Doc2vec을 활용한 기억 속 영화 검색

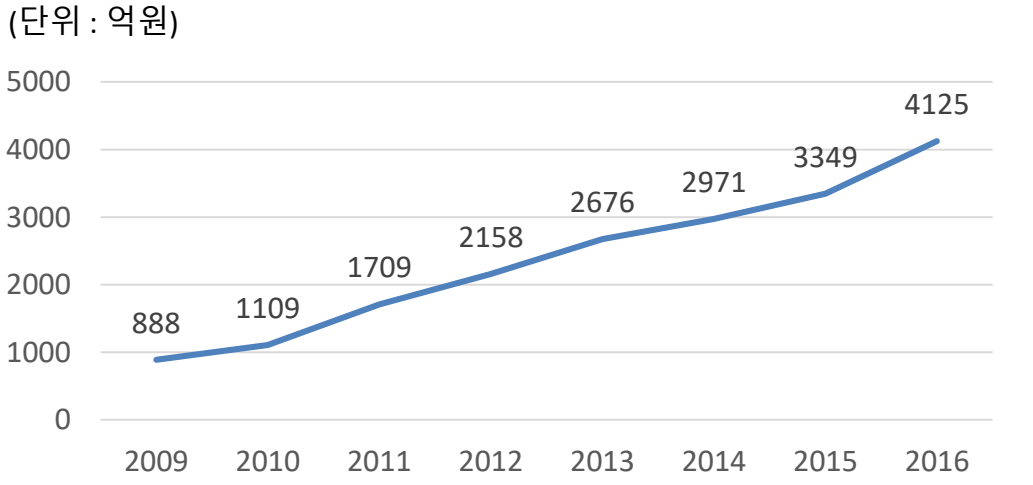
2017. 11. 17

장진규, 박정인 (울산과학기술원 융합경영대학원)
정윤혁 (울산과학기술원 경영학부)

Introduction

- **성장하는 맞춤형 콘텐츠 시장**
  - 디지털 온라인 시장 매출은 매년 평균 25% 성장 중
  - 축적된 데이터 기반으로 개인화 마케팅으로 발전

전체 디지털 온라인 시장 매출 규모



출처 : 영화진흥위원회

Introduction

- **데이터 기반 추천 알고리즘**
  - **협업 필터링**

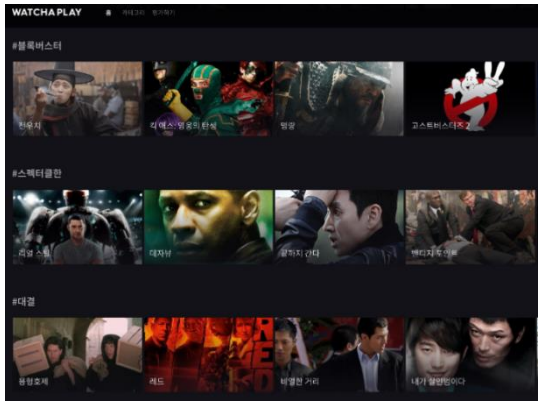
    사용자의 평가 내역을 이용한 비슷한 선호도의 사람들이 선택한 것을 추천

    선호하는 영화를 바탕으로 사용자 간의 연관성 파악
  - **콘텐츠 기반 필터링**

    영화의 특징을 분석하여 비슷한 특징을 가진 영화 추천

    Clustering, Neural Network

넷플릭스



왓챠플레이

Introduction

- ‘그 영화 제목이 뭐 였지?’
  - 찾고 싶은 영화의 제목, 배우 이름 등 구체적인 정보가 떠오르지 않는 경우
  - 지인에게 물어보거나, 인터넷 검색 이용
  - 정제되지 않은 언어를 이용한 검색의 어려움

좀비영화제목좀..

비공개 | 질문 119건 | 질문마감률 100% | 질문채택률 96% | 2017.11.04. 13:33 | 조회수 66

어떤 동굴에서 초록색가스가 나와서 감염되고 주인공들이 그 구멍으로 가려는데 그쪽 군대는 막 안된다고 하고 등장인물은 카메라맨하고 인터뷰 나온 여자와 발굴하는 아빠랑 딸 이렇게 나와요 영화제목이 뭐죠?

