3-A. Data Mining

셀레늄(selenium)을 이용해 네트워크에서 뉴스를 크롤링한다. 과거 데이터를 자동으로 수집하고 senti 점수를 계산하는데에 이용된다.
백 테스트를 보다 편리하게 수행하기 위히여, 날짜별 텍스트 데이터로 분류한다.

5. Experimental Results Ang Discussion

- 논문에서는 1379 시장의 2012년 11월 6일부터 지금까지 SSE와 SZSE와 같은 표준 sentimental 요소를 선형 회귀분석한 결과.

TABLE II
THE REGRESSION RESULT OF THE STANDARD SENTIMENTAL FACTOR WITH SSE

standard sentimental factor	11138.27 ***
p	(3.48e-12)

TABLE III
THE REGRESSION RESULT OF THE STANDARD SENTIMENTAL FACTOR WITH SZSE

standard sentimental factor	39796.6 ***
p	(<2e-16)

- 피어슨 상관계수는 0.18731로(SSE 포함) 시장 지수와 표준 sentimantal이 약한 상관관계에 있다. 그러나 조정된 요소를 통한 분석으로 피어슨 상관계수가 0.26119로 개선되었다.

Pearson correlation coefficient		
	SSE	SZSE
random from uniform	-0.00049399	-0.0005597
random from normal	0.00017451	0.00016782
temperature	-0.025135	-0.063723
standard sentimental factor	0.18731	0.22595
adjusted sentimental factor	**0.26119**	**0.28472**

TABLE IV

- 2015년 2월 11일부터 2015년 9월 11일까지 139일의 중국 주식시장 폭락 기간 동안 동일한 방법으로 계산된 피어슨 상관계수는 중간 정도의 상관관계를 보인다. 이후 개선된 방식으로 계산했을 때에는 0.58815의 높은 상관계수가 도출되었다.

Pearson correlation coefficient		
	SSE	SZSE
random from uniform	0.0011922	0.0010882
random from normal	0.00055439	0.00050227
temperature	0.10125	0.064122
standard sentimental factor	0.36284	0.37204
adjusted sentimental factor	**0.58815**	**0.58042**

TABLE V
PEARSON CORRELATION COEFFICIENT FROM 2015/2/11 TO 2015/9/11

- 시계열 분석 결과, 이 기간 동안 주식 시장 붕괴, sentimental 요소는 시장 지수와 거의 동
일하다.

- 2781개의 단어에 1(+) 또는 -1(-) 또는 0(보통)을 준 뒤, senti-score가 0.1 미만인 경우를 제거하니, 452개의 단어에서 73.0088%의 정확도를 갖고 있었다.

Fig. 2. Time Series of 7 Days Average Senti-score and SSE during the crash