

18 - 1 인공지능 과제 2

Clustering & Word Embedding

조수필 조교

jessay@hanyang.ac.kr

카카오톡 ID : jessay

Word Embedding

- Word Embedding : 1차원 단어-> n차원 벡터로 변환.

- 초기 one-hot-encoding 방식

- ex.

문장이 [인공지능, 연구실, 파이팅] 이고,

이에 대한 사전을 '인공지능' = [1 0 0] , '연구실' = [0 1 0] , '파이팅' = [0 0 1]

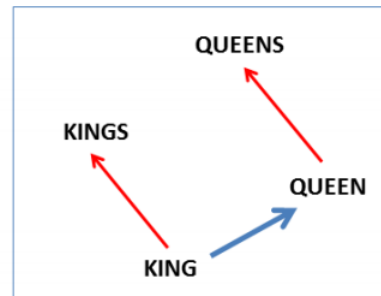
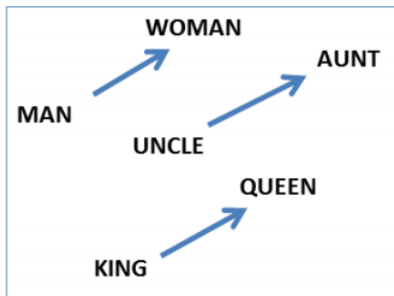
으로 만들었다면,

[인공지능, 연구실, 파이팅] → [[1 0 0] , [0 1 0] , [0 0 1]] 로 변환 가능.

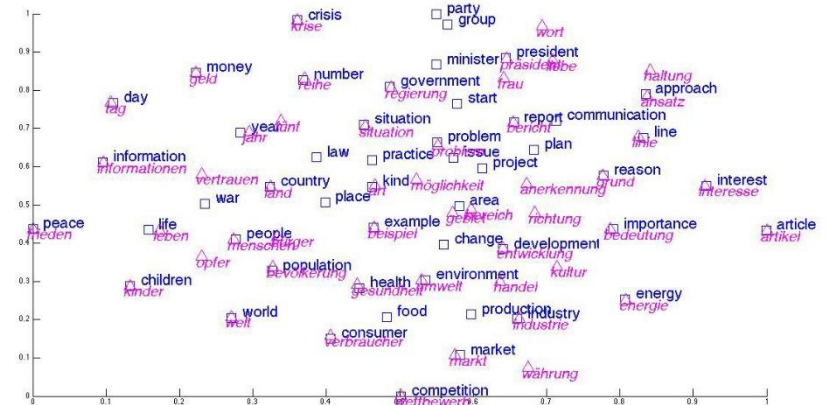
- 단어 간의 의미론적 차이, 연관 관계를 이해할 수 없다는 문제점

Word Embedding

- Word Embedding to Vector
 - 단어의 **문맥상 의미** 자체를 다차원 공간에 '벡터화' (Continuous Word Embedding)
 - 기법 : NNLM, RNNLM, **CBOW**, **Skip-gram**
 - **Ex) KING + (WOMAN – MAN) = Queen**





(Mikolov et al., NAACL HLT, 2013)



과제 목표

- GoogleNews corpus(말뭉치)를 통해 학습된 300차원의 Word2Vec Embedding Vector가 실제로 단어의 의미에 맞게 분포되었는지 파악하기!
- Complete Link Clustering을 이용하여, 과제에 주어진 단어 338개의 Word Embedding Vector를 Clustering 진행.
 - Similarity 는 cosine similarity & euclidean similarity를 사용.


$$\text{sim}(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$


$$d(x, y) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2}$$

- Clustering이 끝났을 때, 주어진 threshold 로 cluster를 분할한 다음 각각의 cluster 에 대해 분석함.