영화제목찾기

비공개 | 질문 12건 | 질문마감률 90% | 질문채택률 70% | 2017.11.14. 23:29 | 조회수 33

영화제목찾기
80년대 (혹은 70년대)영화로 기억하는데요 스파이액션영화쯤입니다. 미국영화구요 미국내 특수부서(존재 지않는것으로 되어있는)인물의 액션영화입니다
정확히는 기억나지 않는데
포트녹스인가요 그 미국금보관시설요 거기에서 한요원이 미국정규군과 액션을 펼치는 장면이 마무리장면 쯤이었습니다.
거의끝부분에 그 요원의 스승 노인이 물위를 달리는 장면이 인상깊죠-한국인이라는 설정이었던거 같아요 혹시 제목 기억나시면 답부탁드려요

영화 찾아주세여

비공개 | 질문 4건 | 질문마감률 100% | 질문채택률 100% | 2017.04.27. 09:35 | 조회수 177

몇년도영화인지는 모르겠고, 10대(?) 여자주인공이 나오는 영화였고요. 외국영화였습니다
주인공이 공책이나 건물 바닥,또는 계단에 그림을 자주 그렸던것같아요..추상적으로??(기억은 잘 안나지만)
여자애가 그리는 그림체 느낌이 약간 팀버튼스러웠어요.
영화의 분위기는 전체적으로 삭막한 브라운계열이였고요, 애니메이션 아니였습니다.
cg가 많아보이긴 했어요.

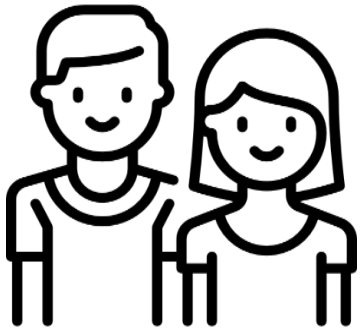
영화 찾아주세요!!

비공개 | 질문 8건 | 질문마감률 100% | 질문채택률 100% | 2017.04.14. 22:43 | 조회수 100

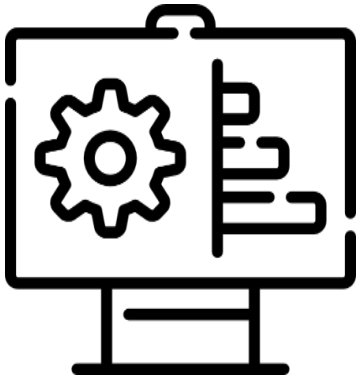
이거 완전 옛날에 봤던 영화거든요?
그게 해외영화고 내용이 엄마와 아들을 주제로 다룬 영화같은데
제가 기억하기로는 스릴러같습니다. 정원같은곳에서 파티를 하는데 엄마가 화장실로 들어가는데 그 화장실은 정말 큽니다 흰색이구요 그리고 아들이 들어오는데 아들이 이상해지고 문을 확 닫아버리고 엄마 손이 꼈던걸로 기억하고 그리고 엄마는 아들을 포함한 다른 아이들과 숨박꼭질을 합니다 그리고 지하방으로 내려가서 무엇을 찾는데 그게 죽은 아들 같습니다 진짜 찾아주세요ㅜ 다시 한번 보고싶은데 기억이 안나네요..

Research Purpose

- **질문자가 찾고 싶어 하는 기억 속의 영화 찾기**
  - 정제된 언어가 아닌 자연어 기반의 검색
  - 실시간으로 이루어지는 질의응답



남자가 여러 모습으로 여자 앞에 나타나는 영화 찾아주세요!

