영어 SF 소설 코퍼스 구축과 디지털인문학적 분석 사례

한수미
영어영문학과 & 디지털인문예술전공
(한림대학교)

한국근대영미소설학회 연찬회
2023년 12월 16일(토)

Hallym University

발표 자료: https://github.com/SumiHan/DH_SF_corpus_analysis

목차

- 연구 배경 & 목표
- 연구 방법
- 연구 결과 & 토의
- 결론 & 후속 연구 제언

Hallym University

발표 자료: https://github.com/SumiHan/DH_SF_corpus_analysis

연구 배경 & 목표

Hallym University

발표 자료: https://github.com/SumiHan/DH_SF_corpus_analysis

연구 배경 & 목표

“인간과 협업하는 딥러닝 기반 AI 소설 생성 융합 연구” 의 기초 연구

(한국연구재단 일반공동연구지원사업

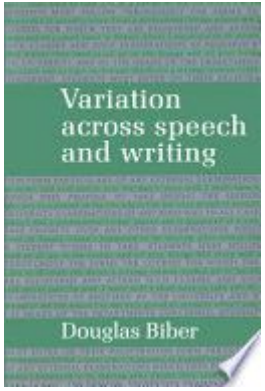
- 문학 텍스트 코퍼스 구축 및 평가
- 문학 텍스트의 언어적 패턴을 이해
- 향후 딥러닝과 결합한 소설 생성에 적용하고자 함
  (인간 소설 vs. AI 소설)

연구 배경 & 목표

코퍼스 언어학에서 디지털 인문학까지

1. 코퍼스언어학: 코퍼스 구축, 평가 및 언어적 패턴 연구

- 코퍼스 구축과 평가(balance, representativeness): COCA 2020
- (구어, 문어) 코퍼스의 언어적 패턴, 장르, 스타일, 표현 방식에 대한 탐구
- Biber (1988), Biber & Conrad (2009), Biber & Egbert (2016)



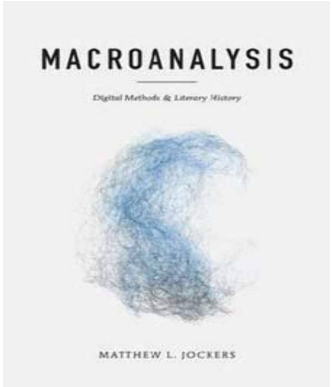
연구 배경 & 목표

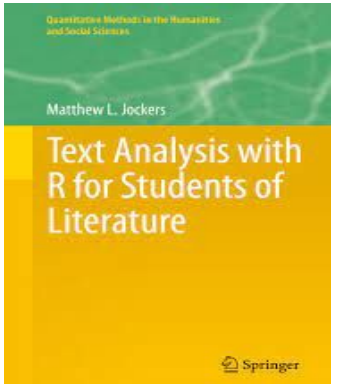
다차원 분석(Multidimenainal Analysis)

- Biber와 공동연구자 연구
- 텍스트 타입 또는 사용역(register)에 따른 어휘문법적 요소의 다차원 분석
- 특정 텍스트 장르별 언어적 특성 파악
- 언어 평가 등에 활용됨

2. 문학 텍스트 연구의 디지털화(디지털 인문학)

- Jockers (2013): Macroanalysis: Digital methods and literary history
- 텍스트를 작가, 주제, 시대, 성별, 지역별 등으로 분류 가능
- 텍스트의 언어적 자질 분석 및 비교 가능
- 군집 분석, 감성 분석, 토픽 모델링 등의 빅데이터 분석 기법 활용





본 발표에서는...

1. SF 코퍼스 구축 및 평가
2. SF 코퍼스의 디지털인문학적 분석 사례(작가 성별 중심으로)

Hallym University

연구 방법

Hallym University

영어 SF 코퍼스 구축 및 평가

- Biber(1993)과 Egbert(2019)의 코퍼스 구축 절차 참고
- 디지털인문학 전문가 3명, 영문학, 국문학 석사/박사 과정생 6명 참여
- 대략 3개월 소요

1. **작품 선정 목적**: (포스트) 아포칼립스적 분위기를 표현하는 문장을 생성하는 AI 알고리즘 학습을 위한 SF 소설

2. **작품 목록 리스트 작성 및 검토**: 대략 20세기~현재까지 출판된 영어 SF 소설, 인류(사회) 멸망의 위기(주제), 디지털화 변경 용이성, 작가 성별 고려

3. **212편 선정하고 각 작품별 텍스트 및 메타정보**(제목, 작가명, 출판년도 등)를 텍스트 파일로 정리

영어 SF 코퍼스 구축 및 평가

메타정보

```
Title:Dagon
Author:H. P. Lovecraft
Gender of Author:Male
Publication Year:1917
Publisher:
Publication Year of the Edition Used:1917
Publisher of the Edition Used:
Index Number:C114
Source Link:https://www.hplovecraft.com/writings/texts/fiction/d.aspx
```

시작

```
*** START OF THIS TEXT Dagon ***
```

텍스트

```
I am writing this under an appreciable mental strain, since by tonight I shall be
no more. Penniless, and at the end of my supply of the drug which alone makes
life endurable, I can bear the torture no longer; and shall cast myself from this
     The end is near. I hear a noise at the door, as of some immense slippery
body lumbering against it. It shall not find me. God, that hand! The window! The
window!
```

끝

```
*** END OF THIS TEXT Dagon ***
```

〈그림 1〉 텍스트 파일 스크린샷

영어 SF 코퍼스 구축 및 평가

- 영어 SF 코퍼스의 **대표성(representativeness)** 평가
- 새롭게 구축된 코퍼스가 해당 도메인(domain)의 언어를 반영하는 지 여부
- 기존 SF 코퍼스와의 어휘 비교를 진행: The SF Nexus Corpus
  - Wermer-Colan & Kopaczewski (2022): https://huggingface.co/datasets/SF-Corpus)
  - 403편의 영어 SF 소설 코퍼스(1908~2015)

