훈련과정명	인공지능 자연어처리(NLP) 기업데이터 분석 전문가 양성과정 (A반)				
교과목	인공지능 자연어 처리 이론 및 실습				
실시일		성명		점수	

번호	문제	답
1	다음 중 자연어 처리 (NLP) 영역이라 보기 어려운 것은? ① Machine Translation ② Named Entity Recognition ③ Image Captioning ④ Image Recognition	
2	자연어 처리 (NLP)를 위해서는 많은 양의 학습 데이터가 필요하다. 자연어 처리를 위한 학습용 데이터를 무엇이라 하는가? ① dictionary (딕셔너리) ② corpus (코퍼스, 말뭉치) ③ token (토큰) ④ vocabulary (어휘 사전)	
3	문장의 형태론적인 구조를 파악하기 위해서는 각 단어의 문법적 역할인 품사를 확인할 필요가 있다. 동일한 단어라도 품사에 따라 그 의미가 달라지기 때문이다. 문장에 쓰인 단어에 품사를 부여하는 것을 무엇이라 하는가? ① Chunking ② Lemmatize ③ POS tagging ④ Tokenize	
4	다음 중 단어의 다의어를 파악할 수 있고 상.하위어의 계층적 관계를 파악하는데 유용한 코퍼스는 무엇인가? ① WordNet corpus ② Brown corpus ③ Reuter corpus ④ Penn Treeback corpus	
5	다음 중 두 단어의 유사도를 확인하는 방법 중 적절하지 않은 것은? ① 편집 거리 (edit distance) 알고리즘을 이용하면 두 단어의 형태적 유사도를 측정할 수 있다. ② 워드넷 (WordNet)을 이용하여 두 단어 사이의 최단 path를 측정하	

번호	문제	답
	면 두 단어의 계층적 유사도를 측정할 수 있다. ③ TFIDF를 이용하여 거리를 측정하면 두 단어의 의미적 유사도를 측정할 수 있다. ④ Skip-Gram으로 벡터화된 두 단어의 거리를 측정하면 의미적 유사도를 측정할 수 있다.	
6	두 개의 텍스트 문서를 각각 1차원 벡터의 수치 데이터로 변환했다 (Document to Vector). 두 문서의 유사성을 측정하려고 할 때 적절하지 못한 것은? ① 두 벡터의 내적을 계산한다. ② 두 벡터의 평균을 계산한다. ③ 두 벡터의 유클리디언 거리를 계산한다. ④ 두 벡터의 코사인 값 (cos(θ))을 계산한다.	
7	다음 정규표현식 (regular expression)에 해당하지 않는 문자열은? 정규표현식 : r"^s.+g{2}r$" ① stggr ② sorggrr ③ smyunggr ④ stuvggr	
8	어떤 문장에 사용된 단어들을 one-hot 방식으로 인코딩한 후 word embedding으로 단어를 표현하려고 한다. 다음 설명 중 잘못된 것은? ① One-hot으로 인코딩된 단어들의 내적은 모두 0 이므로 모두 독립적이다. ② One-hot으로 인코딩하면 모든 단어들이 의미적으로 관계가 없는 상태로 초기화된다. ③ Word embedding으로 표현된 단어들은 서로 의미적 관계가 있다. ④ 네트워크 출력층에 one-hot 벡터를 출력할 수 있고 softmax를 사용한다. Softmax는 vocabulary가 커져도 계산량이 적어지기 때문에 사용하기가 편리하다.	
9	다음 중 주제 식별 (Topic Model)에 사용되는 알고리즘은? ① LDA ② HMM ③ WSD ④ CFG	

번호	문제	답
10	TF-IDF에 대한 설명 중 적절하지 못한 것은? ① TF는 term frequence로 문서에 사용된 단어의 빈도수를 나타낸다. ② DF는 document frequence로 단어가 사용된 문서의 개수를 의미하고 IDF는 DF의 역수이다. ③ 어떤 단어의 TF가 높을수록 그 단어는 자주 사용된 것이므로 중요하게 취급된다. ④ 어떤 단어의 DF가 높을수록 그 단어는 여러 문서에 자주 등장하는 단어이므로 중요하게 취급된다.	
11	다음 중 중요 문장 추출 (sentence extraction)에 대한 설명 중 잘못된 것은? ① 문장 간의 유사도 (similarity)를 기반으로 한다. ② 어떤 문장이 다른 문장들과 유사도가 높으면 그 문장은 중요하다고 볼 수 있다. ③ 레스크 알고리즘 (Lesk algorithm)이 사용된다. ④ TextRank 알고리즘이 사용된다.	
12	Positive와 Negative로 레이블이 부여된 영화 리뷰 데이터를 학습(지도학습)하여 리뷰어들의 감성을 분석하려고 한다. 다음 중 적절하지 못한 방법은? ① TFIDF를 이용하여 리뷰 문서를 벡터화한다. ② 워드 임베딩을 이용하여 리뷰 문서를 수치화한 후 LSTM이나 CNN을 이용한다. ③ 딥러닝을 이용하여 binary classification (이진 분류)을 수행한다. ④ LSA (Latent Semantic Analysis) 알고리즘을 이용하여 리뷰 문서들을 Positive와 Negative로 분류한다.	
13	문서에서 특정 인물의 이름, 지역 이름, 조직 이름 등을 발췌할 수 있는 기술은 무엇인가? ① WSD ② CFG ③ NER ④ LDA	

