NATURAL LANGUAGE PROCESSING

LECTURE 0: NLP OVERVIEW

goorm

KAIST AI
Graduate School of AI

DAVIAN

Schedule

구름 3기 자연어 처리 강의 일정

시간 구분	라이브강의: 10am ~ 12pm Quiz: 10am~1pm 퀴즈 해설 라이브강의: 2~4pm 오피스아워: 4~5pm

요일	월	화	수	목	금
날짜					**2/25**
일차					35
강의 제목					NLP Intro, Bag-of-Words
					라이브
					오피스아워
날짜	**2/28**	3/1	**3/2**	**3/3**	**3/4**
일차	36	X	37	38	39
강의 제목	Topic modeling		Word embedding	RNN	LSTM, GRU
				라이브	라이브
	오피스아워		오피스아워	오피스아워	오피스아워
날짜	**3/7**	**3/8**	**3/9**	**3/10**	**3/11**
일차	40	41	X	42	43
강의 제목	Quiz 1	RNNs with Attention		Pre-tokenization (Pre-processing)	Tokenization
	라이브	라이브		라이브	라이브
	오피스아워	오피스아워		오피스아워	오피스아워
날짜	**3/14**	**3/15**	**3/16**	**3/17**	**3/18**
일차	44	45	46	47	48
강의 제목	Transformer	Quiz 2	BERT	Transformer with Huggingface	GPT-1, GPT-2, GPT-3
		라이브		라이브	라이브
	오피스아워	오피스아워	오피스아워	오피스아워	오피스아워
날짜	**3/21**	**3/22**	**3/23**	**3/24**	**3/25**
일차	49	50	51	52	53
강의 제목	Sentence/Document/Sequence/ Token classification (Encoder)	Text generation (Decoder)	Machine Translation	Training Multi-Billion Parameter Language Model	Quiz 3
	라이브	라이브	라이브	라이브	라이브
	오피스아워	오피스아워	오피스아워	오피스아워	오피스아워

INDEX

- What is NLP?
- NLP Applications
- Trends of NLP

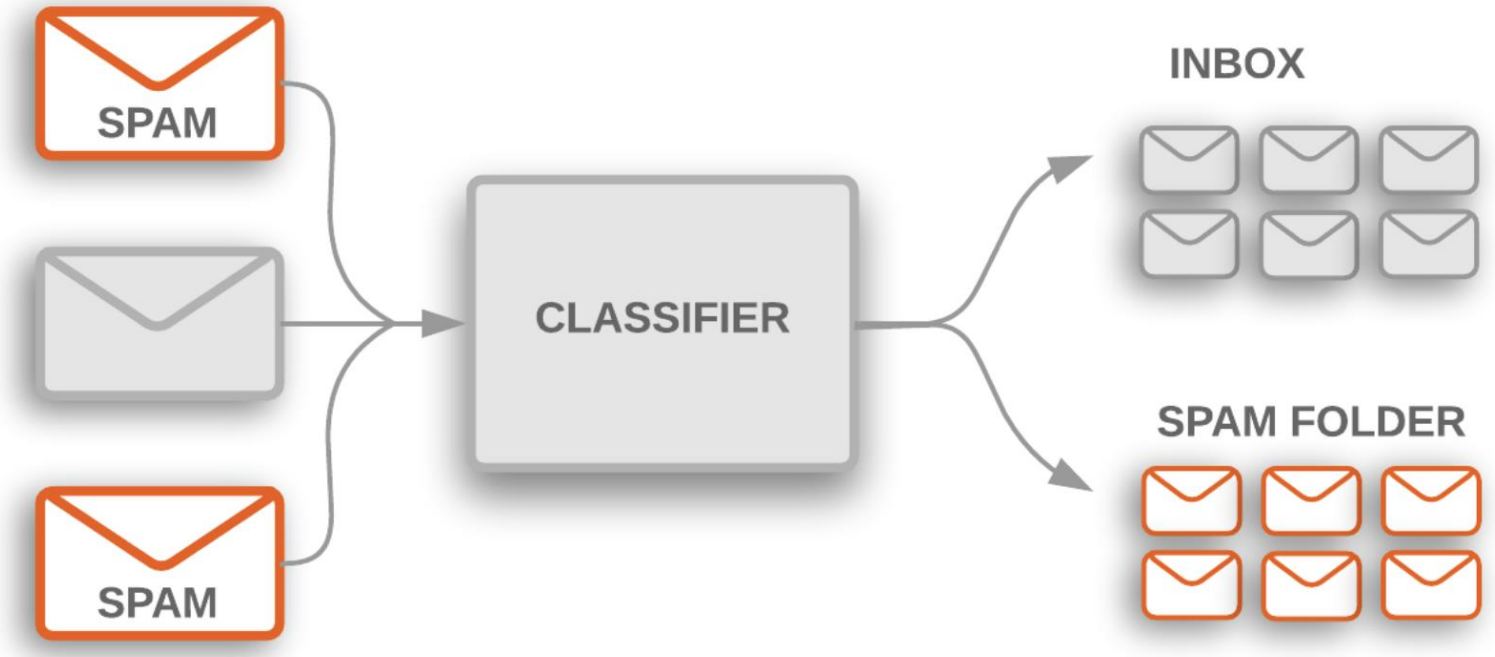
What is NLP?

