| 훈련과정명 | 인공지능 자연어처리(NLP) 기업데이터 분석 전문가 양성과정 (C반) | | | | |
|---|---|---|---|---|---|
| 교과목 | 인공지능 자연어처리 이론 및 실습 | | | | |
| 실시일 | 2019.9.11 | 성명 | | 점수 | |

| 번호 | 문제 | 답 |
|---|---|---|
| 1 | 자연어 처리에 있어서 한국어는 영어와 달리 어려움이 많다. 다음 보기 중에서 한국어 자연어 처리의 어려움 중 2가지를 선택하시오.<br>①한국어는 지나치게 형용사 부사가 많다.<br>②한국어는 조사가 있는 교착어이다.<br>③한국어는 한자조합식 단어가 많다.<br>④한국어는 문자의 순서가 중요하다. | |
| 2 | 다음 중 한국어 형태소 분석 엔진이 아닌 것은?<br>①Kkam ②nltk ③Komoran ④Mecab | |
| 3 | 한국어로 말뭉치로 번역되는 단어. Computer를 이용해서 자연어 분석 작업을 할 수 있도록 만든 문서의 집합을 무엇이라고 하나요? | |
| 4 | 다음 중 Text Mining이나 Web Scrapping시 많이 사용하는 것으로 기준이 되는 문자열을 다른 문자열과 Pattern Matching해서 결과값을 추출 또는 변경하기 위한 기준식(expression)을 무엇이라고 하는가?<br>①조건식 ②구문식 ③형태식 ④정규표현식 | |
| 5 | 다음 중 원핫인코딩(one-hot encoding)에 대한 설명으로 틀린 것은?<br>① 원핫인코딩은 단어를 숫자로 표현하는 방법 중 하나로 unique한 단어 수를 차원수로 하는 벡터로 구성된다.<br>② 새로운 단어 수가 증가 할수록 원핫벡터의 차원수도 함께 증가한다.<br>③ 단어로 벡터로 표현하여 단어간의 유사도를 측정하기에 적합한 방법이다.<br>④ 원핫벡터를 다른 표현으로 희소벡터(sparse vector)라고 한다. | |
| 6 | 분석에 사용하기 위한 문장들은 노이즈 데이터, 불용어 같은 분석에 필요없는 데이터가 들어있는 경우가 일반적이다. 그래서 반드시 Model에 해당 데이터를 넣기 전에 반드시 처리해야 하는 작업을 무엇이라고 하는가?<br>①불용어 처리 ②전처리 ③모델 처리 ④데이터 처리 | |

| 번호 | 문제 | 답 |
|---|---|---|
| | | |
| 7 | 다음은 무엇을 설명하고 있는가?<br>설명 : 단어들을 수치화 한 후 이를 기반으로 단어들 사이의 거리를 계산해서 문서 간의 단어들의 차이를 계산하는 것<br><br>①문서 유사도 ②토픽 모델링 ③연관분석 ④워드 임베딩 | |
| 8 | 전처리 과정 중에서 각 단어의 품사가 명사, 동사, 형용사 인지를 알아내는 작업을 무엇이라고 하는가?<br>①단어 유형 Tagging ②유사도 Tagging<br>③품사 Tagging(POS Tagging) ④연관관계 Tagging | |
| 9 | 다음 중 단어의 빈도수를 기준으로 글씨의 크기를 결정해서 시각화를 하는 개체를 무엇이라고 하는가?<br>①워드 클라우드 ②산점도 ③파이플롯 ④워드 그래프 | |
| 10 | 다음과 같은 수치들이 측정되었다. 셔츠를 구매한 사람이 넥타이를 구매할 확률 즉, 향상도의 값을 구하시오.(단, 소수점 2째 자리까지 처리할 것)<br>1)셔츠 구매건수 : 3번<br>2)넥타이 구매건수 : 2번<br>3)전체 거래건수 : 4번<br>4)동시 거래건수 : 2번 | |
| 11 | 이 알고리즘은 단어분리(Subword Segmentation) 할 때 일반적으로 사용하는 알고리즘으로, 기계가 모르는 단어가 있을 때 발생하는 OOV문제를 해결하기 위한 알고리즘이다. 처음 용도는 압축 알고리즘이었으나 자연의 처리에서 단어 분리 알고리즘으로 기본적으로 사용하고 있는 이것은 무엇인가요?<br>①WPE ②N-gram ③BPE ④WPM | |
| 12 | 일반적으로 Computer는 글자를 처리하기 보다 숫자를 더 잘 처리한다. 더구나 자연어 처리에서는 분석을 위해 반드시 문자열을 숫자로 변환하는 작업을 해야 하는데, 여러 가지 방법이 있지만, 그 중에서도 N차원의 매트릭스를 생성해서 숫자로 변환하는 이 알고리즘을 무엇이라고 하는가?<br>①One-hot Encoding ②Label Encoding<br>③Decoding ④정수 Encoding | |
| 13 | | |

