1. 텍스트마이닝 기반의 중심기술 분석을 통한 경전철용 무선전력전송 기술 개발 전략

해당 논문은 "텍스트 마이닝 기반의 중심기술 분석을 통한 경전철용 무선전력전송 기술 개발 전략"에 관한 연구를 다루고 있습니다. 무선전력전송(WPT) 기술의 기술 개발 동향 및 중심 기술을 분석하여 경전철용 무선전력전송 시스템의 기술 개발 전략을 제시하는 내용이 포함되어 있습니다.

무선전력전송 기술
접촉식 전력공급시스템보다 건설비 및 유지보수 비용이 저렴
신뢰성 및 안전성이 높으며 환경 친화성을 향상
미래 교통시스템의 전력공급 솔루션으로 주목
특히, 이 기술은 철도, 우주항공, 자동차, 선박, 모바일 및 전자기기, 로봇 등 다양한 분야에 적용될 수 있는 파급효과

TF-IDF 방법
1. 특허 문서로부터 키워드를 추출하고,
2. 사회 네트워크 분석(SNA)을 통해 각 기술 영역별 중심 키워드를 도출하였습니다.
3. 분석 결과, 무선급전/집전 기술, 고주파 전력변환장치 기술, 궤도 및 인프라 기술이 WPT 기술 개발의 핵심 분야.

온라인 집전과 센서 기술, 자기공진을 이용한 공진코일 및 공진커패시터 기술, ECU(전자제어장치) 기반의 무선전력전송 제어 기술이 중요한 개발 전략으로 제시

중국이 경전철용 WPT 기술 개발에서 가장 활발한 특허 활동
기술의 상용화를 위해 지속적인 투자와 기술 역량 제고가 필요함을 강조

이러한 연구 결과를 바탕으로 향후 국내 경전철용 무선전력전송 기술의 연구 개발 방향성을 제시하고, 국내 산업의 경쟁력 강화에 기여할 것으로 기대

무선전력전송 기술 유망하다네? -> 기술의 트렌드를 파악(요기서 텍스트 마이닝) -> 중심기술이 무엇인지 알아내자 !! -> 미래 방향성 및 실용화 방안 마련 기틀이 될꺼야 !!

텍스트 마이닝

> 키워드 기반의 특허 분석은 각 특허가 담고 있는 기술정보를 키워드 통해 파악하는 것이다.
> 여기서 특허문서는 구조화 되지 않는 자연어로 구성되어 있으며 정량적 분석이 어렵기 때문에,
> 텍스트 마이닝과 같은 전처리 프로세스를 통해 분석이 가능한 구조화 데이터로 변환해야 한다.

특허 => 기술의 발전 트렌드를 분석하기에 아주 좋은 수단.

근데 특허를 분석하기에는 자연어로 구성되어 있어 분석어려우니깐 텍스트 마이닝으로 전처리하여 분석이 가능한 구조화 데이터로 변환하겠다.

본 연구는 특허의 초록을 이용하여 분석이 가능한 데이터 구조 중의 하나인 특허-키워드행 렬(Patent-keyword Matrix)로 구성

사회네트워크 분석

개체들 간의 상호작용을 보기위해 개체 간의 관계를 네트워크로 형성 -> 구조를 시각적으로 표현

본 연구에서는 사화네트워크분석 기법과 키워드 분석을 복합적으로 연계하여 키워드 네트워크분석 기법으로 특허 문서에서 출현하는 단어와 단어 사이의 관계를 도출한다.

또한,

사회네트워크분석에서 네트워크 연결구조의 특성을 파악하기 위해
네트워크 밀도(Density), : 노드 간의 전반적인 연결정도
중심성(Centrality) & 집중도 (Centralization) : 한 노드가 네트워크의 중심에 위치하는 정도
등의 측정 지표를 사용

> 본 연구에서는 특허정보 중 초록문서를 활용하여 키워드를 추출하였다.

```
키워드의 추출 단계에서 단어의 출현빈도 = TF, Term Frequence
```

