**스팸 필터링**

201511054 컴퓨터과학과 임현

**데이터**

* SMS(문자) Spam Data (총 5574개)
* Not SPAM Data : 4827개 (약 86.59%)
* Spam Data : 747개 (약 13.40%)
* 출처 : <https://archive.ics.uci.edu/ml/datasets/sms+spam+collection>

**활용할 예정인 통계 도구**

* 조건부 확률 및 베이즈 정리
* 특정 단어가 나올 경우 스팸일 확률
* 결합 분포 및 상관도 분석
* 특정 단어와 스팸의 연관성 비교

**연구 예상 결과**

* 특정 단어가 나왔을 경우 Spam일 확률
* 특정 단어가 나왔을 경우 Spam이 아닐 확률