<IDEA>

개개인의 말하는 언어 습관에서 단어의 패턴을 찾고 특정 단어 및 습관들이

우울증이나 정서불안에 관련된 disorder들과 얼마나 관련되어 있는 지 분석한다.

Internet forum이나 SNS등에 게시된 post들로부터의 정서 분석은 이루어져 왔지만,

Utterance를 통한 언어습관에 대해서는 연구가 없는 것 같다.

사람의 글쓰기와 말하기에 사용되는 언어와 습관이 다르다고 가정하고,

말하기에서의 언어습관을 분석하여 우울증과 연관 짓는다.

(색깔은 중요도가 아니라 자료 검색한 시기가 다르다는 것을 명시)

1. 위 가정이 성립하려면 개인의 회화와 글쓰기에 사용되는 단어들의 출현빈도가

다르다는 것을 확증해야 한다.

- 관련 논문 찾기가 힘들다

# -https://www.omniglot.com/writing/writingvspeech.htm : Differences between writing and speech

# - 검색결과 자명한 이야기들을 찾을 수 있었다. 예를 들어 글을 쓸 때에는 생각할 시간이 있고 수정이 가능하지만 말하기는 그렇지 않고 바로 실행에 옮겨야 한다는 것이다. 허나 이것이 우리가 연구하고자 하는 우울증과 관련된 단어의 출현빈도에 대한 차이를 나타내는 지는 모호하다. 추가적으로 글을 쓸 때에는 formality가 중요시되기 때문에, 말하기와는 다르다는 것도 예측할 수 있다. 때 이른 결론을 내보자면 말하기와 글쓰기에 사용되는 단어의 출현빈도가 다르다는 것을 확정 짓기는 이를지라도, 자신의 의견을 convey하는 style에서는 확실히 차이가 있다는 것을 알 수 있다.

# -https://www.mtholyoke.edu/acad/intrel/speech/differences.htm (Speaking and Arguing: The Rhetoric of Peace and War) : 이 기사는 speaking보다는 speeach에 중점을 두었다. 이 기사 에 다음과 같은 절이 있다. Oral communication uses words with fewer syllables than the written language, the sentences are shorter, and self-referencing pronouns such as I are common. 이 말에 따르면 우울증과 관련이 없더라도 말하기에서 더 많은 1인칭 대명사가 사용된다.

# - 관련 자료를 찾다가 생각난 것인데, 일반인과 정서불안 환자의 filler사용 빈도가 다르지 않을까? –

# <https://www.researchgate.net/publication/314243349_Depression_Prediction_Via_Acoustic_Analysis_of_Formulaic_Word_Fillers> : 이 논문은 어렵다. 내가 ANOVA에 대해서 잘 몰라서 그런 것 같다. 요약하자면 말하기와 글쓰기의 차이점에 대해서 연구한 논문인데, 그 둘의 구분 지표를 4가지로 하였다. 번역하기가 애매한데, 이야기하고자 하는 것에 대한 개입성(involvement)과 객관성(detachment)의 비교와, 할 말을 잘 만들어서 하는지(integration) 아니면 쪼개어서(fragmentation) 말하는지를 비교하였다. 결과적으로는 말하기에서 involvement와 fragmentation 가 더 많이 발생하였고, 쓰기에서 integration이 더 많이 발생하였다. 허나 예상과 다르게 detachment는 말하기와 글쓰기에 사이에서 유의미하게 다르다는 것을 확증하기 어렵다는 결과가 나왔다.

# 실험은 8명의 여성으로 하였고, 2개의 말하기와 2개의 글쓰기를 task로 하였으며, 그 주제는 하나는 자신의 관한 일상 이야기(personal story)이고 다른 하나는 어떤 것의 관한 설명(explanation)이다. Story에서 involvement가 explanation에서 보다 월등히 많이 발생했고, fragmentation도 involvement보다 큰 차이지는 아니지만 유의미하게 많이 발생했다. 허나 integration은 explanation에 더 많이 발생하였다. Detachment는 explanation의 글쓰기에서 현저히 높게 발생하였고, 나머지에서는 비슷하게 발생하였다.

# Overall, 위에 말한 4가지 척도가 이를 테면 수동태의 사용 빈도, 자기 참조 빈도, and, but 절의 사용 빈도, 접속절, 종속절의 사용빈도에 따른 것인데, 글쓰기와 말하기에서의 문장구조와 스타일이 다르다는 것은 알 수 있으나, 특정 단어에 대하여 사용, 미 사용 여부를 알아내는 것과는 연관성이 떨어지는 논문 같다.