- NLP(Natural Language Processing)
  - A subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.
  - The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them.
  - The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

https://en.wikipedia.org/wiki/Natural_language_processing

NLP Applications

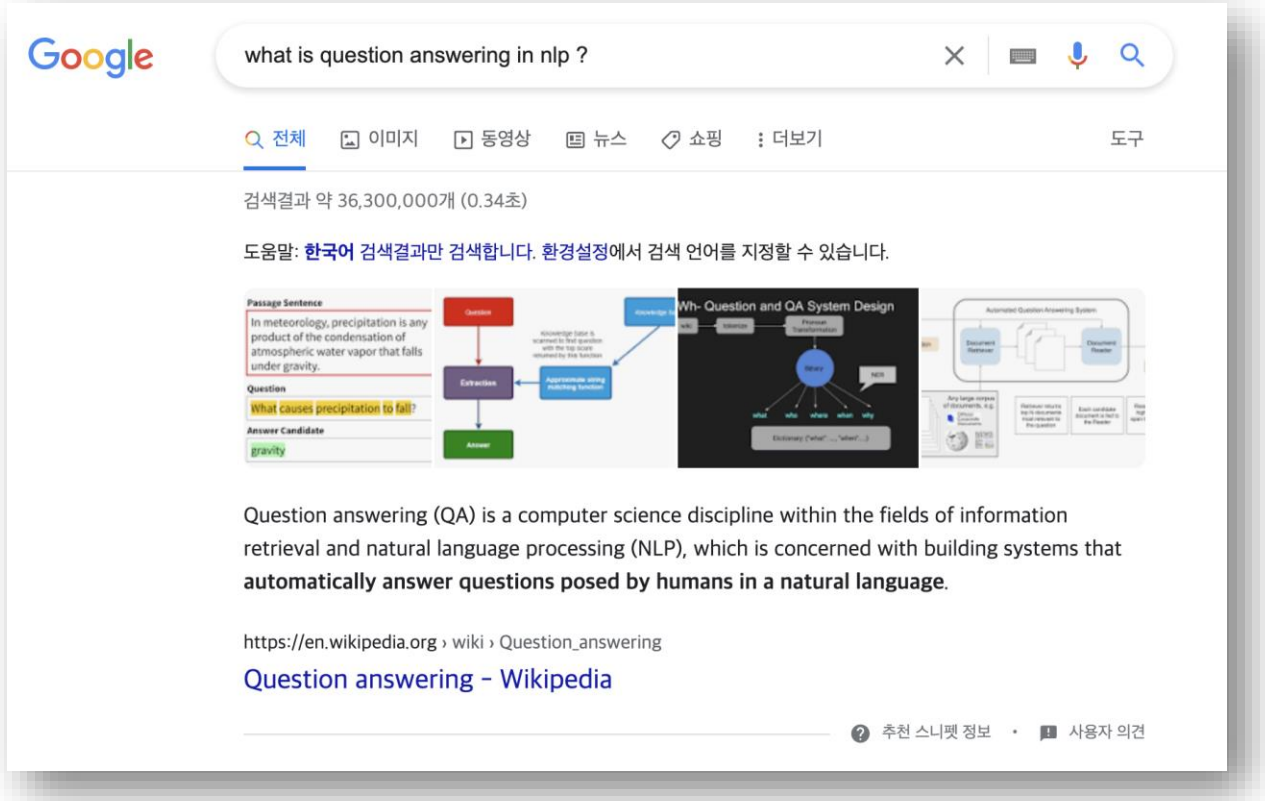
- Text Classification: Spam Detection, Sentiment Analysis, ..
- Question Answering: Search Engine
- Machine Translation
- Chatbot
- Personal Assistant
- Text Summarization