과제 목표

- Complete-Link Clustering 을 이용한 Word Embedding 분석 수행.
- 입력 : Word2vec embedding vector file

WordEmbedding.txt

```
1 secret
2 -4.12597656e-02,2.25585938e-01,2.60009766e-02,-2.10571289e-03
3 confidential
4 -0.14257812,-0.00323486,-0.02075195,-0.24511719,-0.16308594,0
5 controversial
6 4.00390625e-02,1.13281250e-01,2.96875000e-01,-8.85009766e-04,
7 underground
8 -0.19433594,0.02941895,0.12695312,0.03540039,-0.03979492,0.05
9 cover
10 0.09863281,0.13574219,0.05810547,-0.06396484,0.18847656,0.004
11 0.1111
```

- 단어 338 개
- 단어 당 300차원 embedding vector 존재
- 지수 표현과 실수 표현이 섞여 있으니 주의할 것.

과제 목표

- 출력 : cluster number를 첨부한 Word2vec embedding vector file

WordClustering.txt

 : cluster number

```
1 secret
2 -4.12597656e-02,2.25585938e-01,2.60009766e-02,-2.10571289e-03,-2.571
3 3
4 confidential
5 -0.14257812,-0.00323486,-0.02075195,-0.24511719,-0.16308594,0.04150
6 2
7 controversial
8 4.00390625e-02,1.13281250e-01,2.96875000e-01,-8.85009766e-04,-9.948
9 3
10 underground
11 -0.19433594,0.02941895,0.12695312,0.03540039,-0.03979492,0.05883789,
12 1
13 cover
14 0.09863281,0.13574219,0.05810547,-0.06396484,0.18847656,0.0045166,0
15 4
```

클러스터링 평가 방법 예시

- Entropy

Information Gain: Example

■ Data Set

Gender	Car Owner Ship	Travel Cost(\$)/km	Cost Income Level	Transportation
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Cheap	Medium	Train
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car

Target Variable

■ Data Set impurity

$$Entropy = -0.4\log(0.4) - 0.3\log(0.3) - 0.3\log(0.3) = 1.571$$

4B,3C,3T

Gender	Class
Male	Bus
Male	Bus
Male	Bus
Male	Car
Male	Train

3B,1C,1T

$$Entropy = -\frac{3}{5}\log(\frac{3}{5}) - \frac{1}{5}\log(\frac{1}{5}) - \frac{1}{5}\log(\frac{1}{5}) = 1.371$$

Gender	Class
Female	Bus
Female	Car
Female	Car
Female	Train
Female	Train

1B,2C,2T

$$Entropy = -\frac{1}{5}\log(\frac{1}{5}) - \frac{2}{5}\log(\frac{2}{5}) - \frac{2}{5}\log(\frac{2}{5}) = 1.522$$

■ Information Gain of attribute Gender

$$1.574 - (5/10 * 1.371 + 5/10 * 1.522) = 0.125$$

- 각 Cluster 별 entropy를 측정 후, weighted sum을 통해 전체 엔트로피 계산.
- WordTopic.txt 자료 참조! (각 단어의 "class"가 기재되어 있음)
ex) [curiosity] = { secret , confidential , forbidden, agenda, ... }

클러스터링 평가 방법 예시

- 실루엣 지표

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

- i 는 하나의 개체(item)
- $a(i)$ 는 같은 클러스터에 속한 요소들과 i 의 거리의 평균
- $b(i)$ 는 i 번째 개체와 다른 클러스터에 속한 요소들 간 거리들의 평균을 클러스터마다 각각 구한 뒤, 이 가운데 가장 작은 값
- 각 cluster의 응집도를 파악 가능.

과제 수행 조건

- WordEmbedding.txt 에서 단어 별 Embedding Vector를 가져옴.
- complete link clustering 을 수행. (**외부 라이브러리 사용 불가!**)
 - Clustering 시 사용하는 similarity는 Cosine, Euclidean 을 각각 사용!
- 해당 결과를 출력 형식에 맞게 파일로 저장 (파일명 : WordClustering.txt)
- Complete link clustering을 마친 후,

similarity threshold = 0.2 , 0.4 , 0.6 , 0.8 로 하여 cluster 분할 진행.

과제 수행 조건

- Clustering 분석 결과(ex : 엔트로피, 실루엣, 기타 등등...) 를 콘솔에 출력할 것.
 - **엔트로피는 필수적으로 분석할 것!** 실루엣 or 기타 분석은 개인의 판단!
 - 개인이 설정한 지표 or 분석 또한 중요하게 평가 진행함.
 - 출력 양식은 중요 X, 대신 가독성 있게 출력할 것!
 - 해당 값은 보고서에 함께 기재 후, 보고서에서 자세히 분석할 것!
- 분석 결과를 통해, 어떤 threshold와 similarity가 좋은지 판단해 볼 것.