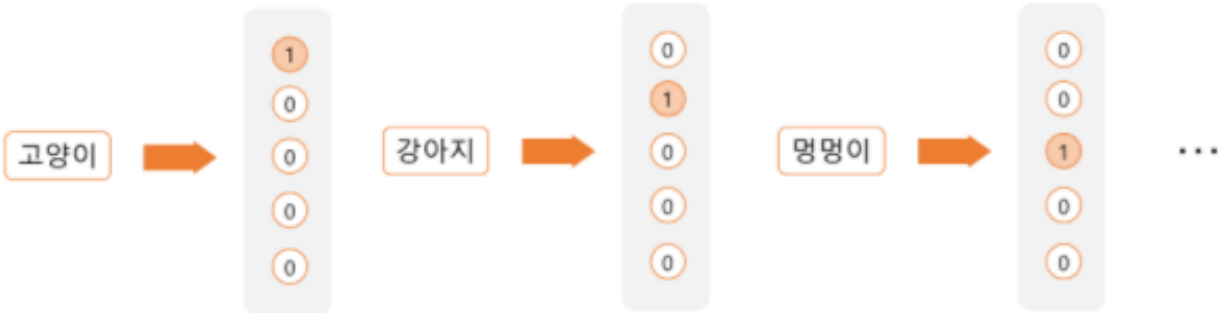


뷰티 인사이드

Word embedding

- **Feature representation**
  - 언어의 속성을 표현하는 두가지 방식 Sparse, Dense representation

- **One hot encoding (Sparse)**
  - 속성이 가질 수 있는 모든 경우의 수를 각각의 독립적인 차원으로 표현
  - 유사하거나 반대의 의미를 갖는 단어들의 관계를 반영 할 수 없음
    (강아지 : 멍멍이 = 강아지 : 자유주의)



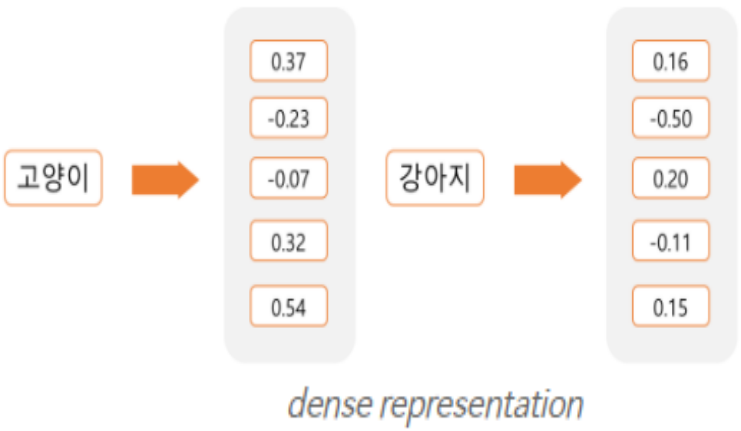
one-hot encoding

*https://dreamgonfly.github.io - 쉽게 씌여진 word2vec

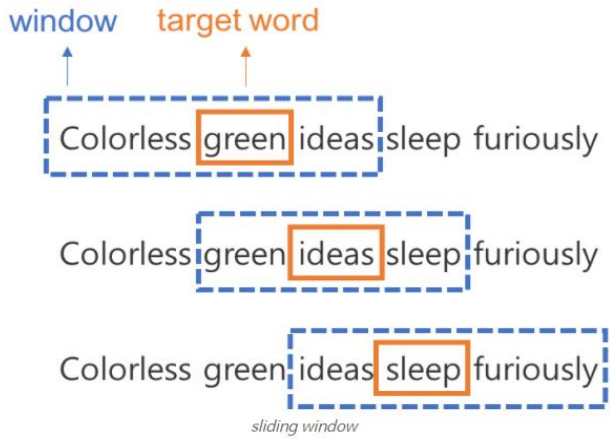
Word embedding

- **Word2vec (Dense)**
  - 한 단어의 정보가 여러 차원에 분산되어 있음 (Distributed representation)
  - 각각의 속성을 독립적인 차원이 아닌 사용자가 정한 차원에 대응시켜서 표현

- **Word2vec의 단어 학습**
  - 단어의 주변을 보면 그 단어를 안다 – J.R. Firth (1957)
  - CBOW 방식 : 맥락(context)을 통한 타겟 단어 예측 (Sliding window)
  - 맥락을 통해 파라미터를 학습하며, 그 파라미터가 벡터로써 표현



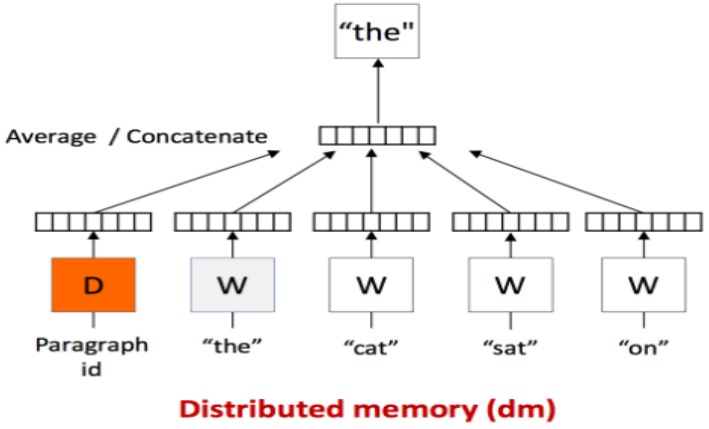
*https://dreamgonfly.github.io - 쉽게 씌여진 word2vec