https://developers.google.com/machine-learning/guides/text-classification

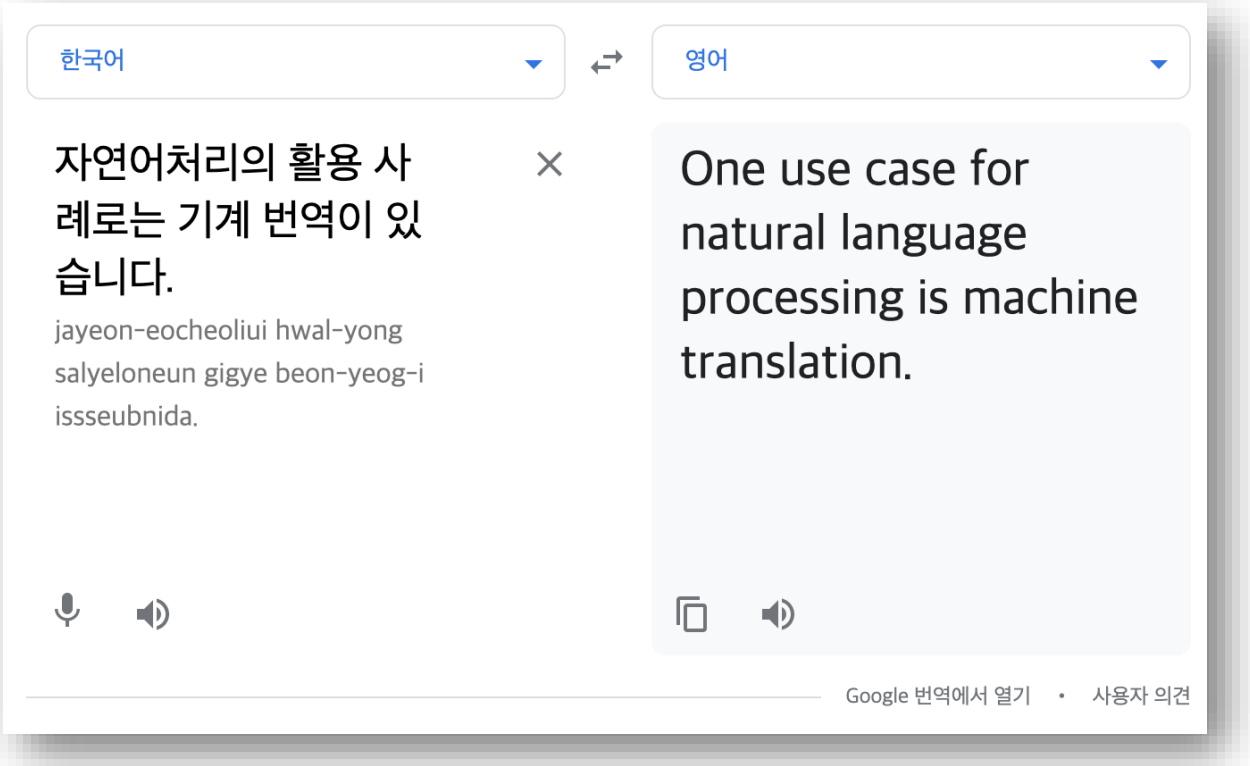
NLP Applications

- Text Classification: Spam Detection, Sentiment Analysis, ..
- Question Answering: Search Engine, ..
- Machine Translation
- Chatbot
- Personal Assistant
- Text Summarization



