**자연어처리론 숙제 4**

Skipgram을 이용한 단어 유사도

**Description**

단어임베딩 기법 중 Skipgram 알고리즘을 이용하여 단어 벡터를 학습하고, 학습된 단어 벡터를 이용하여 입력 단어와 가장 유사한 단어를 찾아주는 프로그램을 작성하시오.

**제공되는 프로그램**

- 숙제 4는 단어 벡터 학습에 사용할 수 있는 skipgram.py 파이썬 프로그램을 제공합니다. 이 프로그램은 Tensorflow 딥러닝 프레임워크를 이용하여 skipgram 알고리즘을 구현하였고 학습 코퍼스를 지정하면 단어 벡터를 학습할 수 있습니다. 속도향상을 위하여 학습은 Noise-Contrasive Estimation Training 기법을 이용하여 진행됩니다.

전체 프로그램은 다음과 같이 5개 Step으로 구성됩니다.

**Step 1.** 학습 코퍼스 로딩 및 단어 Tokenization.

**Step 2.** 단어 사전 구성 및 word -> word\_id 변환.

**Step 3.** 학습 과정에 사용할 mini-batch examples 생성.

**Step 4.** Skipgram 모델 구현을 위한 Tensorflow Graph 구성.

**Step 5.** 학습 실행 및 학습 결과 저장 (Tensorflow checkpoint)

**제공되는 학습 코퍼스 및 질의 단어**

1. **Text8 코퍼스 (영어) (96MB)**

* 파일명: text8.txt
* 다운로드 주소:
* <https://drive.google.com/open?id=1wGTovpEwX3Dm8vIMSqK-fP6YtMIE39xu>

1. **나무위키 덤프 코퍼스 (한국어 형태소 단위) (818MB)**

* 파일명: morphs\_namu\_small.txt
* 다운로드 주소:
* <https://drive.google.com/open?id=1L03WsLI33aBTLQMOspIJRaAiCboCMFZS>

1. **Text8 코퍼스 질의 단어**

* query\_words\_text8.txt (Skipgram.zip 파일에 포함됨)

1. **나무위키 질의 단어**

* query\_words\_namu.txt (Skipgram.zip 파일에 포함됨)

**To-do 리스트**

**Task 1: text8 코퍼스 단어 유사도 구하기**

* skipgram.py 프로그램으로 **text8 코퍼스**의 단어 벡터를 학습하시오.
* (num\_steps 값을 10~50만으로 설정)
* 단어 유사도 계산을 위한 Tensorflow operation을 작성하고 질의 단어 **query\_words\_text8.txt** 파일에 등록된 단어들에 대하여 각각 가장 유사한 단어 8개를 계산하여 **result\_text8.txt**에 출력하시오.

**Task 2: 나무위키 코퍼스 단어 유사도 구하기**

* skipgram.py 프로그램으로 **나무위키 코퍼스**의 단어 벡터를 학습하시오.
* (num\_steps 값을 100만으로 설정)
* 단어 유사도 계산을 위한 Tensorflow operation을 작성하고 질의 단어 **query\_words\_namu.txt** 파일에 등록된 단어들에 대하여 각각 가장 유사한 단어 8개를 계산하여 **result\_namu.txt**에 출력하시오.

**입출력 파일 예시**

**입력 파일 예시 (\*그림1 참고)**

query\_word#1

query\_word#2

…

query\_word#20

**출력 파일 예시 (\*그림2 참고)**

query\_word#1

similarity\_score similar\_word#1

similarity\_score similar\_word#2

similarity\_score similar\_word#3

similarity\_score similar\_word#4

similarity\_score similar\_word#5

similarity\_score similar\_word#6

similarity\_score similar\_word#7

similarity\_score similar\_word#8

query\_word#2

similarity\_score similar\_word#1

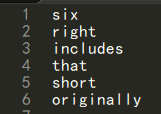


그림 1. query\_words\_text8.txt 입력파일 예시

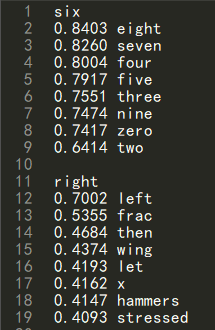


그림 2. result\_text8.txt 출력파일 예시

**실험환경**

* **Python 3.x (한국어 UTF8 처리가 용이함)**
* **Tensorflow 1.x (GPU 버전이 아니라도 상관없음)**
* **Tensorflow를 사용하는 관계로 프로그램 작성은 Python으로 해주시기 바랍니다.**

**제출 일자:** 5월 21일(월) 자정(AM 12:00)까지 late 없음

**제출 방법:** 작성한 프로그램의 소스와 result\_text8.txt, result\_namu.txt 파일을 (\*데이터파일 제외) 압축하여 아래의 메일로 제출

[rhkdgh412@gmail.com](mailto:rhkdgh412@gmail.com)

작성한 코드에 대한 간단한 설명이 포함된 파일을 포함하여야 함.

**기타:** 문의 사항이 있는 경우 위의 메일 주소로 문의하거나 R908호에 조교(허광호)를 찾아오시기 바랍니다.

**참고자료**

**Tensorflow 기본**

<http://learningtensorflow.com/lesson2/>

**Tensorflow 작동원리**

<https://www.tensorflow.org/versions/r1.2/get_started/mnist/mechanics>

**Tensorflow를 이용한 word embedding**

<http://adventuresinmachinelearning.com/word2vec-tutorial-tensorflow/>

<https://www.tensorflow.org/tutorials/word2vec>