| 번호 | 문제 | 답 |
|---|---|---|
| | 영미권에서는 stopword로 불리는 것으로, 분석 시 큰 의미가 없는 단어로서 보통 전처리 과정에서 제거한다. 이런 단어를 무엇이라고 하는가? | |
| 14 | 전처리 과정에서 다루는 것으로 Token화가 된 단어들의 접사(affix)를 제거하여 같은 의미를 형태소의 기본형을 찾는 방법을 말한다. 다음 중 이런 작업을 할 수 있도록 NLTK에서 제공하는 있는 개체를 고르시오.<br>①PorterTokenizer ②Text ③Corpus ④PorterStemmer | |
| 15 | 다음 중 정규표현식 연산자 중 틀린 것은?<br>① * : 0번 이상 발생을 의미<br>② + : 1번 이상 발생을 의미<br>③ $ : 문자열이나 행의 끝을 의미<br>④ ^ : 한 자리의 문자 | |
| 16 | 다음 중 자연어 전처리 단계에 해당하지 않는 것은?<br>① 토큰화<br>② 품사태깅<br>③ 감성분석<br>④ 불용어처리 | |
| 17 | 1. 다음은 source.txt에 담겨있는 데이터들이다.<br>1cake - Right<br>jelly 12hey - Wrong<br>maybe12 - Wrong<br>3 joy - Wrong<br>4432 - Right<br>23b - Right<br>5555b – Wrong<br><br>2. 위의 데이터들을 기준으로, 아래와 같이 숫자로 시작하고 Right로 끝나는 출력결과물이 나올 수 있도록 하는 가장 적절한 표현식을 고르시오.<br>1cake - Right<br>4432 - Right<br>23b – Right<br><br>①r'(^[0-9]+₩w*) - (Right$)'<br>②r'(^(₩w)+₩w?) - (₩w....)$' | |

| 번호 | 문제 | 답 |
|---|---|---|
| | ③r'(Right$) - ([A-Za-z0-9]+)'<br>④r'₩w{3,4} - ₩w{5})' | |
| 18 | 다음 중 Python에서 기계학습을 할 수 있도록 만든 Module을 고르시오.<br>①Pandas ②NumPy ③Matplotlib ④Scikit-learn | |
| 19 | 다음 중 BeautifulSoup에서 파싱할 때 사용하는 파서가 아닌 것은?<br>①lxml ②html.parser ③html5lib ④myparser | |
| 20 | 다음 중 품사태깅(PoS Tagging)에 대한 설명으로 틀린 것은?<br>① 품사태깅은 각 토큰에 품사정보를 부착하는 작업이다.<br>② 품사태깅은 분석시 불필요한 품사를 제거하기 위해 사용된다.<br>③ 품사태깅은 품사를 기준으로 구문 분석하기 위해 사용된다.<br>④ 품사태깅은 품사 빈도를 측정하여 문서의 의미를 파악할 수 있다. | |

**[ 정답 ]**

| 번호 | 정답 | 번호 | 정답 |
|---|---|---|---|
| 1 | ②, ③ | 11 | ③ |
| 2 | ② | 12 | ① |
| 3 | Corpus, 코퍼스 | 13 | 불용어 |
| 4 | ④ | 14 | ④ |
| 5 | ③ | 15 | ④ |
| 6 | ② | 16 | ③ |
| 7 | ① | 17 | ① |
| 8 | ③ | 18 | ④ |
| 9 | ① | 19 | ④ |
| 10 | 1.34 | 20 | ④ |