*https://dreamgonfly.github.io - 쉽게 씌여진 word2vec

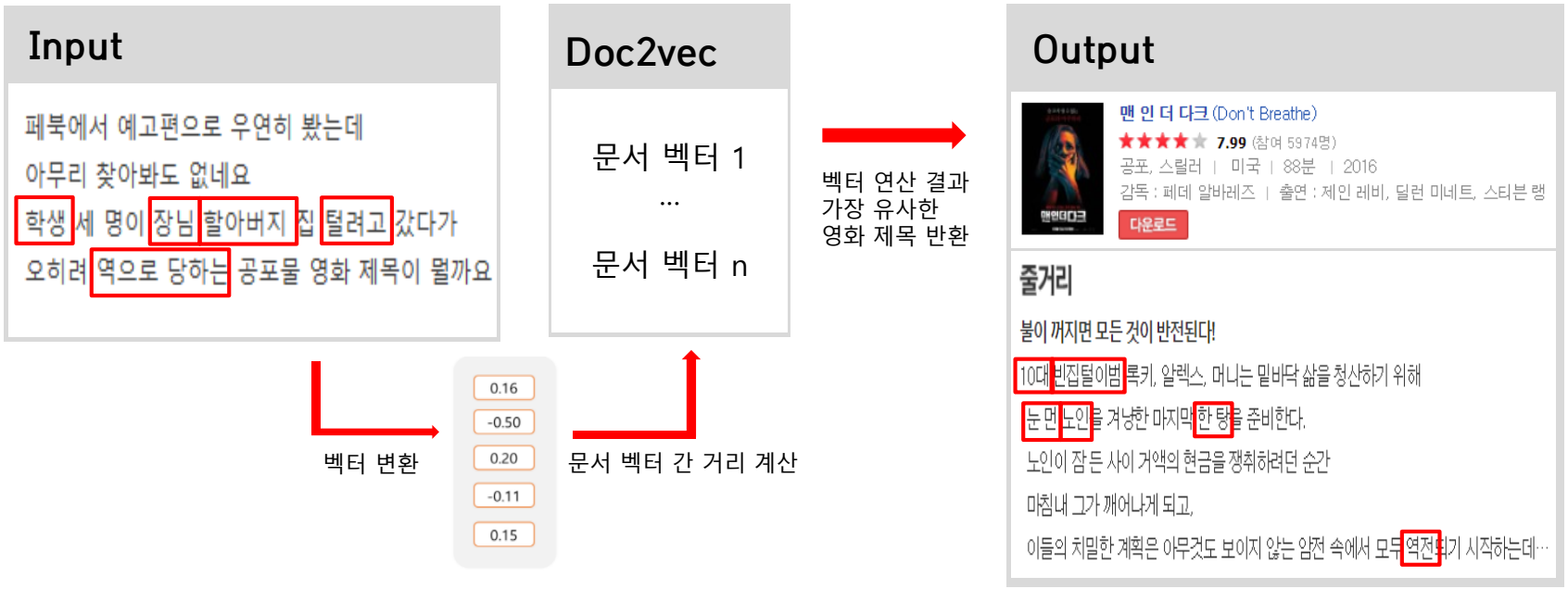
Word embedding

- **Doc2Vec**
  - DM 방식: 단어를 학습 시 각각의 학습 단계를 벡터에 기억시키고
    학습된 최종 벡터를 해당 문서의 벡터로 정의
  - 문서(문단)를 하나의 단어 처럼 학습
  - 각 문서마다 고유 아이디의 역할을 하는 새로운 특징 벡터(feature vector) 존재



* Le, Mikolov, Distributed Representations of Sentences and Documents, ICML, 2014.
(원문에서는 doc2vec 대신 **paragraph vector**라는 이름을 사용)

Procedure



- **질문자가 찾고 싶어 하는 기억 속의 영화 찾기**
  - 질문으로부터 키워드 추출 및 벡터화
  - Doc2Vec 알고리즘을 활용해 쿼리(Query) 문서 학습
  - 해당 질문과 가장 유사한(가까운) 문서 벡터를 가진 영화 제목 반환

Data collection

영화 찾아주세요 2017.11.02.