# 위 논문에서 detachment의 신뢰성이 떨어지는 이유 중 하나는 detachment가 참조한 subject가 4개뿐이 안되기 때문이다. (passive, past perfect, Indirect quotes, Literary expression)

# -또 생각난 것인데, 우울증 환자에 관련해서 말하기와 글쓰기 사이에 특정 단어 출현 빈도의 차이에 대한 연구가 없다는 것이 오히려 좋을 수도 있다. 내가 하면 되니까.

- 다르다는 것을 찾기 힘들면, 같다는 것을 찾아보자.

2. 2018년도 가장 최근에 출간된 언어와 우울증에 관한 논문이 있다. 이것은 절대적 언어와 정서불안의 상관관계에 조사했다.

이 논문에서는 1인칭 대명사와 부정적 감정표현의 출현 빈도가 정서불안과 연관 있다는

이때까지의 연구를 넘어 절대적인 단어, 절대론적 사고가 정서불안에 깊게 관련되어

있다는 사실을 밝혔다.

- multilevel mixed-effect model이 어떤 알고리즘인지 공부하자.

- 절대적 언어사용과 절대론적 사고의 연관관계에 대한 연구를 찾아보자.

3. 1번이 성립한다면 사람들의 회화에서 데이터를 수집해야 하는데, 정서불안 환자에

wearable device를 제공하여 일상대화를 모두 녹음하는 것은 현실적으로 불가능하기에,

정서불안 환자의 인터뷰나 동영상과, 통제집단으로 일반 사람들의 인터뷰나 일상생활을

담은 동영상을 비교하는 방법을 생각할 수 있다. 위 사람들의 회화를 어떻게 수집, 분류하여 데이터화 할 것인가는 고민 해봐야할 문제이다.

4. 2번 논문에서 사용된 LIWC : text 분석 method에 대하여 알아보자.

-The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods

위 논문은 text분석에 대한 자세한 알고리즘이 설명되지는 않는다.

-LIWC에 관한 다른 논문을 찾았는데, 아직 읽지 못했다.

-The Development and Psychometric : 이 논문이 저번에 못 읽은 논문인데 skim으로 읽어보니 LIWC가 단어를 분류하는 방법과, LIWC의 탄생 이유, 역사, 발전과정 등을 담고 있는 듯 하다

-LIWC 2007 version 에 대한 논문이 있어서 새로 읽어 보았다.

*Natural speech* . The speech samples included diverse transcripts from multiple contexts,

including people wearing audio recorders over days or weeks, strangers interacting in a waiting

room, couples talking about problems, and open-air tape recordings of people in public spaces. 이런 구절이 있다. 내가 필요로 하는 데이터를 가지고 있을 수도 있으니 참고하자.

5. 단어와 우울증 관련 기사 및 논문

# ①https://www.prima.co.uk/diet-and-health/healthy-living/news/a42519/people-with-depression-speak-differently/ : People with depression speak differently – these are the signs to look out for

--**More 'all or nothing' (aka absolutist) vocabulary**, including 'always', 'nothing' or 'completely'

# ②https://www.learning-mind.com/signs-of-depression-speech/ : How the Signs of Depression Show in Your Speech and Word Choice

- This suggests that **a depressed person is fixated on themselves**. They find it hard to focus on anyone else as the depression is overwhelming them. : 그들이 그들에게 집중하지 않게 하는 방식이 하나의 치료방법이 될 수 있지 않을까?

+ She also used metaphors to describe her darkest feelings. : 우울증 환자들은 은유법을 많이 사용한다고 한다.

6. 우울증과 과도한 은유법 사용 사이의 관계

<https://rosehendricks.com/2017/05/29/depression-its-metaphors/> :  [Conceptual Metaphor Theory](http://escholarship.org/uc/item/2p93q7nr#page-3)

7. 말하기와 우울증 관련 논문들

-Depression Speaks Automatic Discrimination between Depressed and Nondepressed

=> 발화과정에서의 비 언어적인 형질에 초점을 맞춘 논문이다.

=> 우울증 환자와 그렇지 않은 사람을 발화과정에서 비 언어적인 관점에 비추어 구분하는 것이 연구 목적이다. 우울증 여부를 제외한 성별, 받은 교육의 정도, 나이 등 모두 동일하게 구성하였다.

=>Diagnostic and Statistical Manual of Mental Disorders (DSMIV) : 알아보자

=> 이탈리아인인 실험대상들은 자신들의 일상생활 및 가족 등에 대하여 이야기하게 하여 자발적인 말하기에 대한 샘플을 얻어 냈고, 또한 Tale에 대하여 읽기를 통해 낭독 연설(read speech) 샘플을 얻어 내었다.