> 그러나, TF만을 사용하면 모든 문서에서 공통적으로 많이 출현되는 단어가 가장 중요하게 나오게 되는 문제점이 있다.
>
> 이에 수사, 관사 등의 단어를 제거하고
>
> 출현빈도가 높은 단어의 중요도를 계산하는 기법인 TF-IDF(Term Frequece-Inverse Document Frequence)를 사용

```
TF-IDF는 단순 출현빈도를 나타내는 TF에 문서 출현 빈도인 DF를 역으로 곱함으로 문서에서 출현하는 공통적인 단어를 효율적으로 제거할 수 있다.
```

> 이를 이용해서 단어를 추출해내고,
>
> 추출된 키워드 간의 상관관계를 분석, 직접적으로 연관성이 높은 단어들을 도출하고 이를 중심 키워드로 정의
>
> 근데 또, 중심 키워드만으로는 의미가 너무 일반적이거나 개념을 표현하기 모호하자나요,,
>
> 단어들의 상관계수의 평균을 계산하여 이를 넘는 단어만을 채택한 후, 기술의 개념을 설명할 수 있는 “multi-word phrases”를 만들었다.

```
이 그래서 이 단어들을 자료와 전문가의 의견을 토대로 단어를 합쳐서 중심기술로 도출해냄.
```

LDA 기반의 토픽모델링을 이용한 철도차량용 무선급전시스템 연구 동향 분석

철도차량용 무선급전시스템과 관련 된 연구 논문의 초록 데이터에
LDA(Latent Dirichlet allocation) 알고리즘 기반의 토픽모델링(Topic Modeling) 기법을 적용
세부 연구 키워드를 정의
추출된 세부 연구 키워드에 대해
Hot 토픽과 Cold 토픽을 규명(Seo and Lee, 2015)함으로 써
논문을 활용한 연구 측면의 정량적인 동향 분석을 시행하 고자 한다.

> 본 연구에서는 철도차량용 무선급전시스템과 관련된 논문 데이터에 대한 사회연결망분석(Social Network Analysis), 토픽 모델링(Topic Modeling), 시계열 분석(Time Series Analysis)을 통해 연구의 발전 과정과 동향을 정량적으로 분석한다.

사회네트워크분석은

```
위와 동일함.
```

토픽모델링 -> 이게 이 논문에서의 텍스트 마이닝 기법.

> 토픽모델링은 전체 문서 집합에서 의미 구조를 발견하기 위해 사용되는 텍스트마이닝기법

```
LSA = 토픽 분석 초기 방법론
pLSA += 확률

하지만 앞의 두 방법은 모두 문헌-용어 행렬만을 입력값
        => 문헌 내에 주제가 어떻게 분포하고 있을지는 고려하지 않음.

이에 본 연구에서는 pLSA의 베이지안(Bayesian) 버전이라고 할 수 있는 LDA 알고리즘(토픽모델링 중 가장 진보)을 기반으로 연구
```

LDA

> 문헌 내의 잠재된 주제를 추정하는 것으로 어떤 단어들을 어떠한 주제에서 선택하여 배치할 것인지 각각의 파라미터(Parameter)로 모델링하는 기법
>
> ```
> 관찰된 변수 ex... 문헌, 단어 --->  잠재 변수 ex...문헌의 구조
> 그니깐 쉽게 말하면,  치킨 피자 삼겹살 -> 음식을 도출 / 헬스 러닝 필라테스 -> 운동을 도출하겠다. 이 말인듯.
> ```

시계열 분석

> 미래 유망 연구분야를 선정 하기 위해 추출된 토픽별로 시간의 흐름에 따른 트렌드를 파악하는 것이 필요
>
> 시계열 분석은 시계열 데이터 에서 통계적인 의미를 추출하기 위해 과거의 자료에서 관측된 값을 기준으로 미래의 값을 예측하는 방법론

> ```
> 시점(독립변수)과 빈도(종속변수)로 이루어진 데이터
> ```

> 본 연구에서는 토픽별로 해당되는 문서들의 사후확률(Poste- rior) 값을 연도의 분기별로 평균낸 분기별 비중 평균값을 종속 변수로 사용하고 시점을 독립변수로 사용하여 추세를 파악하고자 한다.

```
빈도에 사후확률 / 분기별 비중 평균값을 집어넣었대용
```