영화좀 찾아주세요 거의기억안나서 사실 찾기불가능할수도있지만 지식인들이라면..처음에..대충 생각나기론 애들이있는데 애들이 보드게임??하고있다가 그게 약간 정글...

#보드게임에 #빨려들어갔다 #후크선장

지식Q&A > 영화 | 답변수 1 · 추천수 1 | 답변 TOP 5 xedziihn

우주영화 찾아주세요오 2017.11.03.

좀 예전영화인데요 부자가 우주선을 타고 여행을가는데 착오가 생겨서 행성(?)에... 외계 생명체 역시 무차별적인 전쟁을 시작하는데... 찾으시는 영화는 [애프터 어스] 입니다.

지식Q&A > 공포, 스릴러 영화 | 답변수 1 · 추천수 1 | 답변 TOP 20 keroro73

탈출영화 찾아주세요! 2017.10.28.

... 영화 제목좀 찾아주세요 내공100 클로버필드 10번지 감독 댄 트라첸버그 출연 메리 엘리자베스 윈스티드,존 굿맨,존 갤러거 주니어 개봉 2016.04.07. 미국, 103분

지식Q&A > 공포, 스릴러 영화 | 답변수 1 · 추천수 0 | 답변 TOP 20 영화학생(lalf85)

영화 찾아주세요~ 2017.10.30.

... 올타리?에 걸려서 죽게되고 남편도 딸이 죽으니까 떠나는 영화엿는데 기억이 안나요 이거 무슨영화인가요? 바람과 함께 사라지다 감독 빅터 플레밍 출연 클락 게이블...

#영화

지식Q&A > 역사영화 | 답변수 1 · 추천수 0 | 답변 TOP 20 cheory73

맨 인 더 다크

Don't Breathe, 2016

관람객 8.43 기자·평론가 7.04

네티즌 7.99 내 평점 등록

개요 공포, 스릴러 | 미국 | 88분 | 2016 , 10.05 개봉

감독 페데 알바레즈

출연 제인 레비(록키), 딜런 미네트(알렉스), 스티븐 랭(눈 먼 노인) 더보기

등급 [국내] 청소년 관람불가

줄거리

불이 꺼지면 모든 것이 반전된다!

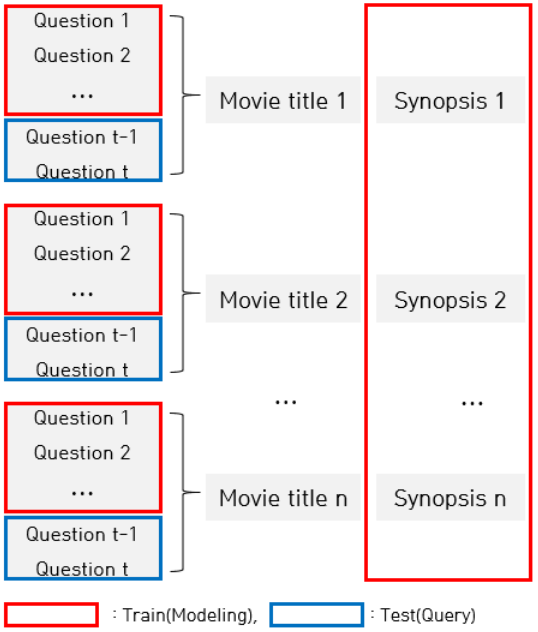
10대 빈집털이범 록키, 알렉스, 머니는 밑바닥 삶을 청산하기 위해
눈 먼 노인을 겨냥한 마지막 한 탕을 준비한다.
노인이 잠 든 사이 거액의 현금을 쟁취하려던 순간
마침내 그가 깨어나게 되고,
이들의 치밀한 계획은 아무것도 보이지 않는 암전 속에서 모두 역전되기 시작하는데…

