Word2vec C++버전은 linux에서 makefile을 통해 사용해야 하기 때문에 window에서 사용할 수 있는 python버전 Gensim을 선택하였습니다. (Gensim install할 때 오류가 많이 생겨서 시간이 오래 결렸지만 아래 방법대로 하면 오류 없이 잘 진행될 것입니다.)

Gensim is known to run on Linux, Windows and Mac OS X and should run on any other platform that supports Python 2.6+ and NumPy. Gensim

**Install**

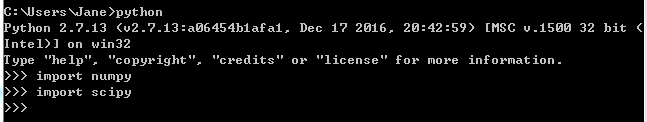
1. Install Gensim 전에 먼저 Install numpy+mlk & scipy 를 해야안다. **(python 2.7 64bit)**

방법 <https://www.youtube.com/watch?v=SrRh1V1Z748>

다운로드 사이트 http://www.lfd.uci.edu/~gohlke/pythonlibs/

(주의: Windows에서는 pip로 install하면 안됨! Install 할때 x32, x64를 잘 보고 다운받기)

실행했을 때 다음과 같이 error없으면 성공!



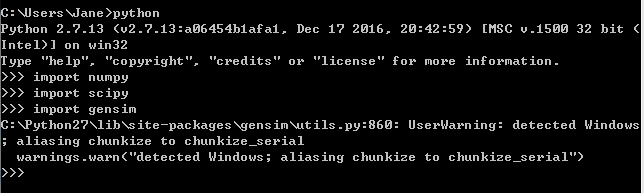
1. Install Genism

방법은 1번과 같다.

다운로드 사이트 <https://pypi.python.org/pypi/gensim#downloads>

(주의: Windows에서는 pip로 install하면 잘 안됨!)

실행했을 때 다음과 같이 error없으면 성공!



**중국어 단어 Similarity** (Gensim-word2vec) 아래 링크 참고

<http://blog.csdn.net/jerr__y/article/details/52967351>

<http://blog.csdn.net/xiaoquantouer/article/details/53583980>

(고급) http://blog.csdn.net/qdhy199148/article/details/51754631

**Word2Vec함수 parameter**

model = Word2Vec ( sentence, size = 400, window = 5, min\_count = 5, workers = multiprocessing.cpu\_count() )

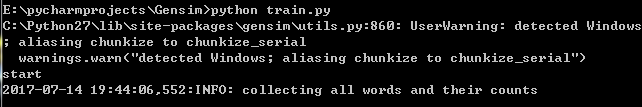
|  |  |
| --- | --- |
| sentence | 트레이닝 할 파일(형태소 분석기를 통해 분사 처린 된 파일) |
| size | 매개 단어의 vector dimensiom |
| window | scan크기, window=5이면 앞5개단어와 뒤5개단어만 생각함 |
| min\_count | 최저빈도, min\_count=5이면 트레이닝 파일 중 5번보다 적게 나타나면 버린다 |
| workers | 트레이닝 프로세스 개수 |

**train.py: word2vec 모델 트레이닝 하기, model저장하기**

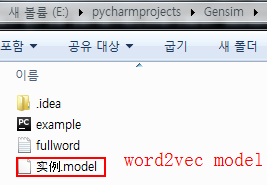
입력: fullword.txt / fullnoun.txt

출력: fullnoun.model word2vec model

실행 예시: python trian.py



결과:



**sim.py: training한 model을 통해 단어들의 similarity구하기**

입력: fullnoun.model

출력: 단어들의 유사도 계산

실행 예시: python sim.py

C:\Users\Jane\AppData\Local\Temp\1499687288(1).png