〈표 1〉 코퍼스 비교

	작품수	단어수(token)	타입수(type)	공통 타입수(type)
영어 SF 코퍼스	212	1,252,957	51,018	35,151 (68.90%)
The SF Nexus Corpus (비교 코퍼스)	403	28,380,932	154,619	약 70%가량의 단어 포함

영어 SF 코퍼스 분석 절차

- 영어 SF 코퍼스의 언어적 특성 분석: **코퍼스언어학과 자연언어처리(Natural Language Processing, NLP) 기술을 활용**
- 파이썬(Python), 분석관련 라이브러리 사용

1. **기술 통계(빈도) 분석**: 어휘문법적 언어적 요소의 빈도와 분포; textstat(https://github.com/textstat/textstat) 사용(44개의 언어 특성 요소 추출가능)
2. **토픽 모델링**: 코퍼스 내의 주요 주제 식별(Blei, 2012); gensim, mallet topic modeling(idamallet) 사용
3. **다차원 분석(multi-dimensional analysis)**: 텍스트나 언어 사용의 다양한 특징을 파악하고 분석하는 통계적 방법으로 장르나 사용역(register) 분류(Biber, 1988)

연구 결과 & 토의

성별에 따른 언어적 특성을 중심으로

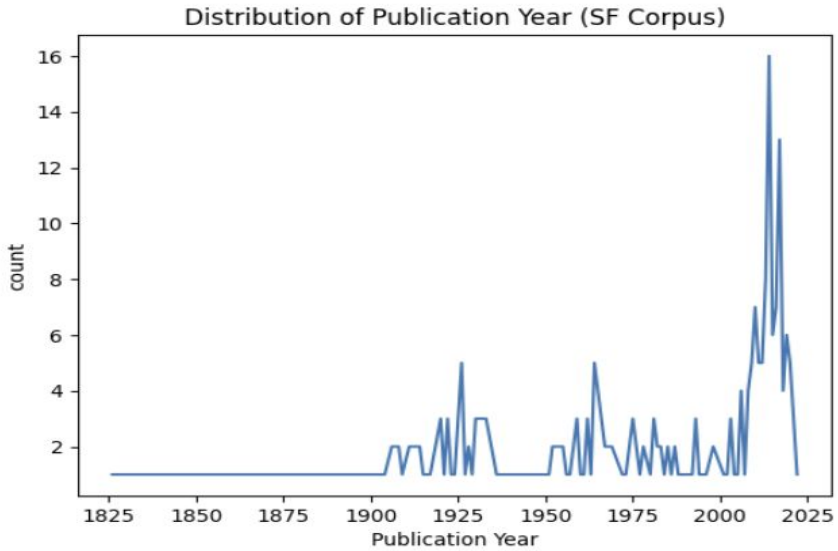
Hallym University

연구 결과 & 토의: 기술 통계

- 출판 연도별 분포

  - 1826년 ~ 2022년

  - 주로 20세기, 21세기 작품(예외: The Last Man (1826), Underground Man(1896))

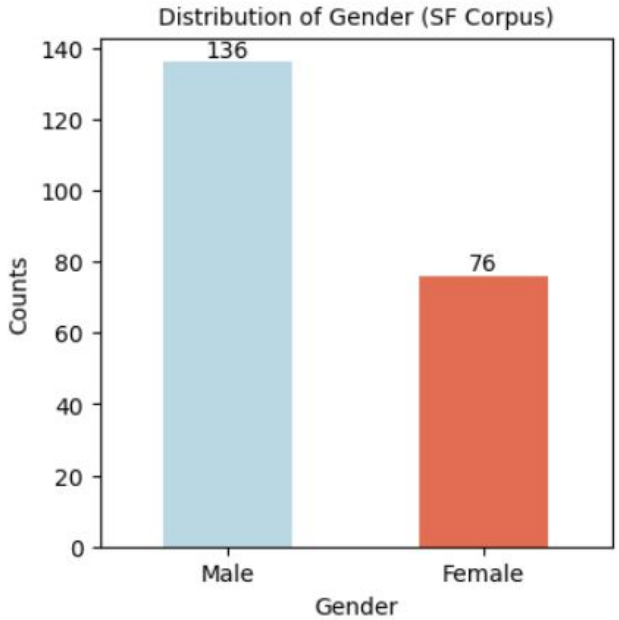


〈그림 2〉 출판연도별 텍스트 수

연구 결과 & 토의: 기술 통계

- 성별 분포

  - 남성 작가(136, 64.15%) vs. 여성 작가(76, 35.85%)

