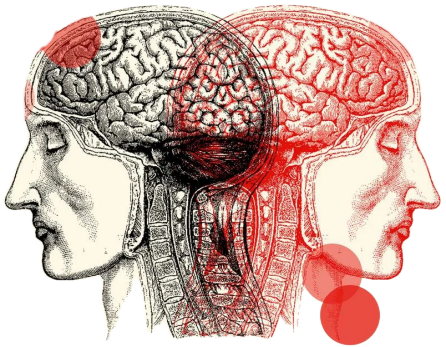
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Francisco Ribeiro | up202104797
Marisa Azevedo | up202108624
Ricardo Ferreira | up201907835
Toni Grgurevic | up202408625

Mental Disorders

Data Collection and Preparation

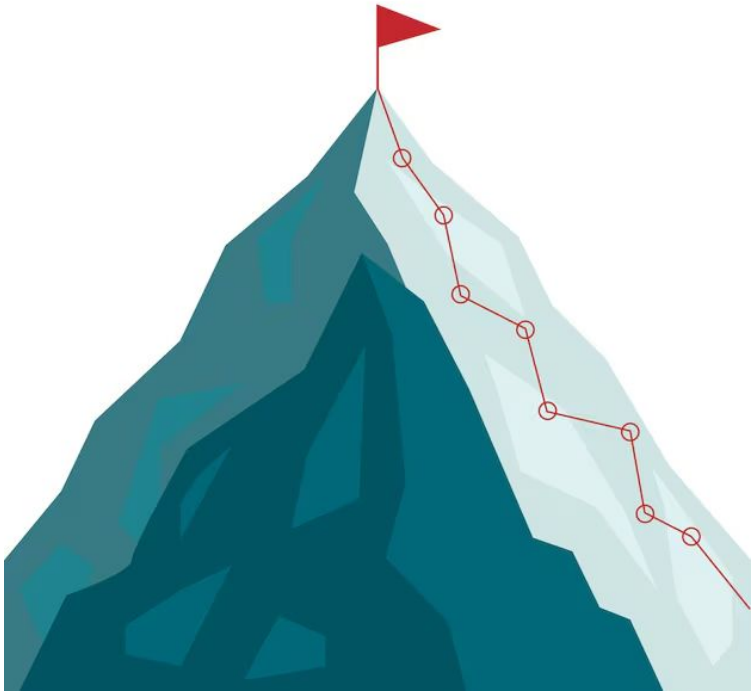


Objectives

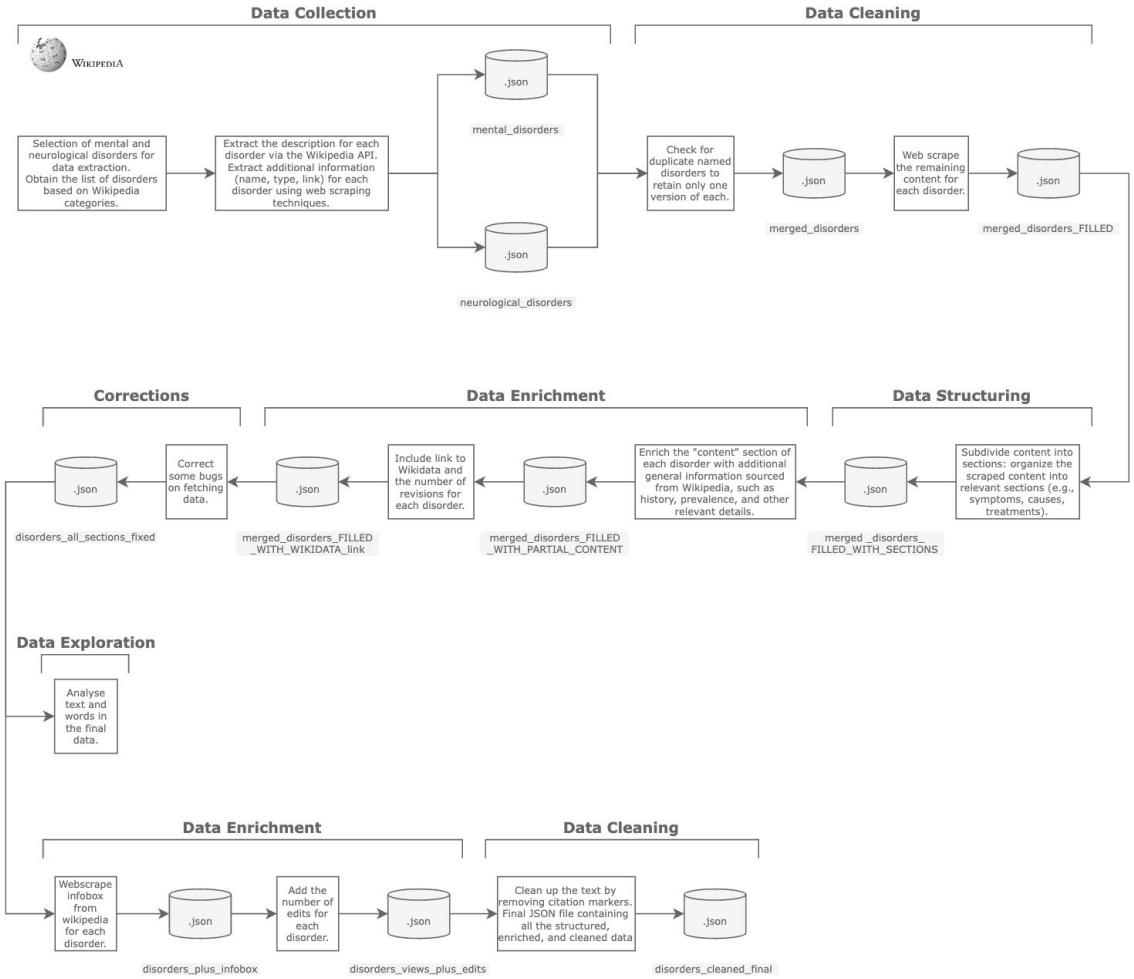
The project consists in design and implementation of an information processing and research system of Mental Diseases.

