뉴스에서는 키워드를 뽑아 내는 것 보다 주제(Theme)를 뽑아 내는 것이 더 현명하지 않을까?

키워드 추출에 대한 생각

키워드는 해당 문서내에 등장하는 중요한 핵심 단어라고 할 수 있다. 그것과 관계된 내용일 가능성이 높다. 그러나 키워드가 언제나 기사의 핵심 주제라고는 할 수 없다. 예를 들어

“트럼프, 김정은 ‘밀당’.. 역사적 정상회담 이뤄냈다.”

라는 기사의 제목의 키워드는 아래와 같이 선정되어 있었다.

```
  "keyword": [
                "트럼프",
                "정상회담",
                "김정은",
                "북미",
                "화염과분노"
            ],
```

‘북미’는 이 기사의 키워드라고 할 수 없다. ‘화염과 분노’ 역시 좀 어색하다.

‘북미’라는 단어가 등장하거나 ‘북미’와 관련된 뉴스 기사를 찾고 싶을 때는 ‘북미’라고 검색창에 두들겼을 때 위 기사가 등장하는 것이 이상하지는 않다. 그래서 언어적으론 키워드가 아니지만 기술적으로는 키워드가 될 수 있다.

‘화염과 분노’도 마찬가지다. 예를들어 아래와 같은 상황이다.

전에 트럼프 미국 대통령이 뭐라 했던 말이 있는데 잘은 기억이 안난다. 하지만 그가 ‘화염과 분노’라는 말을 언급 했던 것은 분명히 기억이 난다.

이 때는 ‘화염과 분노’라고 검색 할 수도 있을 것이다. 이런 상황에서 ‘화염과 분노’는 키워드가 될 수 있지만 기사 내용을 이미 알고 있는 시점에서 ‘화염과 분노’는 그다지 중요한 키워드가 아닐 것이란 말이다.

모르는 내용의 어떤 문서를 찾을 때 키워드를 활용하는 것은 매우 효과적이지만 이미 알고 있는 내용의 문서에 대해 키워드를 선정하라고 했을 때는 일반적인 키워드 추출이 썩 좋은 결과물을 내놓진 않는다고 생각한다. 그래서 나는 ‘합성 명사’를 추출 해 내기 위해 ‘키워드 추출’이 아니라 ‘테마 추출’이 필요하다고 믿고 있다.

예를들어

“트럼프, 김정은 ‘밀당’.. 역사적 정상회담 이뤄냈다.”

이 기사의 경우 아래와 같은 테마가 추출되면 참 좋겠다.

트럼프_김정은_정상회담

3개의 명사가 합성된 새로운 명사다. 이것은 word가 아니라 multi-word이다. 한가지 골 때리는 점은 TextRank와 PMI 구현 방법에 따라 아래와 같은 상황이 발생 할 수도 있다는 것이다.

김정은 트럼프 정상회담
트럼프 김정은 정상회담
정상회담 트럼프 김정은
정상회담 김정은 트럼프

난감하다. 이 경우 JACARD 유사도와 같은 방법으로 서로 비교 했을 때 100% 일치하므로 모두 '똑같은 합성 명사' 취급을 해 주면 좋을 것 같다. 아직 이 해결 방법에 대해선 고민중.

A = { "김정은", "트럼프", "정상회담" }
B = { "트럼프", "김정은", "정상회담" }

A와 B의 유사도는 100%니까 동일한 합성 명사로 간주한다. 더 골 때리는 상황은?

A = { "안희정", "성추행" }
B = { "안희정", "성폭력" }

이거 골치 아프다. 유사도가 50%인데 똑같다고 해야할까 다르다고 해야 할까? 근데 성추행과 성폭력은 엄밀하게 법적으로는 다르지만 '이슈'라는 관점에서는 둘 다 비슷하지 않나? 그러니까 유사도를 100%로 봐야하나 50%로 봐야하나? 100%로 봐도 50%로 봐도 애매한 상황이다. 해결해야 할 문제가 많다.