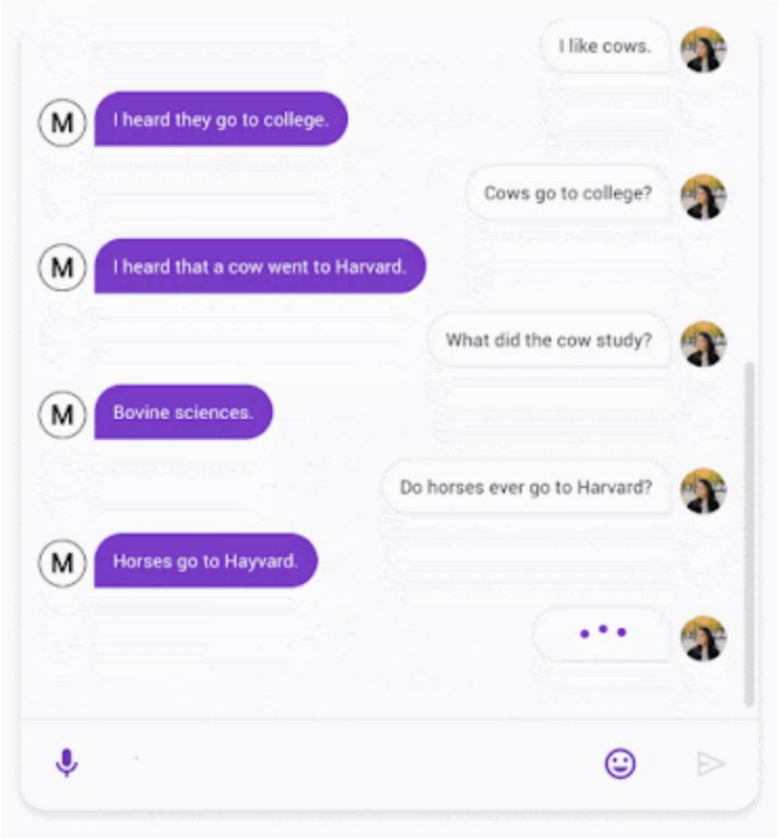
NLP Applications

- Text Classification: Spam Detection, Sentiment Analysis, ..
- Question Answering: Search Engine
- Machine Translation
- Chatbot
- Personal Assistant
- Text Summarization



NLP Applications

- Text Classification: Spam Detection, Sentiment Analysis, ..
- Question Answering: Search Engine
- Machine Translation
- Chatbot
- Personal Assistant
- Text Summarization



Meena (Google, 2020)

