1. (Morphological Similarity) : 형태적 유사성

처음에는 word2vec 모델을 훈련하여 가능한 많은 단어 벡터를 달성해야 한다. 일반적으로 **zh-wiki를 훈련 말뭉치로 선택하여 모델을 훈련**한다.

단어를 가능한 한 하이퍼 평면에서 선형적으로 분리하기에 **적절한 길이의 단어 벡터를 설정한**다. 그러고 나서 몇가지 생소한 말을 생략하고 중요한 정보를 기억하기 위해 **적절한 주파수 임계값을 설정**할 수 있다. 훈련이 끝나고 난 후, word2vec 모델이 완성되는 한편 ,모든 단어들의 단어 벡터들을 알게 되고 그렇다면 마치 사전처럼, 모델에서 단어의 벡터만 찾아보면 된다.

두 단어의 형태학적 유사성은 두 단어의 코사인 거리를 이용하여 계산할 수 있다.

예를 들어, w1과 w2라는 두 단어와 normalize 한 값을 w1',w2'라 정의하면,

$$\text{distance} = \frac{w_1 . w_2}{|w_1||w_2|} = w_1' . w_2'$$

이며, 위의 공식은 두 단어 w1과 w2 사이의 형태학적 유사성을 나타낸다.

우리는 word2vec을 이용하여 두 단어 사이의 유사성을 추정할 수 있지만, 위의 공식을 통해 얻은 것은 단지 형태학적 유사성일 뿐 즉, 그들의 의미적인 부분은 알 수 없다는 것이다.

이러한 상황에 대하여, 상대적으로 좋은 수단인 Senti-score of words를 제안한다.

2) (Semantics Similarity) : 의미론적 유사성

Wordnet은 단어를 기록하기 위해 tree를 사용한다. tree의 구조는 거리를 자연스럽게 정의한다. Wordnet을 사용하여 단어들의 의미론적 유사성을 계산할 때, 우리는 계산할 단어를 나타내는 두 개의 노드를 연결하는 최단 경로를 계산하여, 그것의 역수를 유사성으로 간주한다.

단어와 그 자체의 유사성은 1, 두 단어를 연결하는 경로가 없는 경우 두 단어 사이의 유사성을 0으로 정의한다. 위의 정의에서 의미론적 유사성은 0과 1 사이의 값이 될 것이고 그 값이 클수록 두 단어는 의미론적으로 유사하다는 것이다.

D. Senti-score Computation

1) Sentiment Lexicon(감정 어휘)

일반적으로 감정 분석에서, 우리는 우리의 모델에 **어떤 단어가 긍정적이고 부정적인지를 알려줄 수 있는 감성 어휘(Sentimet lexicon)**를 만들 것이다.

구체적인 감성 어휘(sentiment lexicon)를 단어의 감성을 평가하기 위한 작은 집합으로 정의하여 우리 모델이 금융시장을 나타내는 감성을 산출할 수 있도록 한다.

우리가 초기화하는 감성 어휘(sentiment lexicon)와 특정 단어의 유사성을 계산하여 그 단어의 정서를 반영하는 것이 가능한 것이다.

예를 들어, 우리는 금융 뉴스에 자주 등장하는 100개의 단어들을 선택하고 label로써 구체적인 정서의 단어를 갖고 있다. 여기서 계산을 좀 더 공평하게 하기 위해, 우리는 시장에 대해 긍정적인 태도를 가진 50개의 단어를 선택하고 부정적인 태도를 지닌 다른 50개의 단어를 선택한다.

2) Senti-score of Words

위의 계산을 통해서, 우리는 모든 단어로부터 200차원의 벡터를 갖게 될 것이다.
첫 번째 100차원은 Word2Vec과 sentiment lexicon(감성어휘) 100개의 단어와 유사하다.
그들은 감성어휘(sentiment lexicon)의 label단어와 형태학적 유사성을 보여준다.
반면에 또 다른 100차원은 Wordnet과 감성 어휘(sentiment lexicon)의 100개 단어와의 유사성으로, 라벨 단어와의 의미적 유사성을 나타낸다. Sentiment lexicon은 시장에 대한 태도를 보여주므로, 위의 유사성을 근거로 하여 우리는 특정 단어의 태도, 감정을 평가할 수 있다.

그 감정을 **양적**으로 나타내기 위해, 우리는 **Senti-score**이라는 값을 정의한다.
이 연구에서 우리는 모든 단어의 Senti-score 값을 계산하기 위해 similarity vector를 사용한다. 구체적인 단계는 아래와 같다.
(1) Word2vec 유사성과 WordNet 유사성을 각각 사용하고, 그런 다음 상호협력 필터링을 사용하여 이를 긍정적 또는 부정적 단어로 분류한다.
(2) Senti-score를 계산하려면 Word2vec 유사성을 사용한다.
구체적인 과정과 이면의 이유는 다음과 같다.

우리가 감성어휘(sentiment lexicon)와 단어들과의 유사성을 고려할 때, 형태학적, 의미적 유사성을 고려해야 하고, 그것은 협업 필터링을 통해 찾을 수 있다.
우선, 우리는 상위 n개의 유사한 단어들을 찾기 위해 Word2vec 유사성을 지닌 처음 100차원을 사용한다. 우리가 처음 100차원에서 골라낸 가장 큰 가치를 지닌 이 n개의 단어들은 우리가 계산해야 할 목표 단어와 가장 형태적으로 유사한 n개의 단어일 것이다.
그러고 난 후, 우리는 마지막 100차원에서 이 n개의 단어들의 WordNet 유사성을 비교한다. 이 n개의 단어 중에서 워드넷 유사점이 가장 큰 상위 m을 골라낸다면 이 단어들은 형태학적으로나 의미론적으로 모두 목표 단어와 유사한 상위 m개의 단어가 될 것이다.