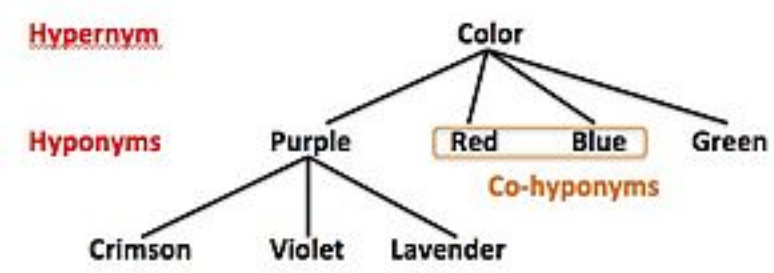
CS372 Project: First Proposal

한수민, 박주회

1. Background

Hypernym이란 주어진 단어의 의미적 부분이 해당 단어보다 상위적 계층에 속해 있는 것을 의미한다. 이와는 반대로 hyponym이란 의미적 부분이 하위에 있을 때를 의미한다.



위와 같은 예시에서, Purple, Red, Blue, Green이라는 단어는 모두 색을 의미하는 단어이므로, Color의 Hyponym이라 할 수 있다. 반대로 Color는 해당 단어들의 hypernym이 된다.
NLTK에서는 많은 수의 영어 단어에 대해서 이러한 hypernym, hyponym 분석을 사용할 수 있는 라이브러리를 제공하고 있다. 본 프로젝트에서 우리는 이러한 hypernym-tree를 이용해 뉴스기사에 대해 다양한 분석을 시도해보고자 한다.

```
>>> dog = wn.synset('dog.n.01')
>>> dog.hypernyms()
[Synset('canine.n.02'), Synset('domestic_animal.n.01')]
>>> dog.hyponyms()
[Synset('basenji.n.01'), Synset('corgi.n.01'), Synset('cur.n.01'), Synset('dalmatian.n.02'), ...]
>>> dog.member_holonyms()
[Synset('canis.n.01'), Synset('pack.n.06')]
>>> dog.root_hypernyms()
[Synset('entity.n.01')]
>>> wn.synset('dog.n.01').lowest_common_hypernyms(wn.synset('cat.n.01'))
[Synset('carnivore.n.01')]
```

(NLTK 에서 hypernym과 hyponym의 사용 예시)

2. Research Proposal

2.1 Hypernym-tree를 이용한 뉴스 기사의 주제 및 뉴스관 연관성 분석

어떤 뉴스 기사의 단어들을 각각에 대해 hyponym을 검색하는 식으로 하면 각 단어마다 상위 개념이 있을 것이므로 어떠한 일련의 node와 edge로 이루어진 linked list가 만들어 질 것이다. 그러한 각 단어마다 생기는 Linked-list와 다른 단어들로부터 만들어진 list들에 공통적으로 포함되는 단어를 찾아서 연결시키는 방식을 반복한다면 일종의 트리 구조를 형성하게 될 것이다. 예를 들어 위의 예시에서 Crimson-Purple-Color와 Lavender-Purple-Color 두 단어군 모두 Color를 향하게 되므로 이 두 단어로부터 나오는 Tree 는 위에서 보여진 트리와 같은 형태를 띠게 된다.

이러한 hyponym 을 이용한 tree 구조가 나오게 되면 frequency를 계산 하지 않더라도 얼마나 더 많은 Child node를 가지고 있는지를 통해 소위 "인기있는" node를 파악할 수 있다. 즉, 특정 article에서 나오는 단어들을 분석하여 hyponym으로 발생하는 linked-list를 기반으로 한 tree를 만들게 되면 해당 article의 주제와 관련한 키워드들이 인기있는 node 로 나타날 것으로 예상된다. 또한 단어에 대한 frequency를 이용하여 node에 대한 weight 까지 추가한다면 해당 기사에서 사용된 단어들의 의미망 분석에 대해 보다 더 정확한 정보를 얻을 것으로 예상된다. 따라서, 이러한 Hypernym-tree를 이용하여 뉴스의 주제를 파악하고 키워드를 뽑아내는 것이 목표이며, 더 나아가서는 다양한 뉴스들에서 나오는 hyponym-tree를 통해 알 수 있는 인기있는 키워드를 분석하여 뉴스를 비슷한 뉴스끼리 묶어 주는 알고리즘을 개발해 보는 것을 목표로 한다.

2.2 Hypernym-tree 구조 분석을 이용한 기자들의 글쓰기 스타일 분석

뉴스를 작성하는 기자마다 문장을 서술하는 방식, 같은 단어를 반복하는 빈도, 비유 등을 사용한 같은 개념을 설명하는 방법 등 개인의 글쓰기 스타일이 다르게 나타날 수 있다. 그러한 부분들을 hyponym tree의 구조를 분석하여 어느 정도 유사성을 포착할 수 있다고 생각된다. 따라서, 해당 프로젝트는 hypernym-tree의 구조를 분석하여 기사들간의 글쓰기 스타일을 특징지을 수 있는 방법을 찾아내는 것이 목표이다. 이와 같은 경우에는 짧고 여러 개의 기사를 분석하는 것 보다 꾸준히 특정 주제에 관한 칼럼을 작성하는 것 보다 긴 글을 분석하여 좀 더 섬세하게 표현되는 Tree들의 모양 패턴을 분석하는 것이 더 의미를 가질 것이라 예상된다.