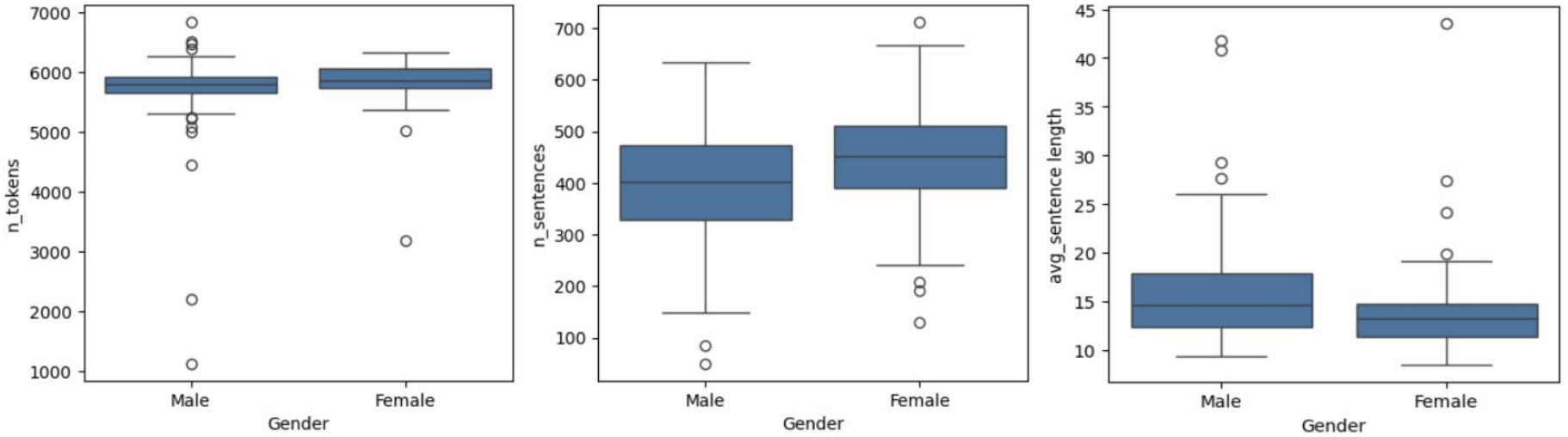


〈그림 3〉 작가 성별 텍스트 수

연구 결과 & 토의: 기술 통계

- 성별 언어 특성 분포

  - textstat (44 linguistic features 파일): Basic Statistics(기본 통계), Readability(읽기 난이도)

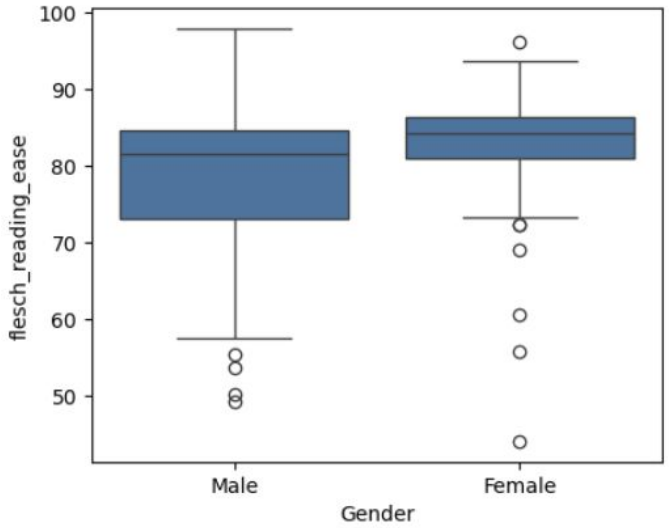


<그림 4> 단어수, 문장수, 문장 평균 길이

연구 결과 & 토의: 기술 통계

- 성별 언어 특성 분포

  - textstat (44 linguistic features): Basic Statistics(기본 통계), Readability(읽기 난이도)



〈그림 5〉 Flesch Reading Ease (읽기 난이도)

연구 결과 & 토의: 기술 통계

- ChatGPT로 생성한 남성/여성 SF 텍스트 언어 특성 분석

  - Generation date: December 15, 2023

  - Prompt: You are a (fe)male writer. Please write a science fiction about artificial intelligence.

  - 생성 텍스트 파일

  - LFTK 패키지 사용

다양한 부분에서 차이점을 살펴볼 수 있음; 성별에 따른 문체 분석 등의 정보가 LLM 학습에 사용된 듯

	Title	Gender	n_tokens	n_unique_tokens	n_sentences	avg_sentence_length	flesch_reading_ease	n_entities	readting_time_average
0	The Hear of Astra	Female	473.0	226.0	23.0	20.565217	55.216	35.0	1.971
1	Echoes of Tomorrow	Male	926.0	379.0	50.0	18.520000	57.483	72.0	3.858

연구 결과 & 토의: 기술 통계

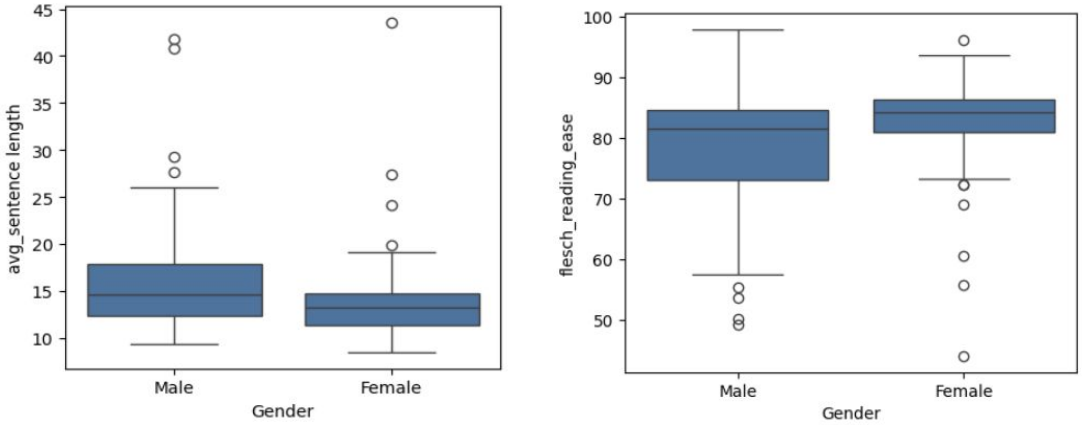
- **통계 분석(이집단 독립표본, two-sample t-test)**

  - **평균 문장 길이(남성/여성 작가의 평균 문장 길이에는 통계적으로 유의미한 차이가 있음)**

  - Men authors were more likely to write longer sentences than female authors (t = 2.67, p = 0.008).

  - **Flesch Reading Ease(남성/여성 작가의 텍스트의 읽기 난이도에는 통계적으로 유의미한 차이가 있음)**

  The readability of female texts is higher than that of male texts (t = - 3.425, p = 0.001).