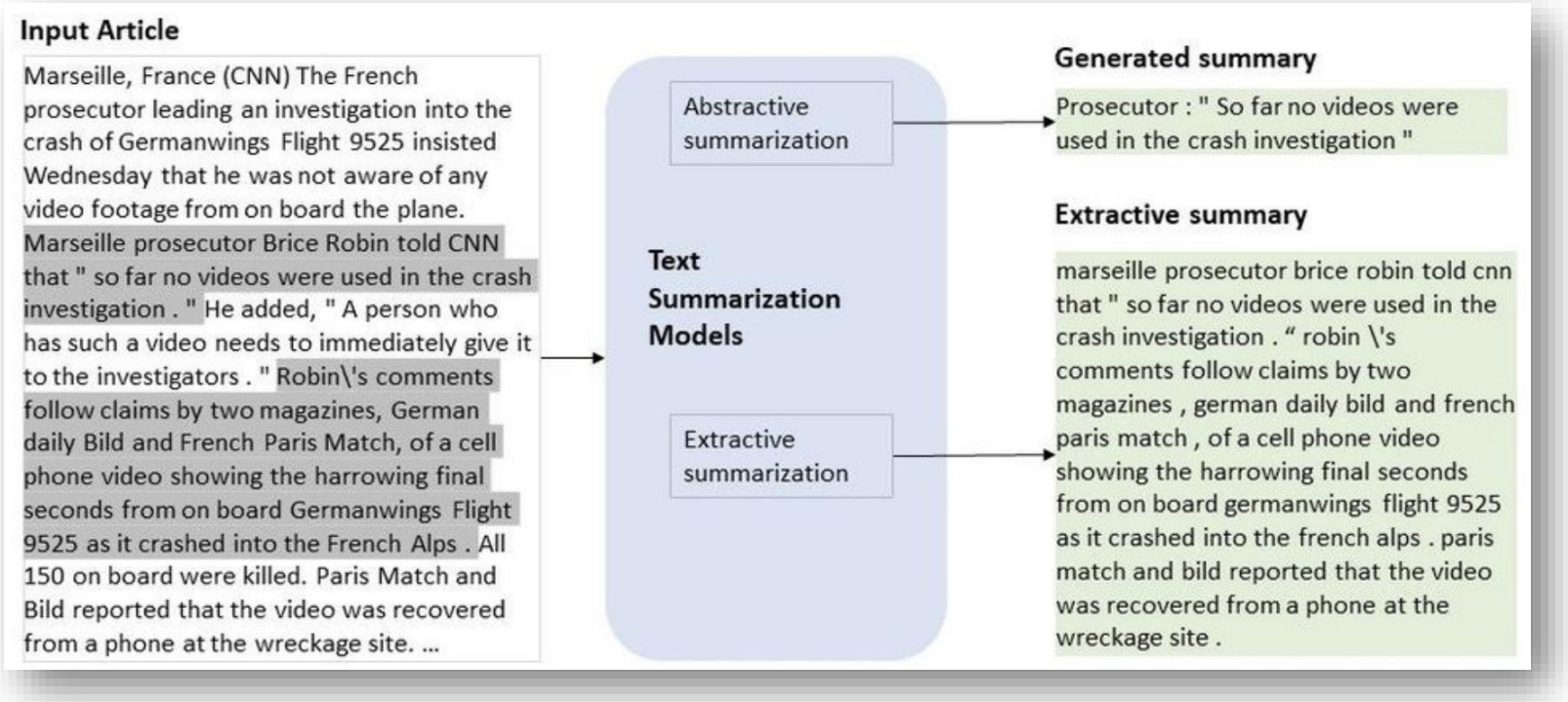
NLP Applications

- Text Classification: Spam Detection, Sentiment Analysis, ..
- Question Answering: Search Engine
- Machine Translation
- Chatbot
- Personal Assistant
- Text Summarization

Siri (Apple)

NLP Applications

- Text Classification: Spam Detection, Sentiment Analysis, ..
- Question Answering: Search Engine
- Machine Translation
- Chatbot
- Personal Assistant
- Text Summarization



https://techcommunity.microsoft.com/t5/ai-customer-engineering-team/bootstrap-your-text-summarization-solution-with-the-latest/ba-p/1268809

Academic Disciplines related to NLP

- Natural language processing (major conferences: ACL, EMNLP, NAACL)
  - Includes state-of-the-art deep learning-based models and tasks
  - Low-level parsing
    - Tokenization, stemming
  - Word and phrase level
    - Named entity recognition(NER), part-of-speech (POS) tagging, noun-phrase chunking, dependency parsing, coreference resolution
    - Semantic relation extraction
  - Sentence level
    - Sentiment analysis, machine translation,
  - Multi-sentence and paragraph level
    - Entailment prediction, question answering, dialog systems, summarization

Academic Disciplines related to NLP

- Text mining (major conferences: KDD, The WebConf (formerly, WWW), WSDM, CIKM, ICWSM)
  - Extract useful information and insights from text and document data
    - e.g., analyzing the trends of AI-related keywords from massive news data
  - Document clustering (e.g., topic modeling)
    - e.g., clustering news data and grouping into different subjects
  - Highly related to computational social science
    - e.g., analyzing the evolution of people's political tendency based on social media data

- Information retrieval (major conferences: SIGIR, WSDM, CIKM, RecSys)
  - This area is not actively studied now.
  - It has evolved into a recommendation system, which is still an active area of research.