인물감지에 대한 생각

TextRank를 하다보니 발견한 문제점이다. 글이라는 것이 결코 똑같은 단어만 써 가며 쓰는 것이 아니다. 영어는 특히 더 그렇다. 처음에 James라고 지칭 했다가 나중에 가서는 He, Him, His 등의 대명사로 표현한다. 아래의 경우를 보자.

바른미래당 **하태경 최고위원**이 8일 성폭행 혐의를 받고 있는 **안희정 전 충남지사**를 “긴급체포해야 한다”고 밝혔다.

하 최고위원은 이날 국회에서 열린 원내정책회의에서 “**안 전 지사**가 상습 강간범이라는 게 확인이 됐다”며 이같이 밝혔다.

하 최고위원은 “추가 피해자가 없다고 했는데 어제 저녁 뉴스에 추가 피해자가 나왔다”며 “더 악질적 범죄라는 게 확인됐다. 또 도주의 우려도 있다”고 했다.

하 최고위원은 하태경을 의미하는 말이다. 근데 TextRank를 해 보면 ‘최고위원’이 키워드로 추출되는데 당연히 ‘하태경’보다는 최고위원이 더 많은 간선을 가지니까 그렇다. 이게 TF-IDF에서는 ‘하태경’이 나름 추출 되는데 TextRank에서는 그렇지가 못하다. 이것도 유사도를 사용해서 비슷하게 취급하면 어떨까?

A = { “하”, “최고위원” }
B = { “하태경”, “최고위원” }

형태소 분리를 할 때 저렇게 분리 해 준다면 얼마나 좋을까. 아쉽다. 자연어 처리를 통해 문장에서도 ‘주어’를 찾으면 좋을 것 같다. 각 문장에서 주어들 끼리 유사도를 비교해서 동일한 것들은 동일하게 취급하는 것이다. 그렇게 한다면 ‘하 최고위원’이 정확히 누구를 말하는 것인지도 알 수 있을 것이다.