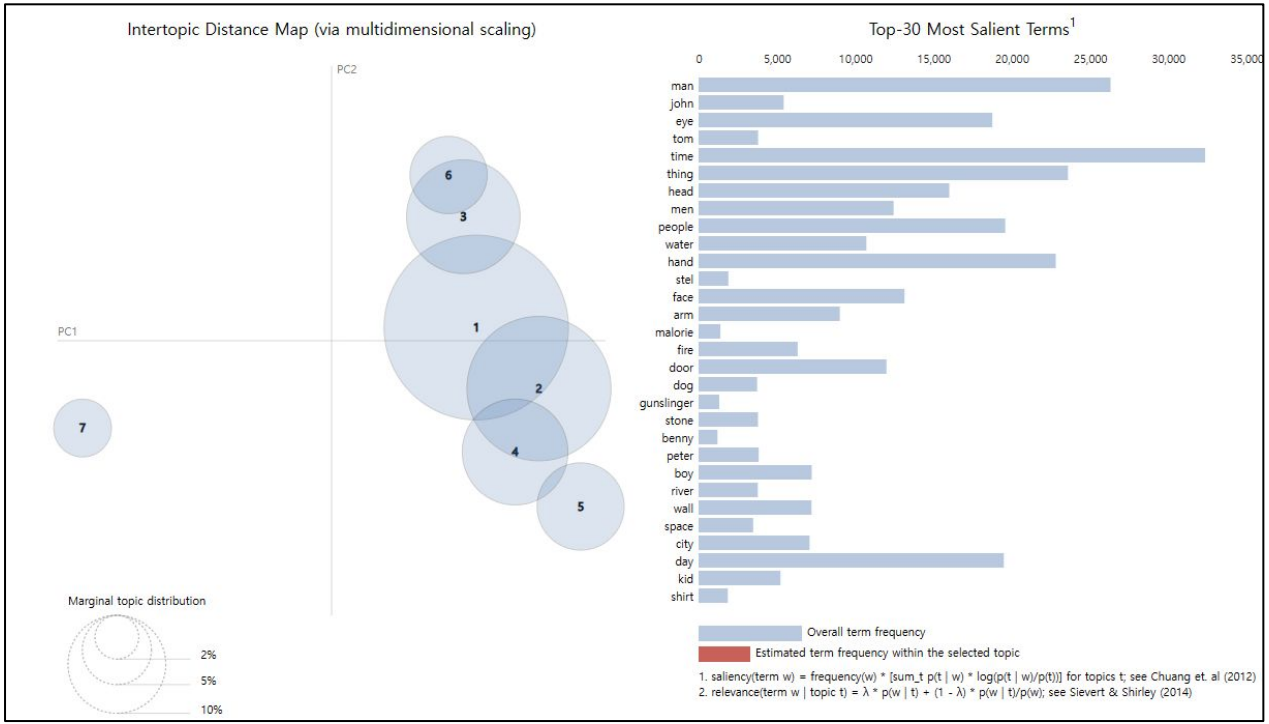
연구 결과 & 토의: 토픽 모델링

- 토픽 모델링: 대규모 텍스트 데이터에서 숨겨진 주제 구조를 발견하는 프로세스의 한 종류(Blei, 2012; Blei, Ng, & Jordan, 2003)
- 잠재 디리클레 할당(LDA): 토픽 모델링에서 흔히 쓰이는 비지도학습 알고리즘의 한 종류; 텍스트에서 단어의 출현이 곧 문서의 주제에 의해 결정된다고 가정하고, 자주 나타나는 단어의 그룹을 하나의 토픽으로 간주. 토픽은 텍스트 속 단어의 출현 비율에 의해 정해짐
- 주제의 분석을 통해 SF 문학의 다양한 측면을 이해하고, 인간과 AI가 어떻게 이 주제들을 창작에 활용할 수 있는지 탐색할 수 있음
- 남성과 여성 작가들이 사용하는 어휘/주제 선택에서 일반적인 성별 차이를 보여줄 수 있으며, 이는 그들의 작품에서의 테마, 스타일, 그리고 이야기 구성에 영향을 미칠 수 있음

주의사항: 텍스트가 적어서 토픽 모델링 결과는 참고만 하는 것으로!