=> The feature extraction is performed with OpenSmile [19], a publicly available tool commonly adopted for the inference of social and psychological phenomena from speech : OpenSmile 알아보기

=> The experiments follow a leave-one-out approach: All subjects except one are used to perform the feature selection and to train the SVM aimed at doing the actual classification. : leave-one-out 알아보기

=> 비언어적인 특징에 중점을 두고 우울증 환자와 일반인을 75%의 확률로 가려낼 수 있다고 함. 이것은 인터뷰같이 진행되는 우울증 검사( Hamilton Rating Scale for Depression (HRSD), 이나 벡 우울증 척도 검사)등의 부작용을 줄여줄 수 있다.

=> 얻어진 샘플들은 일단 speech와 non-speech 부분으로 수동적으로 나뉘고, 나뉘어진 speech 부분(intervals)을 vector로 바꾸는 특성(feature)추출이 이루어진다. 그 후 linear kernel SVM에 넣어져서 우울증환자인지 통제집단인지 구분되게 된다. 이때 특성 추출은 OpenSmile을 사용했다.

=> Opensmile : speech로부터 사회적, 심리적 현상을 추측하기 위해 사용되는 common한 public tool

=> 자발적인 자기 이야기보다 동화 소리내어 읽기에서 자신에 대한 개입이 적고 생각할 시간의 여유의 필요성이 적기 때문에 더욱 더 높은 확률로 우울증환자와 일반인을 분류해낸다.

=> 여기서 말하는 분류 확률이 높다는 것은 우울증 환자가 우울증 환자인지 그리고 일반인이 일반인인지 분류해 낼 확률이 75%이상이라는 것이다.

=> 다르게 말하자면 자발적인 말하기에서 일반인을 우울증환자라고 오인할 가능성이 컸다.

즉, 우을증환자와 일반인을 비교하는 실험이 아니고, 각 사람이 병을 가지고 있는지 아닌지에 대한 분류를 하는 실험이다.

=> 결론적으로는 동화를 따라 읽을 때 가장 분류율이 좋았는 데, 그 이유는 말하는 데 계획을 세울 필요가 없어서 일반인이 우울증 환자처럼 마음이 동요하는 것처럼 보이는 현상이 줄었기 때문이다.

-Detecting Depression in Speech A Multi-classifier System with Ensemble

=> multi classifier system 개발에 중점을 둔 논문이다.

=> SVM 알아보기 : SVM is considered the current state-of-the-art classifier and has

been effectively used in modeling speech information for depression recognition

=> 단어에 관한 연구보다는 , 다양한 speech 상황에서 억양의 변화, 발화 속도, 톤의 변화, 불분명한 발음에 대한 연구 논문들과 새로운 방식의 우울증 관련 utterance 분류 방법을 소개하는 듯 하다.

=> kappa statistic 알아보기 : Smaller values of kappa (k) indicate a high diversity, k=0 indicates independent classifiers and k=1 indicates identical classifiers.

=> 긍정적, 중립, 부정적인 기사, 인터뷰, 그림, 비디오 클립을 통해 각각의 측정 방법에 대한 실험 대상의 반응을 연구했다.

=> BDI (벡의 우울증 자기진단 검사) score로 의사에게 우울증 판정을 받은 환자들을 평가했고 대부분 14점 이상이었다. 0-9점:우울하지 않은 상태10-15점:가벼운 우울 상태16-23점:중한 우울 상태24-63점:심한 우울 상태

=> 통제집단의 BDI score는 14 이하였다.

=>Voice features can be divided into acoustic and linguistic features [44]. In this experiment, only acoustic features were extracted since we wanted to analyses the characteristics of depressed speech regardless of the language used. : 단어나 언어에 대한 것보다는 발화의 음향학적 특질에만 중점을 두고 연구했다.

-Detection of Depression in Speech

=>k겹 교차검증에 대해 알아보자 : k-fold cross validation

=> 우울증 환자, 일반인, 고위험군(우울증에 걸릴 확률이 높은) 사람, 이렇게 세 부류의 그룹을 실험대상으로 하여 인터뷰, 이야기와 긍정적, 부정적, 중립적 단어 읽기, 긍정적, 부정적, 중립적 그림 묘사하기를 통해 실험대상들을 관찰했다.

