Introduction to Text Classification

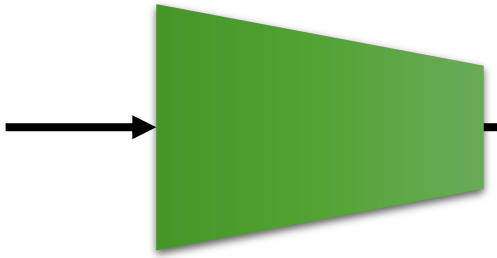
Ki Hyun Kim

nlp.with.deep.learning@gmail.com

What we need

- 텍스트를 입력으로 받아 원하는 항목에 대한 수치를 출력하는 것
  - e.g. 감성 분석(sentiment analysis), 주제 분류(topic classification)

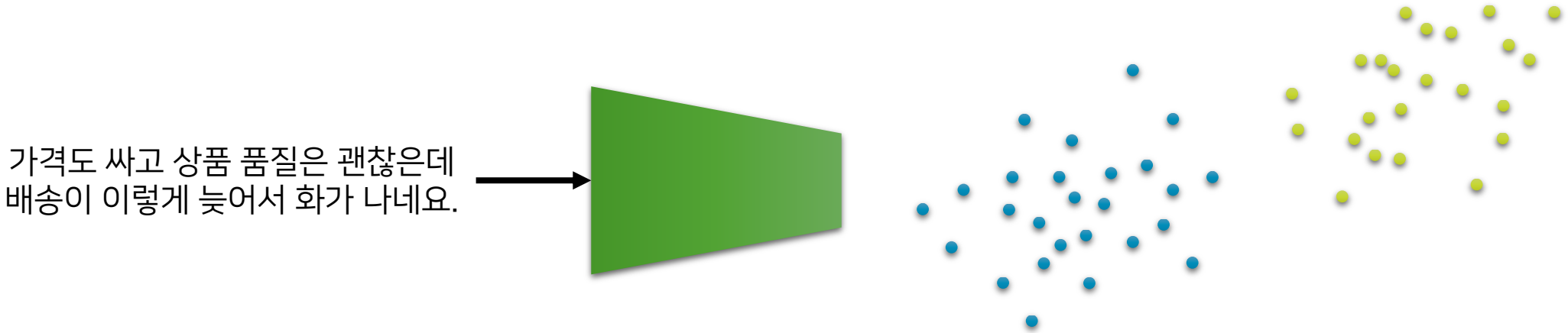
가격도 싸고 상품 품질은 괜찮은데
배송이 이렇게 늦어서 화가 나네요.



항목	수치
품질	긍정
배송	부정
가격	긍정
종합	부정

What we need to do (in specific)

- 문장을 latent space에 projection하여 decision boundary를 찾는 것



In Probabilistic Perspective,

- 문장이 주어졌을 때, 문장이 속할 클래스의 확률 분포 함수를 approximate.

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \, \mathbb{E}_{\mathrm{x} \sim P(\mathrm{x})} \Big[\mathbb{E}_{\mathrm{y} \sim P(\mathrm{y}|\mathrm{x})} \big[\log P(\mathrm{y}|\mathrm{x}; \theta) \big] \Big]$$

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$$

$$\begin{aligned} \hat{\theta} &= \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log P(y_i|x_i; \theta) \\ &= \underset{\theta \in \Theta}{\operatorname{argmin}} - \sum_{i=1}^{N} \log P(y_i|x_i; \theta) \end{aligned}$$

$$\mathcal{L}(\theta) = - \sum_{i=1}^{N} \log P(y_i|x_i; \theta)$$

$$\theta \leftarrow \theta - \alpha \nabla_{\theta} \mathcal{L}(\theta)$$

We will talk about

- RNN을 활용한 Classifier

- CNN을 활용한 Classifier

- RNN과 CNN을 활용한 Text Classification Project
  - RNN, CNN Classifier Model 실습
  - TorchText를 활용한 data loader 실습
  - Model을 학습시키기 위한 Trainer 실습
  - 학습(train.py)과 추론(classify.py) 코드 실습