More on topic modeling process; 김기연 & 한수미(2022) 논문 참고

연구 결과 & 토의: 토픽 모델링

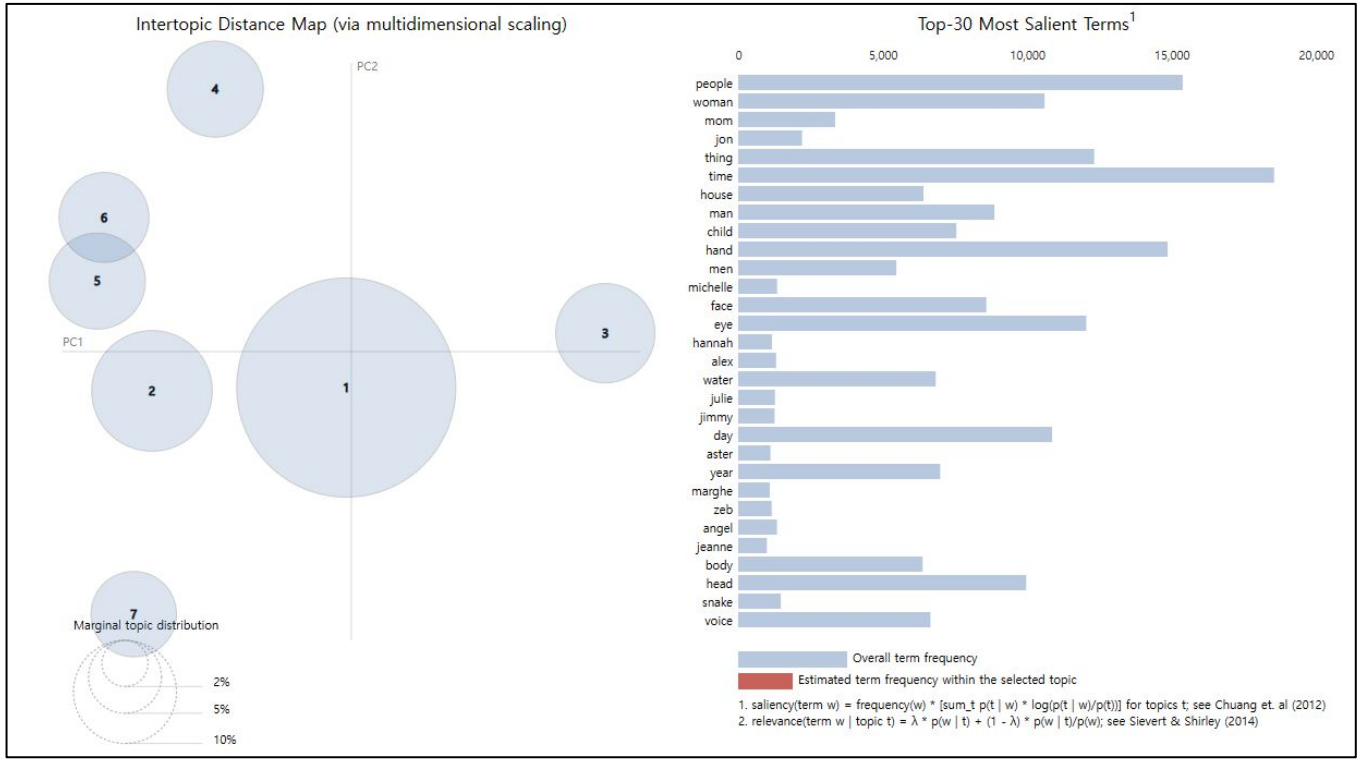


- 주제 7 그룹: 좀 더 나눠져야할 듯

- 주요 30개 단어: man, john, eye, tom, time, thing, head, men, people, water, hand, stel, face, arm, malorie, fire, door, dog, gunslinger, stone, benny, peter, boy, river, wall, space, city, day, kid, shirt

〈그림 6〉 남성 텍스트 - t-SNE 시각화

연구 결과 & 토의: 토픽 모델링



- 주제 7 그룹: 적절한 분포
- 주요 30개 단어: people, woman, mom, jon, thing, time, house, man, child, hand, men, michelle, face, eye, hannah, alex, water, julie, jimmy, day, aster, year, marghe, zeb, angel, jeane, body, head, snake, voice

〈그림 7〉 여성 텍스트 - t-SNE 시각화

연구 결과 & 토의: 토픽 모델링

- 공통 단어: 일반적인 SF 소설에서 흔히 볼 수 있는 요소들로, 인간, 물리적 환경 등을 나타내는 단어 ("time", "thing", "man", "people", "hand", "face", "eye", "water", "day" 등)

- 남성 텍스트 단어: 물리적 공간, 액션, 그리고 남성적 이미지와 관련된 단어로, 이는 남성 작가들이 더 기술적이고, 동적인 요소를 강조하는 경향 보여줌
("john", "tom", "head", "men", "stel", "arm", "fire", "door", "dog", "gunslinger", "stone", “peter", "boy", "river", "wall", "space", "city", "kid", "shirt" 등)

- 여성 텍스트 단어: 인간관계, 감성적 요소, 그리고 인물들의 내적 세계를 나타내는 경향, 여성 작가들이 인간 중심의 이야기, 감정의 표현, 그리고 세밀한 인물 묘사에 더 집중하는 경향을 나타낼 수 있습니다.
(“woman", "mom", "house", "child", "michelle", "hannah", "julie", "aster", "angel", "jeane", "body", "snake", "voice" 등)

연구 결과 & 토의: 다차원 분석

- **다차원 분석(Multi-Dimensional Analysis, MDA)**

  - 언어학에서 텍스트나 언어 사용의 다양한 특징을 파악하고 분석하는 통계적 방법

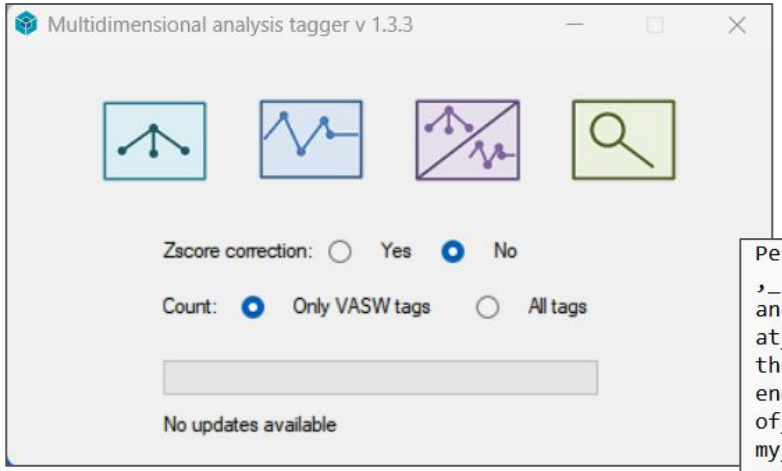
  - 요인 분석(Factor Analysis)을 바탕으로 텍스트나 언어 사용 패턴 식별하고, 특정 언어 스타일이나 사용역(register)를 분석(Biber, 1988; Biber & Egbert, 2016; Hardy & Friginal, 2016; Huang & Ren, 2020; Qian, 2022)

  - **Multidmensinal Aanalysis Tagger**(MAT, Nini, 2019)