Trends of NLP

- Text data can be considered as a sequence of words. Each word is represented as an embedding vector such as Word2Vec or GloVe.
- RNN-based models such as RNN, LSTM, and GRU deal with sequential data.
- As attention and Transformers are released, RNN is replaced with self-attention.
- Recently, self-supervised models such as BERT is trained with a large dataset and fine-tuned with various downstream task.

Course Overview

- Bag-of-Words
- Topic Modeling
- Word Embedding
- RNN
- LSTM, GRU
- RNNs with Attention
- Pre-tokenization
- Tokenization
- Transformer
- Transformer with Huggingface
- GPT-1, GPT-2, GPT-3
- BERT
- Text Classification (Encoder)
- Text Generation (Decoder)
- Machine Translation
- Quiz 1, 2, 3

NATURAL LANGUAGE PROCESSING

LECTURE 1: BAG-OF-WORDS

goorm

KAIST AI
Graduate School of AI

DAVIAN

Word Embedding

1. 벡터가 어떻게 의미를 가지게 되는가

구분	백오브워즈 가정	언어 모델	분포 가정
내용	어떤 단어가 (많이) 쓰였는가	단어가 어떤 순서로 쓰였는가	어떤 단어가 같이 쓰였는가
대표 통계량	TF—IDF	-	PMI
대표 모델	Deep Averaging Network	ELMo, GPT	Word2Vec

임베딩을 만드는 세 가지 철학

https://ratsgo.github.io/embedding/

Word Embedding

- **백오브워즈 가정**: 저자의 의도가 단어 사용 여부나 그 빈도에서 드러난다고 보는 가정.
  - TF-IDF (Term Frequency-Inverse Document Frequency): 어떤 단어의 주제 예측 능력이 강할 수록 가중치가 커지고, 그 반대의 경우 작아짐.

$$TF-IDF(w) = TF(w) \times \log(\frac{N}{DF(w)})$$

  - Deep Averaging Network (Iyyer et al., 2015): 문장에 속한 단어의 임베딩을 평균을 취해 문장의 임베딩을 만듦.

- **언어 모델**: 단어의 등장 순서를 학습해 주어진 단어 시퀀스가 얼마나 자연스러운지 확률을 부여함.

- **분포 가정**: 단어의 의미는 주변 문맥을 통해 유추해볼 수 있다고 보는 가정
  - PMI (Pointwise Mutual Information): 두 단어(A, B)가 얼마나 자주 같이 등장하는지에 관한 정보를 수치화

$$PMI(A,B) = \log \frac{P(A,B)}{P(A) \times P(B)}$$

  - Word2Vec: 특정 다긴 단어 주변의 문맥, 즉 분포 정보를 함축함.

https://ratsgo.github.io/embedding/

INDEX

- Bag-of-words representation
- Naïve Bayes Classifier

Bag-of-words Representation

- Bag-of-Words(BoW): a text is represented as the bag of its words, disregarding grammar and even word order but keeping multiplicity
- Modeling BoW
  1) Build vocabulary with unique words.
  2) Encode each word with one-hot vector.
  3) Represent text data as a sum of one-hot vectors.

Naïve Bayes Classifier

- Naïve Bayes Classifier: a simple probabilistic classifiers based on applying Bayes' theorem.
- Bayes' theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Naïve Bayes Classifier

- Consider there are C classes for each document d.
- $P(c|d)$: probability that d belongs to c
- Applying Bayes' theorem,

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

- Dropping the denominator,

$$P(c \mid d) = P(d \mid c)P(c)$$

Naïve Bayes Classifier

- d can be considered as a sequence of words $w_1, w_2, \ldots, w_n$.

$$P(d \mid c)P(c) = P(w_1, w_2, \ldots, w_n \mid c)P(c)$$

- Applying the chain rule,

$$P(d \mid c)P(c) = P(c)\Pi_{w_i \in W} P(w_i \mid c)$$

- Chain rule:

$$\begin{aligned} \mathrm{P}(X_4, X_3, X_2, X_1) &= \mathrm{P}(X_4 \mid X_3, X_2, X_1) \cdot \mathrm{P}(X_3, X_2, X_1) \\ &= \mathrm{P}(X_4 \mid X_3, X_2, X_1) \cdot \mathrm{P}(X_3 \mid X_2, X_1) \cdot \mathrm{P}(X_2, X_1) \\ &= \mathrm{P}(X_4 \mid X_3, X_2, X_1) \cdot \mathrm{P}(X_3 \mid X_2, X_1) \cdot \mathrm{P}(X_2 \mid X_1) \cdot \mathrm{P}(X_1) \end{aligned}$$