- **Selenium을 활용한 크롤링**
  - 네이버 지식in (영화 질문과 답변, 해당 영화 정보) - 총 수집 116377 건

Preprocessing & Modeling

- **Preprocessing**
  - 50자 미만 질문, 2건 미만 질문 영화 제거 (정보 부족)
  - 이상치, 중복, 불용어 제거
  - KoNLPy 패키지를 활용한 POS-tagging (명사, 형용사, 동사)
  - 총 38614건 Q&A, 영화 3866건 사용

- **Modeling**
  - gensim 패키지를 활용한 doc2vec 구현
  - 시놉시스, 질문의 80%을 활용한 모델 학습
  - 20%의 질문으로 Test 진행
  - 한 영화에 대한 여러 질문들에 같은 Tag 부여
  - 가장 유사한 문서 상위 N개의 Tag 반환



Results

- **Results**

  - 모델 정확도 **(57.6% ~ 73.5%)**

    (찾고자 하는 영화가 결과 상위 1~10개 이내에 있는 경우)

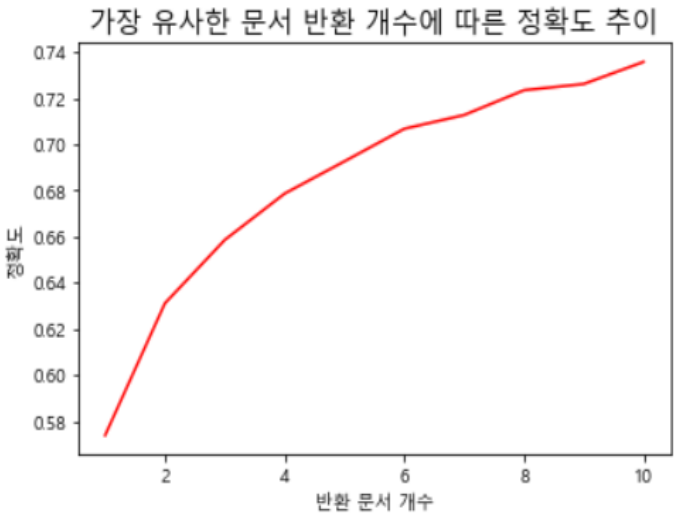
- **결과 예시**

Input

['보이다', '할아버지', '죽다', '보상금', '받다', '도둑', '보상금ㄱ', '훔치다ㄱ', '들어가다', '지하', '여자', '있다', '스릴러']

output

[('맨 인 더 다크', 0.1963629275560379),
('종횡사해', 0.1927681565284729),
('프리즌 히트', 0.19275511801242828),
('스턱', 0.18937425315380096),
('1408', 0.188787579536438),
('로스트 하이웨이', 0.17300422489643097),
('다크 엔젤', 0.17231591045856476),
('페이백', 0.16933536529541016),
('휴먼 센티피드', 0.16795529425144196),
('파커', 0.1645517796278)]



Discussion

- **의의**
  - 의미 기반 검색을 통한 새로운 검색 서비스 방안 제시
  - 음성인식 기술과 결합하여 AI 비서의 새로운 검색 서비스로의 확장 가능성
  - 쇼핑, 여행, 도서 검색 등 다양한 분야와의 접목 가능성
- **한계점**
  - 이미지를 통해 영화를 찾고자 하는 경우
  - 충분한 데이터가 없는 경우 (짧은 질문, 시놉시스 부재 등)
  - 영화 리뷰 데이터 등 추가적인 데이터를 통한 보완 필요

이미지로 영화를 찾고자 하는 경우

Q&A

감사합니다.

Appendix

모델 학습 파라미터

prameter

dm 'distributed memory' (PV-DM) is used. size is the dimensionality of the feature vectors.
window is the maximum distance between the predicted word and context words used for prediction within a document.
alpha is the initial learning rate (will linearly drop to min_alpha as training progresses).
seed = for the random number generator.
min_count = ignore all words with total frequency lower than this.
iter = number of iterations (epochs) over the corpus.
hs = 1, hierarchical softmax will be used for model training.
negative = negative sampling will be used, the int for negative specifies how many "noise words" should be drawn

```
# modeling
model = Doc2Vec(doc_list, alpha=0.1, size= 300, window = 6, min_count = 1, workers=1, seed = 42, iter=5, hs=1,negative=3)
```