    - SF 소설 코퍼스를 자동으로 POS 태깅, 67개의 어휘문법적 언어 자질을 수치화함

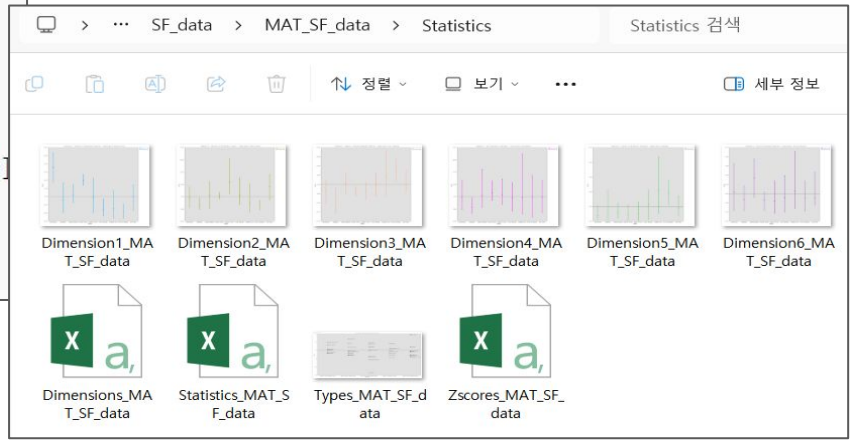
    - Biber(1988, 1989)에 따른 **6개 차원(Dimension)과 8개 텍스트 타입**으로 분류 시각화 가능

연구 결과 & 토의: 다차원 분석




```
Penniless_NN
,_,
and_ANDC
at_PIN
the_DT
end_NN
of_PIN
my_FPP1
supply_NN
of_PIN
the_DT
drug_NN
which_WDT [WHSUB]
alone_RB
makes_VPRT
life_NN
endurable_JJ
```

- 텍스트 파일로된 코퍼스 자료 입력
- POS 태깅 파일 생성
- 7개의 어휘문법적 언어 자질 표준 점수 (Z-score) 파일 생성
- 텍스트별 유사 text type 자동 분류



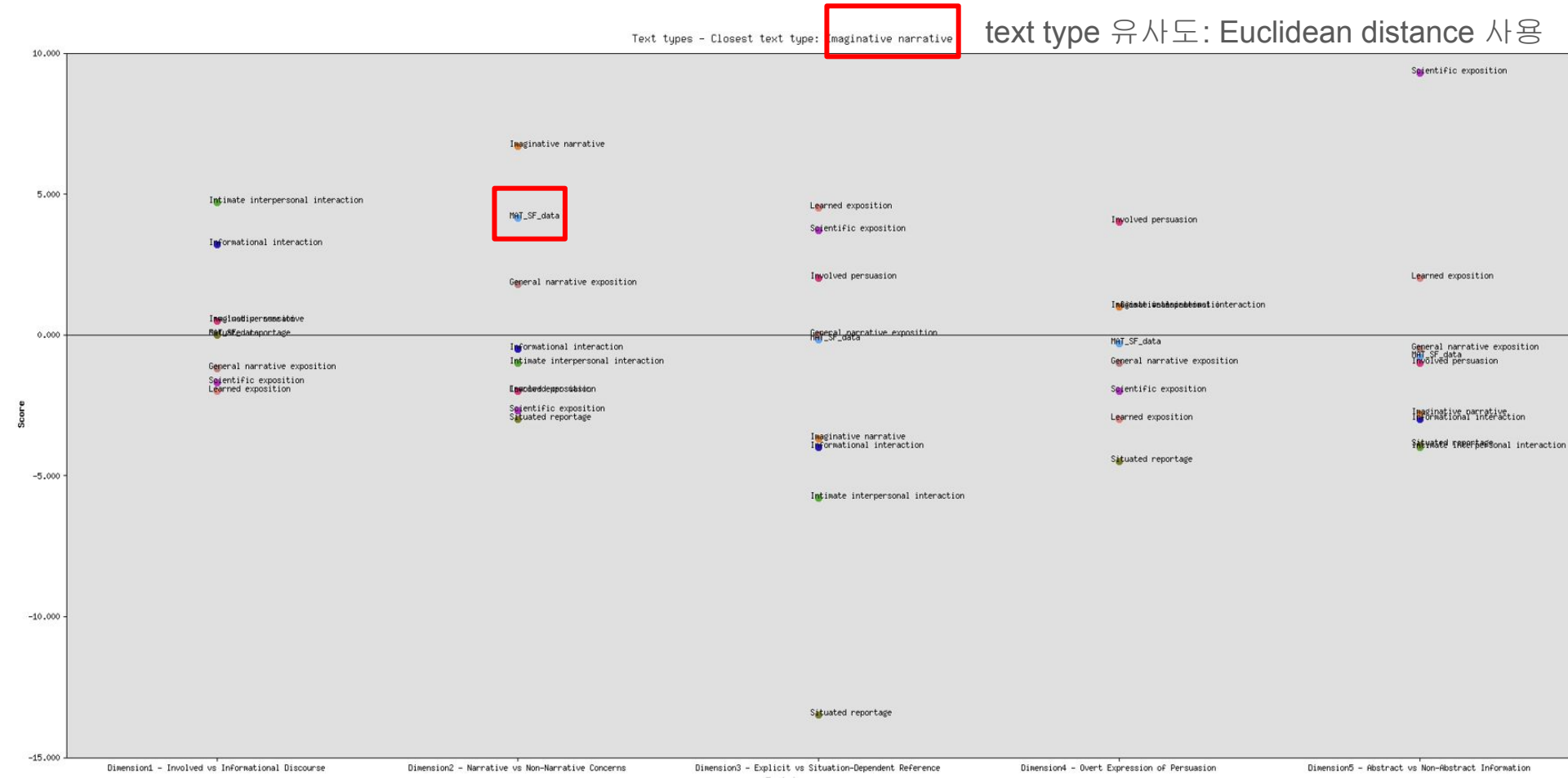
연구 결과 & 토의: 다차원 분석

Z-score

	Filename	Gender	Publication Year	Dimension1	Dimension2	Dimension3	Dimension4	Dimension5	Dimension6	Closest Text Type
113	C114_Dagon	Male	1917.0	-7.39	1.42	3.96	-3.42	0.38	0.16	General narrative exposition
114	C115_Nyarlathotep_H_P_Lovecraft	Male	1920.0	-5.55	0.42	4.21	-6.51	-2.42	0.48	General narrative exposition
115	C116_The_Whisper_in_the_Darkness	Male	1930.0	-3.99	1.82	2.78	-0.02	-0.20	1.17	General narrative exposition