Naïve Bayes Classifier Example

- For a document d, which is a sequence of words w, and a class c,

		Document (d)	Class (c)
Training	1	me free lottery	Spam
	2	free get free you	Spam
	3	you free scholarship	Inbox
	4	free to contact me	Inbox
	5	you won award	Inbox
	6	you ticket loterry	Spam
Test	7	you free loterry	?

$$P(d \mid c)P(c) = P(c)\Pi_{w_i \in W} P(w_i \mid c)$$

$$P(c_{Spam} \mid d_7) = P(c_{Spam})P(w_{you} \mid c_{Spam})P(w_{free} \mid c_{Spam})P(w_{lottery} \mid c_{Spam})$$

$$P(c_{Inbox} \mid d_7) = P(c_{Inbox})P(w_{you} \mid c_{Inbox})P(w_{free} \mid c_{Inbox})P(w_{lottery} \mid c_{Inbox})$$

- Test document belongs to "Spam" if

$$P(c_{Spam} \mid d_7) > P(c_{Inbox} \mid d_7)$$

- Test document belongs to "Inbox" if

$$P(c_{Spam} \mid d_7) < P(c_{Inbox} \mid d_7)$$

Naïve Bayes Classifier Example

- For a document d, which is a sequence of words w, and a class c,
- We need to obtain $P(c_{Spam}|d_7) = P(c_{Spam})P(w_{you}|c_{Spam})P(w_{free}|c_{Spam})P(w_{lottery}|c_{Spam})$

$$P(c_{Inbox}|d_7) = P(c_{Inbox})P(w_{you}|c_{Inbox})P(w_{free}|c_{Inbox})P(w_{lottery}|c_{Inbox})$$

		Document (d)	Class (c)
Training	1	me free lottery	Spam
	2	free get free you	Spam
	3	you free scholarship	Inbox
	4	free to contact me	Inbox
	5	you won award	Inbox
	6	you ticket loterry	Spam
Test	7	you free loterry	?

$$P(c_{Spam}) = \frac{3}{6} = \frac{1}{2} \qquad P(w_{you}|c_{Spam}) = \frac{2}{10} \qquad P(w_{you}|c_{Inbox}) = \frac{2}{10}$$

$$P(c_{Inbox}) = \frac{3}{6} = \frac{1}{2} \qquad P(w_{free}|c_{Spam}) = \frac{3}{10} \qquad P(w_{free}|c_{Inbox}) = \frac{2}{10}$$

$$P(w_{lottery}|c_{Spam}) = \frac{2}{10} \qquad P(w_{lottery}|c_{Inbox}) = \frac{0}{10}$$

Naïve Bayes Classifier Example

$$P(c_{Spam}\,|\,d_7) = P(c_{Spam})P(w_{you}\,|\,c_{Spam})P(w_{free}\,|\,c_{Spam})P(w_{lottery}\,|\,c_{Spam}) = \frac{1}{2}\times\frac{2}{10}\times\frac{3}{10}\times\frac{2}{10} = \frac{6}{1000}$$

$$P(c_{Inbox}\,|\,d_7) = P(c_{Inbox})P(w_{you}\,|\,c_{Inbox})P(w_{free}\,|\,c_{Inbox})P(w_{lottery}\,|\,c_{Inbox}) = \frac{1}{2}\times\frac{2}{10}\times\frac{2}{10}\times\frac{0}{10} = 0$$

- Therefore, the test document belongs to "Spam".

		Document (d)	Class (c)
Training	1	me free lottery	Spam
	2	free get free you	Spam
	3	you free scholarship	Inbox
	4	free to contact me	Inbox
	5	you won award	Inbox
	6	you ticket loterry	Spam
Test	7	you free loterry	Spam