> 토픽별로 예측 모형의 설명력이 높은 토픽을 대상 으로 Hot 토픽과 Cold 토픽을 분석한다. Hot 토픽은 최근 몇 년 간 관심이 높아지고 있는 토픽을 뜻하고, Cold 토픽은 점차 관 심이 줄어드는 토픽을 의미

여기까지가 사용할 기술에 대한 설명

여기부터는 진짜 과정

1. 데이터수집

생략

2. 데이터 전처리

> 텍스트 분석 위해 R programming의 “tm” 패키지를 전처리과정에서 이용
> 먼저 문장부호를제거(remove Punctuation)하고
> 숫자를 제거(removeNumber)한 후,
> 공백을 기 준으로 구분(Stripwhitespace)하였다.
> 그다음, 특수문자, 조사, 관사 등 불용어(Stopwords)를 제거하는 일련의 과정을 거쳐
> 만들어진 키워드들을 바탕으로
> 문서와 키워드 간의 매트릭스 (Document-Term matrix)를 구축
> 마지막으로 526개의 문서에서 4,308개의단어로이루어진 Document-Term matrix를
> 토대로 토픽모델링 분석을 진행하였다.

3. 동시인용 연결망분석

이 부분 첫번째 논문에선 그냥 상관관계 구해서 인접행렬 했다고만 하고 방법론이 제시가 안됐는데 요기에는 나온거네.
둘 다 연결망 분석을 하는데 여기선 페이지랭크 중심성쓴다면서 수식 넣어놓은 거네.

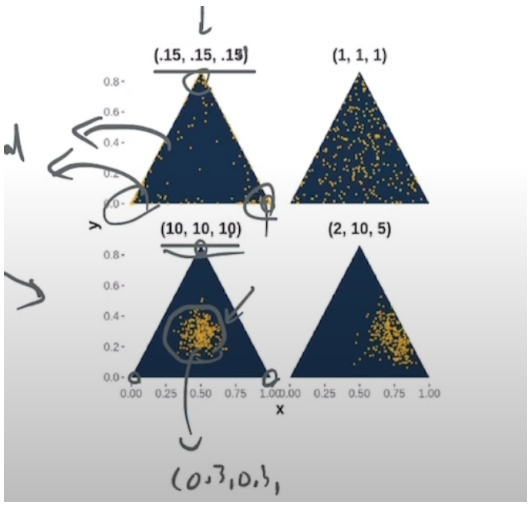
4. LDA 기반의 토픽모델링

본 연구는 논문 초록 정보에 LDA 기반의 토픽모델링을 적용한다.
R programming의 “topicmodels” 패키지를 활용하며,
LDA 모델의 파라미터 추정을 위해 Gibbs Sampling 방식을 사용한다(Hornik and Grün, 2011; Heo, 2012).
토픽을 생성하기에 앞서, 몇 개의 토픽을 생성할지에 대한 기준이 필요하다.
토픽 의 개수는 사전에 주어지지 않기 때문에 이를 고려해야 한다.
이는 언어모델의 복잡도(Perplexity) 알고리즘에 의해 최적의 K개를 도출하였다.

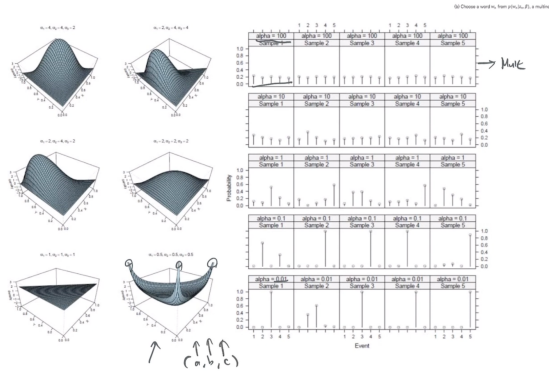
```
-> 이게 무슨말이냐 헬스 마라톤 러닝에서 => 1개를 뽑으면 '운동'이겠지만 '건강'이라는 주제를 뽑을 수도 있자나여
 -> 최적의 K개를 언어모델의 복잡도 알고리즘을 통해서 도출하겠다.



내가 무슨 내용을 쓸 것인지(Topic), 쓸 내용을 어떻게 표현할 것인지(Word)
```

