이미 트레이닝한 genism word2vec 모듈로 유사도 구하기 <http://spaces.ac.cn/archives/4304/> 참고

**모듈소개:**

|  |  |
| --- | --- |
| 트레이닝 데이터 | 위쳇 공식계정(카카오플러스 친구) 문장, 여러영역을 포함 |
| 트레이닝 데이터 크기 | 800만편좌우, 단어수 650억 |
| 모듈내 단어 수 | 352196개, 거의 중국어 단어, 소수의 자주사용하는 영어 단어 |
| 모듈 구조 | Skip+Gram + Huffman Softmax |
| Vector Dimension | 256 |
| 트레이닝 시간 | 서버에서 7일 |
| 기타 | Window=10, min\_count=10, iter=10 |
| 트레이닝 코드 | word2vec = gensim.models.word2vec.Word2Vec(sentences(), size=256, window=10, min\_count=64, sg=1, hs=1, iter=10, workers=25) |

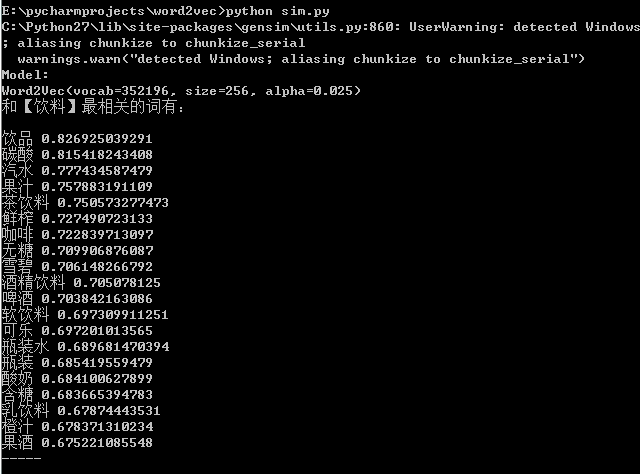
사용방법:

1. Python2.7 64bit (numpy, scipy, genism, pandas)
2. word2vec\_wx, word2vec\_wx.syn1neg.npy, word2vec\_wx.syn1.npy, word2vec\_wx.wv.syn0.npy 와 sim.py를 같은 파일에 넣는다.

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

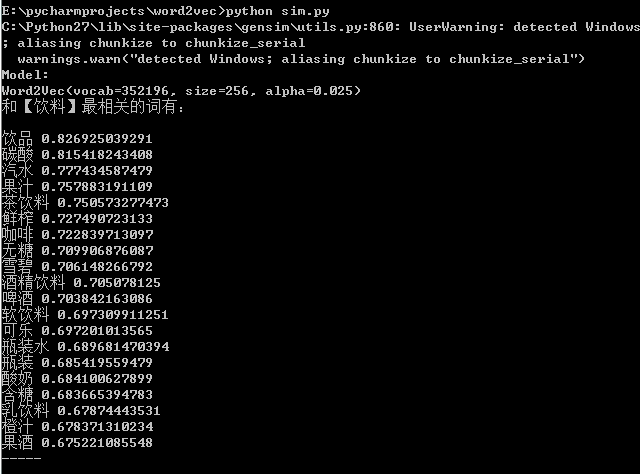
출력: 단어들의 유사도 계산

실행 예시: python sim.py



결과 1:

음료와 가장 유사한 단어 20개를 프린트 한다



음료

1. 음료
2. 탄산
3. 사이다
4. 주스
5. 차음료
6. 신선하게 짠
7. 커피
8. 무설탕
9. 쒜삐(중국사이다 브랜드명)
10. 알코올 음료
11. 맥주
12. 소프트 드링크
13. 콜라
14. 병포장 생수
15. 병포장
16. 요쿠르트
17. 설탕함유
18. 우유음료
19. 오렌지 주스
20. 과일 술

결과 2:

|  |  |  |  |
| --- | --- | --- | --- |
| **饮料+男性 음료+남성** | | **饮料+女性 음료+여성** | |
| 饮品 음료 | 0.665293216705 | 饮品 음료 | 0.69562458992 |
| 女性 여성 | 0.659904241562 | 男性 남성 | 0.658503770828 |
| 碳酸 탄산 | 0.658678352833 | 碳酸 탄산 | 0.644434928894 |
| 酒精饮料 알코올 음료 | 0.658015191555 | 咖啡 커피 | 0.610902011395 |
| 烈性酒 독한 술 | 0.622029542923 | 酒精饮料 알코올 음료 | 0.607605040073 |
| 咖啡因 카페인 | 0.607005953789 | 女性朋友 여성 친구 | 0.604772090912 |
| 软饮料 소프트 드링크 | 0.605351448059 | 茶饮料 차음료 | 0.597746789455 |
| 咖啡 커피 | 0.605114221573 | 含糖 설탕함유 | 0.596823334694 |
| 汽水 사이다 | 0.602664232254 | 果汁 쥬스 | 0.594695448875 |
| 酒类 주류 | 0.595391511917 | 无糖 무설탕 | 0.588315844536 |

|  |  |
| --- | --- |
| **饮料+年轻 음료+젊음** | |
| 饮品 음료 | 0.654176712036 |
| 咖啡 커피 | 0.621722042561 |
| 碳酸 탄산 | 0.619032800198 |
| 喝 마시다 | 0.601421713829 |
| 可乐 콜라 | 0.599813342094 |
| 汽水 사이다 | 0.595042645931 |
| 喝咖啡 커피 마시다 | 0.592226266861 |
| 果汁 쥬스 | 0.588625788689 |
| 茶饮料 차음료 | 0.572879076004 |
| 鲜榨 신선하게 짠 | 0.568604707718 |