번호	문제	답
14	문장 구조에서 단어들의 모임인 구 (phrase)를 하나의 단위로 묶는 것을 무엇이라 하는가? ① Lemmatize ② Stemming ③ CFG ④ Chunk	
15	다음 중 촘스키 (Noam Chomsky)의 계층 문법에 속하지 않는 것은? ① 문맥 의존 문법 (context sensitive grammar) ② 구조 자유 문법 (structure free grammar) ③ 문맥 자유 문법 (context free grammar) ④ 정규 문법 (regular grammar)	
16	nltk 패키지를 이용하여 다음 문장 (text)의 품사를 태깅하려 한다. 다음 중 올바른 명령은? import nltk text = "Seoul is the capital of Korea" ① pos = nltk.pos_tag(text) ② pos = nltk.pos_tag(nltk.word_tokenize(text)) ③ pos = nltk.tagger(text) ④ pos = nltk.tagger(nltk.word_tokenize(text))	
17	문서 혹은 단어를 수치 데이터로 변환하는 방법에 대한 내용 중 잘못된 것은? ① 워드 인코딩 (encoding)은 빈도나 사전을 이용하여 단순히 수치로 변환하는 방식이다. 이 방법은 단어의 의미를 전혀 반영하지 못한다. ② 워드 임베딩 (embedding)은 단어의 의미를 반영하도록 수치화하는 방식이다. ③ 대표적인 워드 임베딩에는 주변 단어의 문맥을 고려한 CBOW와 Skip-Gram 방식이 있다. ④ 워드 인코딩 (encoding)은 학습을 통해 수치 데이터로 변환하는 방식이다.	

번호	문제	답
18	Keras의 Embedding layer에 대한 설명 중 적절하지 못한 것은? ① (문서 개수 x 단어 개수) 구조의 2차원 행렬로 표현된 문서를 입력 받아서 (문서 개수 x 단어 개수 x 단어의 임베딩 벡터) 구조의 3차원 텐서를 출력한다. ② 분석자가 사전에 vocabulary를 생성하기 때문에 out-of-vocabulary (OOV) 문제가 없다는 것이 큰 장점이다. ③ 철자가 동일하면서 뜻이 다른 단어 (예 : bank = 은행, 강둑)의 워드 임베딩 벡터는 동일하다. ④ Embedding layer의 파라메터 (W)는 (vocabulary의 단어 개수 x 출력될 워드 임베딩 벡터의 원소 개수)로 구성되고, W를 이용하면 vocabulary 내의 단어에 대한 임베딩 벡터를 확인할 수 있다.	
19	대표적인 Word2Vec에는 CBOW와 Skip-Gram이 있다. 다음 중 설명이 잘못된 것은? ① Skip-Gram은 특정 단어를 입력 받아 주변 단어가 나오도록 학습하고 CBOW는 주변 단어를 입력받아 특정 단어가 나오도록 학습한다. ② 학습이 완료되면 hidden layer의 출력이 해당 단어의 임베딩 벡터가 된다. ③ Skip-Gram Negative Sampling (SGNS)은 최종 출력에 softmax가 사용되므로 계산량이 적어진다. ④ Skip-Gram으로 생성한 워드 임베딩 벡터는 범용적인 것으로 특정 목적의 네트워크를 학습할 때 pre-training용으로 사용할 수 있다.	
20	TFIDF 같은 빈도 기반 방식은 문서 전체에 대한 통계를 사용하지만 단어 별 의미는 고려하지 못하는 단점이 있고, Skip-Gram 같은 학습 기반 방식은 주변 단어만을 이용하기 때문에 문서 전체를 고려하지 못하는 단점이 있다. 이 문제를 보완하기 위해 빈도 기반과 학습 기반을 혼용한 Word2Vec 기술은 무엇인가? ① GloVe ② ELMo ③ FastText ④ CBOW	

[정답]

번호	정답	번호	정답
1	4	11	3
2	2	12	4
3	3	13	3
4	1	14	4
5	3	15	2
6	2	16	2
7	2	17	4
8	4	18	2
9	1	19	3
10	4	20	1