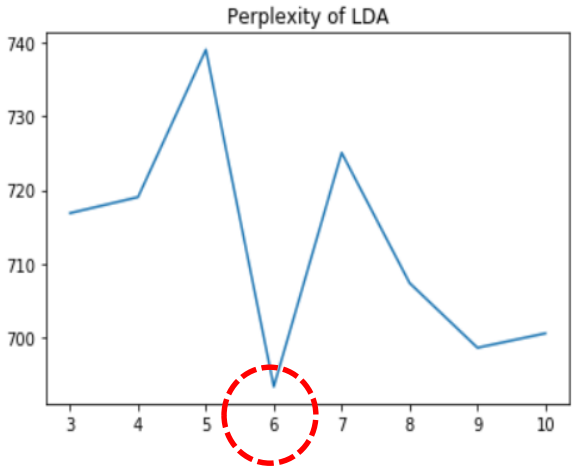
- ✓ **Preprocessing**
  - Select Q&A category from Department of Mental Health
  - Extract nouns, delete noise, etc.
- ✓ **LDA for topic modeling (label)**
  - Questions in Q&A data, survey text
  - Find best number of topics
- ✓ **Word2vec**
  - Question in Q&A data, survey text
  - Get document vector (V_i, V'_i) (average of all word vector)

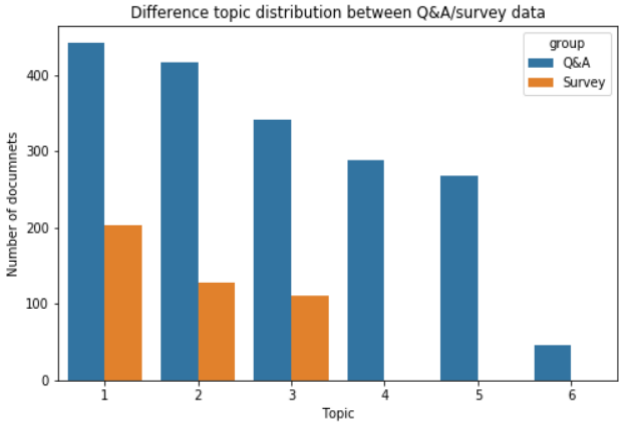
Overview



- Split dataset in two ways
  - ✓ To find out the impact of Q&A data in survey results prediction
- Split A – Train data: All Q&A, Survey 50%, Test data: survey 50%
- Split B – Train data: Survey 70%, Test data: survey 30%
- Classifier - RandomForest, LightGBM

Topic modeling results





Perplexity: The smaller the value is, the better the learning is because it means that the topic model reflects the actual document results well.

Topic 1:
사람 친구 생각 엄마 학교 부모 가족 남자 성격 아빠 → Relationship problem

Topic 2:
생각 정신 정도 병원 우울증 치료 증상 불안 사람 때문 → Depression, unrest

Topic 3:
게임 중독 입원 시간 컴퓨터 알콜 인터넷 본인 하루 치료 → Alcoholism, computer addiction

Topic 4:
대화 상대 자신 사람 집중 생각 친구 관심 고갯짓 반말 → communication

Topic 5:
결벽증 부작용 거식증 페니 샤워 보험 정신 적용 생리 터치 → mysophobia, anorexia

Topic 6:
강박 공포증 확인 복용 벌레 사고 다리 시선 현상 노래 → Obsession, fear

- Set 6 topics by perplexity
- 6 Topics in Q&A data, 3 Topics in survey data

Classification results

Random Forest

test fit report

	precision	recall	f1-score	support
1	0.59	0.85	0.70	101
2	0.61	0.30	0.40	64
3	0.86	0.69	0.77	55
avg / total	0.67	0.65	0.63	220

test fit report

	precision	recall	f1-score	support
1	0.67	0.95	0.79	61
2	0.67	0.32	0.43	38
3	0.86	0.73	0.79	33
avg / total	0.72	0.71	0.68	132

Light GBM

test fit report

	precision	recall	f1-score	support
1	0.66	0.87	0.75	101
2	0.58	0.39	0.47	64
3	0.86	0.69	0.77	55
avg / total	0.69	0.69	0.67	220

test fit report

	precision	recall	f1-score	support
1	0.74	0.92	0.82	61
2	0.64	0.42	0.51	38
3	0.77	0.73	0.75	33
avg / total	0.72	0.73	0.71	132

Spli-a Split-b

- Split-b is slightly higher accuracy than Split-a
- Ensuring the availability of answers from Q&A data with the same topic in the survey results.

Application with R shiny

Conclusion

Implication

- Healthcare service center self-diagnosis service improvement
- Expecting to raise psychological stability through providing relevant case information

Further improvement

- Get professional consultant's advice for facilitating contents
- Updating/Collecting DB related with 'students' cases
- Update model when service can gather enough cumulative text
- Other descriptive analysis when privacy issues are eased (ex : demographic)