- 전체 212편의 텍스트 중
  - 약 56%(118편): Imaginative narrative
  - 약 35%(74편): General narrative exposition

해석 참고:
6개 차원(Dimension)과 8개 텍스트 타입


Text types - Closest text type: Imaginative narrative
text type 유사도: Euclidean distance 사용
Score
10.000
5.000
0.000
-5.000
-10.000
-15.000
Dimension1 - Involved vs Informational Discourse
Dimension2 - Narrative vs Non-Narrative Concerns
Dimension3 - Explicit vs Situation-Dependent Reference
Dimension4 - Overt Expression of Persuasion
Dimension5 - Abstract vs Non-Abstract Information
Imaginative narrative
Intimate interpersonal interaction
Informational interaction
MAT_SF_data
General narrative exposition
Scientific exposition
Learned exposition
Involved persuasion
Situated reportage

Dimensions

연구 결과 & 토의: 다차원 분석

- 남성과 여성 작가의 언어 사용에 있어서 통계적 유의미한 차이가 나타남
- Dimension 1(Involved vs. Informational Discourse) (t= -3.64, p=0.00)

  Dimension 4(Overt Expression of Persuasion) (t = -2.13, p = 0.03)

  Dimension 5(Abstract vs. Non-abstract Information) (t = 2.77, p = 0.01)

  Dimension 6(On-Line Informational Elaboration) (t = 6.21, p = 0.00)

결론 & 후속 연구 제언

Hallym University

요약

- 문학 텍스트 코퍼스의 구축 과정, 평가, 그리고 디지털인문학 분석에 대해서 살펴봄
- 디지털인문학 분석 방법을 적용하여 코퍼스 내의 언어적 구조와 특성 파악
  - 성별에 따른 언어적 특성 확인
- 본 연구는 AI 소설 생성 연구의 기초 연구로, 향후 분석 결과를 인간과 AI가 협업하여 소설을 창작하는 과정에 반영할 예정임

후속 연구 제언

- 텍스트 선정 기준을 명시적으로 설정(문학적 가치, 영향, 역사적 중요성 등 고려)
- 메타정보 수집시 '성별 정보'에 대한 정확성 확인 필요
- 시기별, 주제별 분석을 통해서 문체의 변화나 언어적 패턴 파악 가능
- 작가 문체 스타일을 반영하는 다양한 언어적 자질을 분석(발표 자료)

  (e.g., complexity, coherence, cohesion)
- 추가적인 분석 방법: Factor Analysis, Cluster Analysis, Sentiment Analysis 등
- 문학 텍스트 분석을 위한 파이썬 패키지를 제작하고 공유

참고문헌

김기연 & 한수미. (2022). 텍스트 마이닝 기법을 활용한 구글 플레이 스토어 영어 학습 앱 사용자 리뷰 분석. *디지털콘텐츠학회논문지, 23*(10), 1901-1908.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: CUP.

Biber, D. (1989). A typology of English texts. *Linguistics, 27*(1), 3-43. Stanford Tagger v. 3.1.5. Retrieved from: http://nlp.stanford.edu/software/tagger.shtml.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*(4), 243-57.

Biber, D., & Egbert, J. (2016). Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics, 44(*2), 95-137.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77-84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Egbert, J. (2019). Corpus design and Representativeness. In Sardinha, T. B., & Pinto, M. V. (Eds.), *Multi-dimensional analysis: Research methods and current issues,* 27-42.

Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

참고문헌

Hardy, J. A., & Friginal, E. (2016). Genre variation in student writing: A multi-dimensional analysis. *Journal of English for Academic Purposes, 22,* 119-131.

Huang, Y., & Ren, W. (2020). A novel multidimensional analysis of writing styles of editorials from China Daily and The New York Times. *Lingua, 235,* 102781.

Nini, A. (2019). The multi-dimensional analysis tagger. In Sardinha, T. B., & Pinto, M. V. (Eds.), *Multi-dimensional analysis: Research methods and current issues* (pp. 77-94). New York: Bloomsbury Academic.

Qian, Y. (2022). A stylometric approach to the interdiscursivity of professional practice. *Humanities and Social Sciences Communications, 9*(1), 1-11.

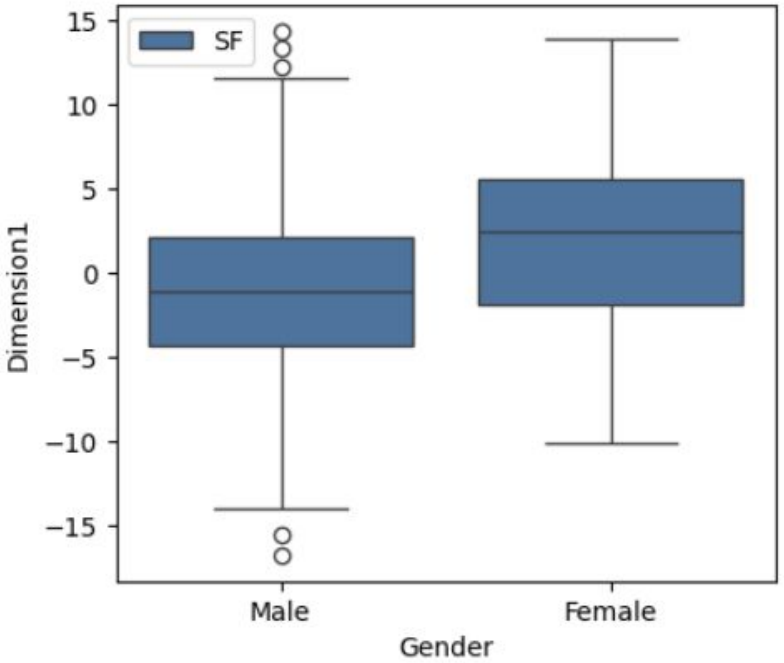
Wermer-Colan, A., & Kopaczewski, J. (2022). The new wave of digital collections: Speculating on the future of library curation. *Transactions of the American philosophical society, 110*(3), 211-241.