=> The linguistic features will not be analyzed since we are aiming at general characteristics for

depressed speech regardless of the language used. Acoustic features could also be categorized into two branches: 이 논문도 언어학적으로는 분석안하고, 음향학적으로 분석함

=>결과가 없다

-https://theconversation.com/people-with-depression-use-language-differently-heres-how-to-spot-it-90877

-https://www.smithsonianmag.com/innovation/can-artificial-intelligence-detect-depression-in-persons-voice-180970702/ : Can Artificial Intelligence Detect Depression in a Person’s Voice? :That suggests that the words a person chooses is a better predictor of depression than how they sound.

-https://medium.com/datadriveninvestor/people-with-depression-use-language-differently-c9dce24b9230 : People With Depression Use Language Differently

=> This type of content is commonly found in written and spoken communications of those who are severely depressed and who are having suicidal thoughts.

<절대론 논문>

In an Absolute State: Elevated Use of

Absolutist Words Is a Marker Specific to

Anxiety, Depression, and Suicidal Ideation

-요점 :

절대적 언어 사용 -> 절대론적 사고 -> 정서적 장애

1인칭 대명사와 부정적 감정 표현이 많이 포함된 글이 해당 작성자의

우울증, 정서불안, 자살 가능성 등을 representive한다고 조사 되었었는데,

이 논문에서는 위 두가지보다 절대적인 function word와

절대론적 사고가 더 밀접하게 관련 있다고 말한다.

특히 통제집단보다 불안, 우울, 자살 증세가 있는 집단의 글에서

절대적인 단어가 더욱더 빈번히 출현하였고, 특히 자살에 관한 글에서

불안, 우울보다 의미 있게 많은 절대적 언어가 출현하였다.

또한 위와 같은 정서 불안에서 벗어난(recovery) 집단의 글에서 조차 통제집단보다 많은

절대적 단어를 사용했다. 다만 긍정적 감정에 대한 단어는 해당 recovery집단에서

가장 많이 출현했고 그 뒤를 통제집단이 따랐다.

또한 가정2에 대한 실험에서 절대적인 단어 사용은 심리학적 불안 증세보다는

절대론적 사고를 반영한다는 결과가 나왔다. 경계성 성격장애와 Eating disorder인

그룹 A와 외상 후 스트레스 장애와 조현병인 그룹 B 모두 심리적 장애를 가지고 있지만

그룹 A의 절대적 언어 사용이 그룹 B보다 월등히 높았다. 즉 그룹 A는 절대적 사고를

근간으로 하고 있고 그룹 B는 그렇지 않다는 말로써, 절대적 사고는 심리학적 불안

그 자체보다는 특정 심리적 불안 증세와 연관이 깊은 것으로 나타난다.

<LIWC와 텍스트 분석 논문>

The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods

-LIWC의 기본 분석 메커니즘을 설명하고, 사용되는 수많은 언어들을 psychometric, content와 style words, 사회학과 심리학적 의미의 대명사와 시제, 감정표현의 따른 정서분석, 대명사 사용 및 사회적 단어 사용과 social relationship의 관계, 지위 및 계층에 따른 언어 사용의 차이점, group processing, 정직도와 기만도, 파트너 간의 사랑정도, 사람의 사고 방법과, 외향적인지 내향적인지의 따른 언어사용의 차이 등의 근간을 두어 분석한다.

- The Psychometrics of Word Usage : 정신력측정용의 단어 사용은 신뢰성을 얻기 힘들다. 설문지는 같은 질문을 조금 다르게 하여 질문함으로써 답변에 대한 신뢰성을 추구할 수 있지만 정신력측정용 단어에 대하여 신뢰성을 얻기 위해서 반복적인 질문을 하면 사람들은 이전의 대답과 다른 대답을 한다. 자신의 심경이 변했거나, 이전과 같은 반복된 말을 하면 지루하고 또는 상대방이 무례하게 느낄 수 있다고 생각하기 때문이다.

-content versus style words : style words 는 대부분의 문장에서 55%이상을 차지한다. 매우 작은 양이지만 content word와는 다른 뇌의 작용을 거쳐서 쓰여진다. Style word는 사회관계에서 중요하게 여겨지고 그 사람의 상태, 지위, 정직도 등을 나타내어준다.