LDA 알고리즘은 문서-단어행렬 생성을 통해 추출된 문서 내 단어들을
토대로 문서간의 구조를추론하고, 문서내 단어들에 대한 디리클레(Dirichlet) 분포를 통해 K개의 수만큼 토픽을 생성하는 것이다

디리클레 분포
다항분포의 분포 => k=3인 다항분포라면 (0.6, 0.3, 0.1) (0.7,0.1,0.2) 등등이 있을텐데, 이를 모조리 담고 있는 분포가 디리클레 분포



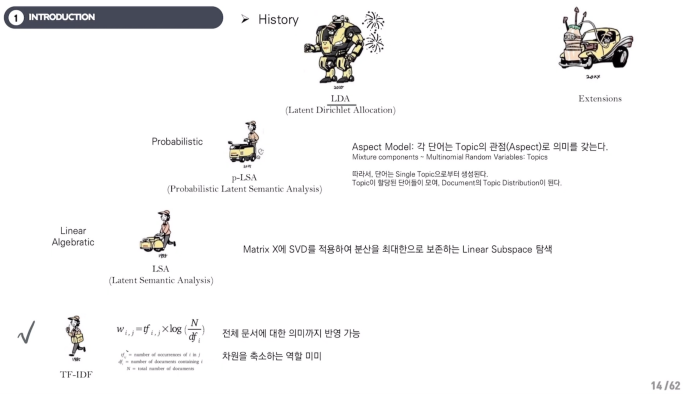
근데 이게, 알파 값에 따라 어떤 다항분포가 잘나올지가 정해짐.



3D로 보면 이렇구영

그래서 디리클레 분포에서 다항분포 하나를 뽑고 (세타가 뽑힌거임) , 여기서 어떤 토픽을 뽑을지 확률로 계산함. (체육인 0.1, 아이돌 0.8, 일반인 0.1)=> 여기서 확률 높은애가 Zn(아이돌)

이 Zn이 또 하나의 확률 분포(레드벨벳, 아이브, 아이유, 마동석 등등으로 이루어진)인데,
여기서 Wn을 샘플링



TF-IDF => 전체 문서에 대한 의미 반영 가능 차원을 축소하는 역할 미미, 잠재된 정보 표현하기에는 역부족.

그래서 나온게 LSA

확률을 더해 보완한게 p-LSA

이게 최종본임. LDA

In sum,

1번의 TF-IDF는 개수 카운팅 -> 전체 문서에 대한 의미를 파악

```
특허정보 중 초록문서를 활용하여 키워드를 추출

그러나, 단어의 출현빈도(TF)만을 사용하면 모든 문서에서 공통적으로 많이 출현되는 단어가 가장 중요하게 나오게 되는 문제점

이에 수사, 관사 등의 단어를 제거하고

출현빈도가 높은 단어의 중요도를 계산하는 기법인 TF-IDF(Term Frequece-Inverse Document Frequence)를 사용

추출된 키워드 간의 상관관계를 분석, 직접적으로 연관성이 높은 단어들을 도출하고 이를 중심 키워드로 정의

근데 또, 중심 키워드만으로는 의미가 너무 일반적이거나 개념을 표현하기 모호

단어들의 상관계수의 평균을 계산하여 이를 넘는 단어만을 채택한 후, 기술의 개념을 설명할 수 있는 “multi-word phrases”를 만들었다.(합쳤다.)
```

2번의 LDA는 잠재된 정보까지 끌어내는 가장 진보된 채첨단 신기술

논문의 초록 데이터에
LDA(Latent Dirichlet allocation) 알고리즘 기반의 토픽모델링(Topic Modeling) 기법을 적용

세부 연구 키워드를 정의
추출된 세부 연구 키워드에 대해
시계열 분석하여
Hot 토픽과 Cold 토픽을 규명(Seo and Lee, 2015)함으로 써
논문을 활용한 연구 측면의 정량적인 동향 분석을 시행하 고자 한다.

이게 무슨말이냐 헬스 마라톤 러닝에서 => '운동'을 뽑아내겠다.