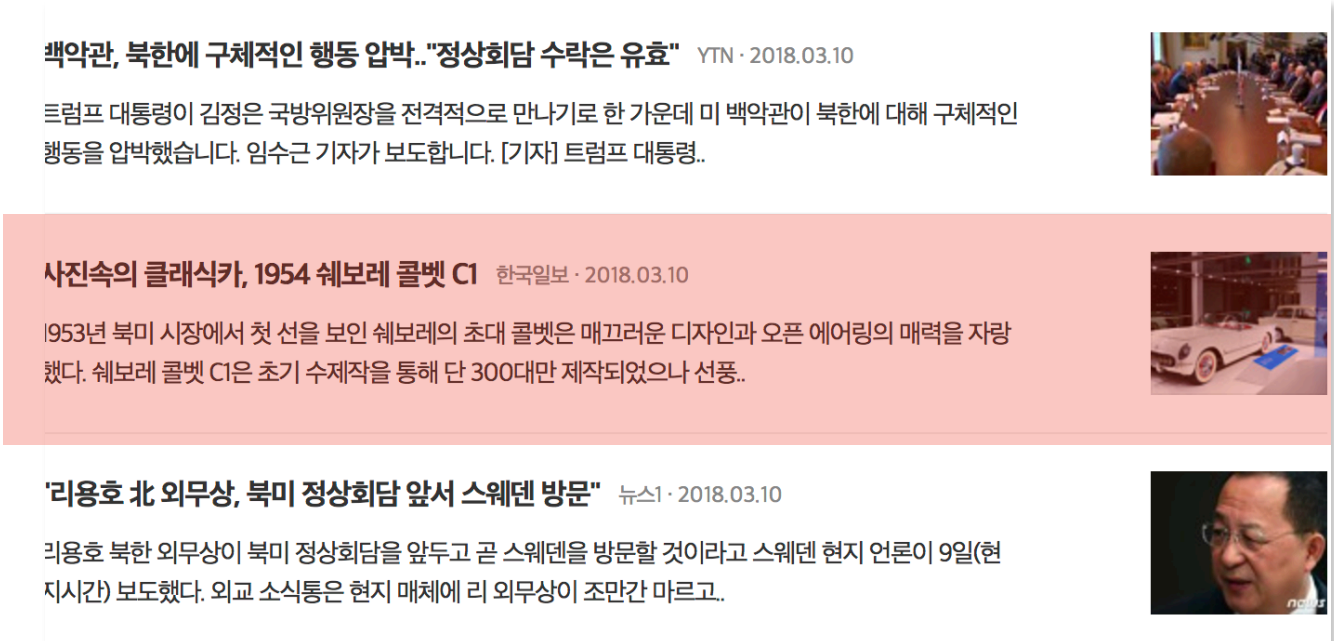
단일 키워드의 문제점

요즘 트럼프와 김정은의 정상 회담이 이슈다. 메인 키워드에서 “트럼프” 키워드를 클릭 했을 때 “김정은 트럼트 정상회담”과 직접 관련된 내용이 나오는게 아니라 간접적으로 “트럼프”가 언급 되었더라도 해당 기사를 표시하고 있다.

좀 많이 아래로 내려가보면 요즘 이슈가 되는 정상 회담 기사는 없고 “관세 폭탄” 관련한 기사가 뜨기도 한다. 말그대로 오로지 “트럼프”를 키워드로 추적 했을 때의 결과이다.

처음 부터 키워드가 #트럼프_김정은_정상회담, #트럼프_관세폭탄 이렇게 나뉠 수 있다면 좋을 것 같다. 좀 더 확실하게 두 주제를 ‘지금 뜨는’지 안 뜨는지 구분 해 줄 수 있으니까 ㅇㅇ. 관세폭탄 보다는 지금 정상 회담이 더 뜨는 키워드 일 수 있으니까... 또 한가지 예를 들어보면

아래 사진은 #북미를 클릭 했을 때 나오는 결과다. 요즘은 이 말이 북한과 미국을 의미하는 건데 미국의 북부 지방을 의미하기도 해서 클래식카에 관한 기사도 리스트에 담겨져 버렸다. 해당 기사는 이번 정상회담과는 전혀 무관계한 내용이었다. ‘북미 판매량’이라는 말이 많이 언급 되던데 그것 때문인 것 같다 ㅋㅋ.



#북미_정상회담 이런식으로 주제가 추출 되었더라면 아마 클래식카 기사는 리스트에 없어서 더 좋았을 것이다. 저 기사에 '정상회담'이라는 단어는 없으니까. 일단 지금은 많의 양의 '최신 기사'로 덮어 버려서 어느정도의 서비스 질은 유지하고 있는 것 같다.

```
"김정숙 여사, 평창올림픽 화장실 청소봉사 어르신들과 오찬"
```

```
"keyword": [
                "화장실",
                "평창올림픽",
                "김정숙여사",
                "평창동계올림픽",
                "감사"
            ],
```

다음 뉴스에서 제공하는 키워드가 썩 좋은 편은 아닌 것 같다. 이 기사가 '화장실'에 대한 이야기를 하고 있는 것은 아니지 않은가... 하지만 검색으로서의 키워드는 좋다고 할 수 있다. 그치만 '감사'의 경우에는 검색 기술적으로도 언어적으로도 썩 좋은 것 같진 않다. 내가 생각하는 테마 추출 방법을 쓴다면 '평창올림픽 화장실 청소 봉사' 또는 '김정숙 여사'가 나오면 가장 이상적일 것이다. 아래는 내가 구현한 TextRank + PMI를 통해 키워드를 추출한 결과다. 지금 다음에서 서비스 중인 키워드 추출과도 비교 해 보았다.

Daum	arf.rumo
보수 변호사단체 "남북정상회담서 북한인권 문제 다뤄야"	
정상회담	한반도 인권
북한인권	북한 주민 인권
보수	남북 정상회담 의제
북핵	주민 인권

아무튼 다르게 나옴. 기사 내용 자체는 남북 정상회담 의제에 북한 주민 인권 문제도 같이 다루자는 내용.