- Attentional Focus: Pronouns and Verb Tense : content word는 그 사람이 어디에 집중하고 있는지 설명해준다. 돈이든 섹스든 친구든. Style word도 어디에 주의하고 있는지 알려주는데 예를 들어 신체적, 정신적 고통을 받는 사람들은 자기 자신에 집중을 많이 하게 되고 그로 인해 1인칭 대명사를 많이 사용하게 된다. 또한 긍정적인 정치적 광고에 1인칭 대명사와 현대, 미래형 동사가 많이 사용되고 부정적 정치적 광고에는 3인칭 대명사와 과거형 동사가 사용된다. 즉 사람이 과거에 집중하고 있는지 미래나 현재에 집중하고 있는지 추측할 수 있다. 사람을 놀릴 때 놀림 받는 사람은 자신에게 집중하고 놀리는 사람도 놀림 받는 사람에게 집중한다. 결국 사용하는 인칭 대명사가 다르다. 또한 남성이 3인칭 대명사를 좀 더 많이 사용한다. 대명사와 시제는 특질과 의도, 과정을 분석하는데 중요한 언어적 요소이다.

- Emotionality: Positive and Negative Emotions : 감정적 단어 사용은 해당 주제에 더 몰두 할 수 있게 해준다.

-Social Relationships : 높은 직책에 있는 사람이 더 말을 많이 하고 1인칭 복수명사도 많이 사용한다. 낮은 직책의 사람들이 질문을 더 많이하고 1인칭 단수명사를 많이 사용한다. 또한 사람들은 예측과는 다르게 그룹에 응집(cohesion)될수록 1인칭 복수명사 사용이 줄어든다.

-Honesty and Deception : 사람들은 거짓말(기만)할 때 더 많은 부정적 감정의 단어와 motion word, 더 적은 exclusive word 와 1인칭 단수 대명사를 사용한다.

-너무 주제가 많아서 한번 읽은 것으로는 다 흡수하기 힘들다.

-연구하고자 하는 주제와는 사뭇 다른 방향의 논문이다. 언어 사용에 따른 수많은 분야에 대해서 얘기하지만, 정서불안에 관한 chapter는 적다. 하지만 conclusion에서 More of the research results have come from labs in the United States working with college-aged students, often in highly contrived settings. Very little work has explored the differences between spoken and written language. 라는 구절이 있는데, 이 의미가 회화와 글쓰기의 차이에 대한 연구 자체가 적다는 것인 지, 아니면 연구를 했으나 그 차이가 미미하다는 것인지 모호하다. 전자일 가능성이 크게 해석되지만, 만약 후자라면 1번 가정이 성립하지 않는다. 또한 전자여도 관련 논문을 찾기 힘들 것 같다.

<LIWC 2015>

2007버전에서 words와 dictionaries가 추가 및 개선됨

<The Development and Psychometric>

- 샘플 파일의 단어를 target word라고 하고 LIWC의 dictionary file에 있는 단어들을 dictionary word라고 한다. 한번의 하나의 파일을 읽고 그 안의 하나의 target word를 dictionary word와 비교하며 match 되었을 경우 해당 dictionary word가 속한 category의 count가 1씩 증가한다. 이때 한 단어는 여러 개의 category에 속할 수 있다. (출력 변수에 대해서는 해당 논문에 표로 참조되어 있다.)

-발전 과정 : 맨 처음에는 미리 존재하는 word lists를 참조했다. (PANAS (Watson, Clark, & Tellegen, 1988), Roget’s Thesaurus, and standard English dictionaries) 그 후 3~6명의 judges들이 머리를 맞대고 최초의 lists를 만들었다. 최초의 list로부터 judge들은 각 단어에 대하여 특정 category에 속할지 아닐지를 고민했고 3명 중에 2명이상이 찬성한 경우에만 해당 category에 속하게 하거나 새로 추가시켰다. 그러나 이렇게 고안된 LIWC의 초반 version은 신뢰성과 타당성이 높지 않고 다른 언어 분석 프로그램에 비해 많이 사용되지 않아 대대적으로 개편하여 LIWC 2007 version이 나왔다.

<참고하면 좋을 글귀 및 문헌>

People who are experiencing physical or emotional pain tend to have their attention drawn to themselves and subsequently use more first-person singular pronouns (e.g., Rude, Gortner, & Pennebaker, 2004).

Whereas personal pronouns provide information about the subject of attention, analyses of the tense of common verbs can tell us about the temporal focus of attention. In the same study of political ads, the authors found that positive ads used more present and future tense verbs, and negative ads used more past tense verbs (Gunsch et al., 2000)

In a reanalysis of six expressive writing studies, Pennebaker, Mayne, and Francis (1997) found that increasing use of causal and insight words led to greater health improvements.

Studies measuring personality in participants through writing samples (Pennebaker & King, 1999) and spoken dialogue (Mehl, Gosling, & Pennebaker, 2006) have shown that some LIWC categories correspond with big-five personality traits.

More of the research results have come from labs in the United States working with college-aged students, often in highly contrived settings. Very little work has explored the differences between spoken and written language.