과제 수행 조건

➤ Clustering 분석 결과(ex : 엔트로피, 실루엣, 기타 등등...) 를 콘솔에 출력할 것.

- **엔트로피는 필수적으로 분석할 것!** 실루엣 or 기타 분석은 개인의 판단!
- 해당 값은 보고서에 함께 기재 후, 보고서에서 자세히 분석할 것!

➤ 예시 :

	Cosine Similarity	Euclidean Similarity
Threshold = 0.2	엔트로피 : --- OOO : ---	엔트로피 : --- OOO : ---
Threshold = 0.4	엔트로피 : --- OOO : ---	엔트로피 : --- OOO : ---
Threshold = 0.6	엔트로피 : --- OOO : ---	엔트로피 : --- OOO : ---
Threshold = 0.8	엔트로피 : --- OOO : ---	엔트로피 : --- OOO : ---

과제 조건

- 사용 언어 : C/C++ , JAVA , PYTHON
- 보고서에 컴파일 방법과 사용 버전을 명시해야 한다.
- 외부 라이브러리 사용 불가 (pip, maven 등)

제출 사항

- 코드 , exe 파일 , WordClustering.txt (Python은 exe 파일 제외)
(파일명 : ex) assignment2_2014000000.c & .exe)
- 과제 보고서 1개 (파일명 : assignment2_2014000000.docx or .hwp)
 - 코드 설명
 - 실험 결과
 - 해당 코드에 대한 컴파일 방법과 사용 버전을 명시
- 점수 비중 : 코드 70% 보고서 30%

과제 주의사항

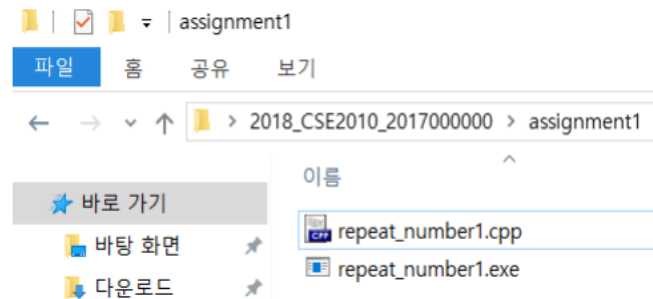
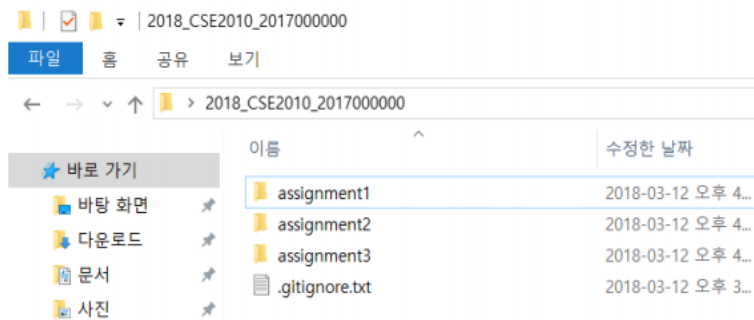
- 코드 및 **exe** 파일 & 보고서 1개 를 제출할 것
- 출력 형식을 **반드시** 준수할 것!
- 파일명을 **반드시** 준수할 것! (ex) assignment2_2014000000.c)
- 파일은 GitLab에 올릴 것!(경로 주의해주세요)
- 기한 : 18/5/31 까지. 추가 제출 X.

학번

과제 주의사항

- 파일은 GitLab에 올릴 것!
 - 경로 : (GitLab init 경로) – (assignment2) – [파일]
 - 파일명 : ex) assignment2_2014000000.c
 - GitLab 업로드 시, **빈 디렉토리가 존재하지 않도록 할 것!**

- 프로젝트는 아래 그림과 같이 관리!
 - 소문자 assignment1, assignment2, ... 로 폴더를 만들어 과제 제출



Thank you!