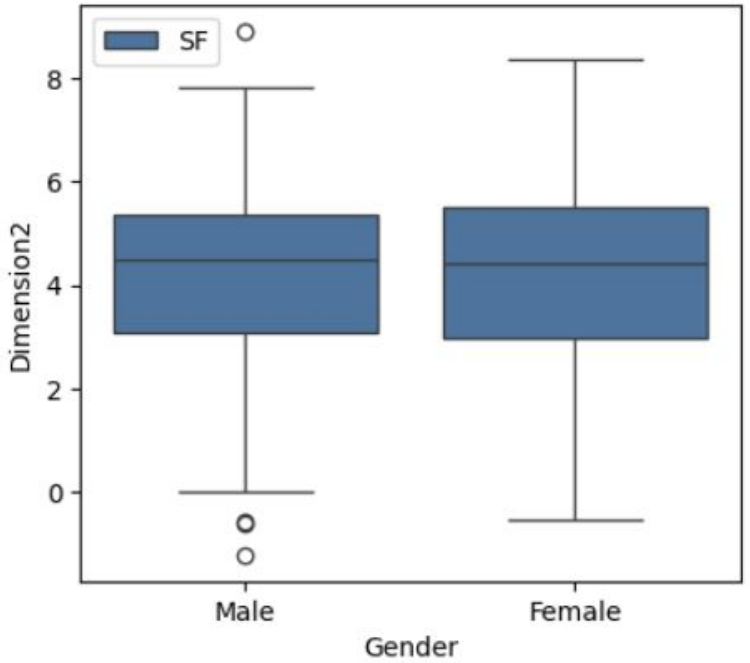
Thank you!

Presentation materials are available here:
https://github.com/SumiHan/Validation_Python_Packages_Text_Analysis

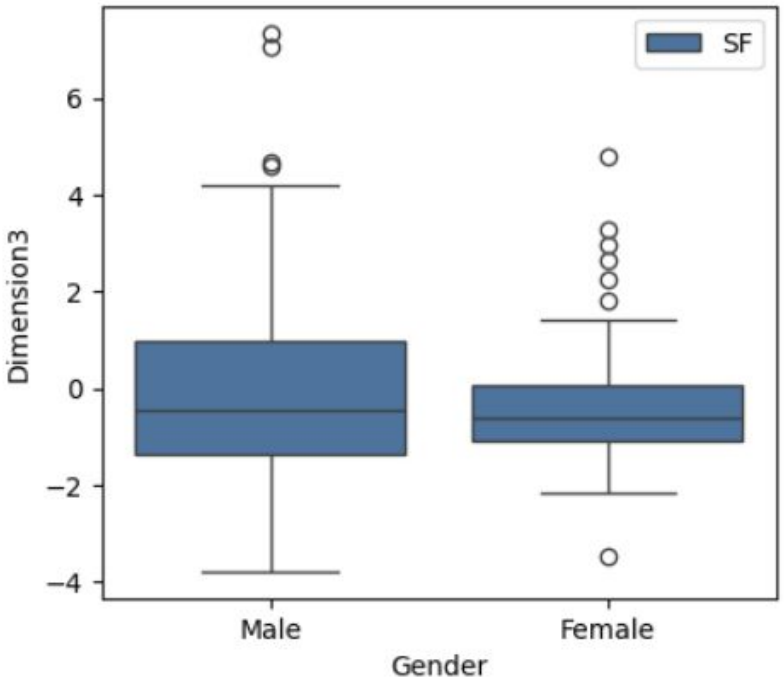
Contact me at sumihan20@gmail.com

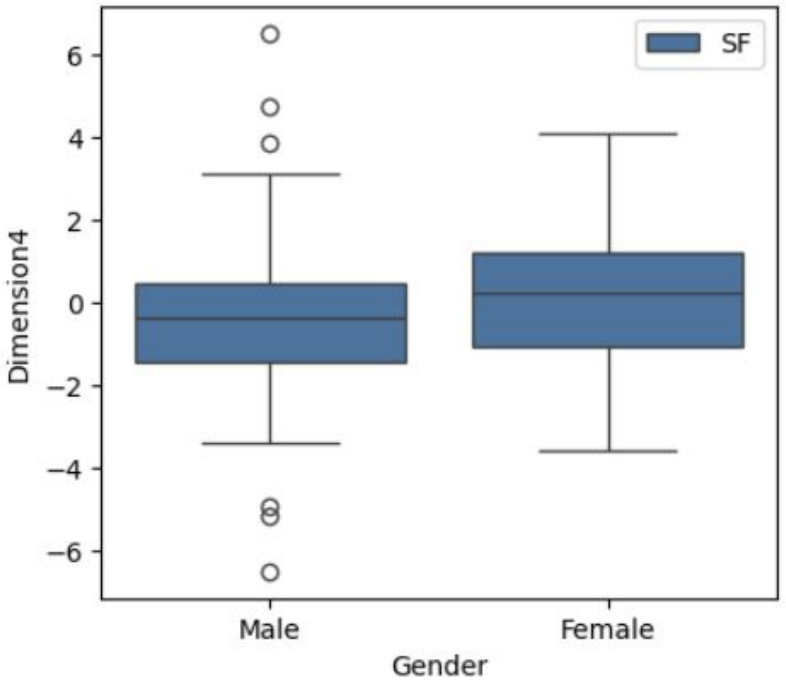
Dimensions by Gender





Dimensions by Gender





Dimensions by Gender