The questions we seek to answer are:

- What is [Disorder Name]?
  - Example: What is Agoraphobia?
- What are the symptoms of [Disorder Name]?
  - Example: What are the symptoms of **Bipolar Disorder**?
- I am feeling [symptom]
  - Example: I am feeling anxious



Pipeline

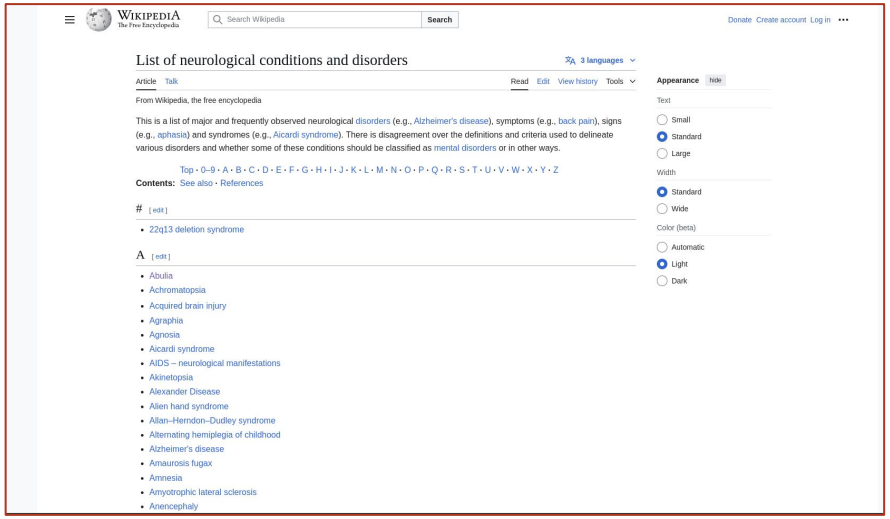

Data Collection
Data Cleaning
WIKIPEDIA
Selection of mental and neurological disorders for data extraction. Obtain the list of disorders based on Wikipedia categories.
Extract the description for each disorder via the Wikipedia API. Extract additional information (name, type, link) for each disorder using web scraping techniques.
.json
mental_disorders
.json
neurological_disorders
Check for duplicate named disorders to retain only one version of each.
.json
merged_disorders
Web scrape the remaining content for each disorder.
.json
merged_disorders_FILLED
Corrections
Data Enrichment
Data Structuring
Subdivide content into sections: organize the scraped content into relevant sections (e.g., symptoms, causes, treatments).
.json
merged _disorders_ FILLED_WITH_SECTIONS
Enrich the "content" section of each disorder with additional general information sourced from Wikipedia, such as history, prevalence, and other relevant details.
.json
merged_disorders_FILLED _WITH_PARTIAL_CONTENT
Include link to Wikidata and the number of revisions for each disorder.
.json
merged_disorders_FILLED _WITH_WIKIDATA_link
Correct some bugs on fetching data.
.json
disorders_all_sections_fixed
Data Exploration
Analyse text and words in the final data.
Data Enrichment
Data Cleaning
Webscrape infobox from wikipedia for each disorder.
.json
disorders_plus_infobox
Add the number of edits for each disorder.
.json
disorders_views_plus_edits
Clean up the text by removing citation markers. Final JSON file containing all the structured, enriched, and cleaned data
.json
disorders_cleaned_final

Data Collection

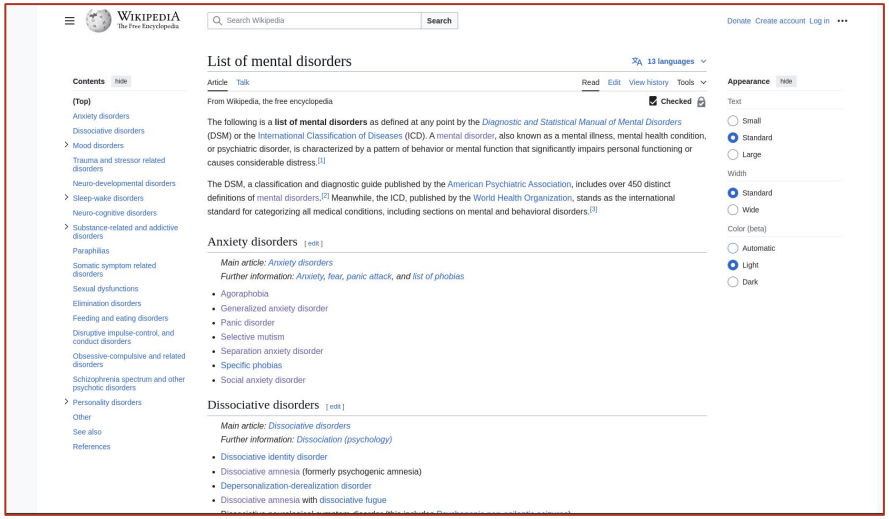
Collecting Information - Start Point

We find out two lists regarding mental disorders and neurological disorders, we scrape that information including name, link and type of disorder, and merged the list into one. We use JSON notation to store the list.

https://en.wikipedia.org/wiki/List_of_neurological_conditions_and_disorders



https://en.wikipedia.org/wiki/List_of_mental_disorders

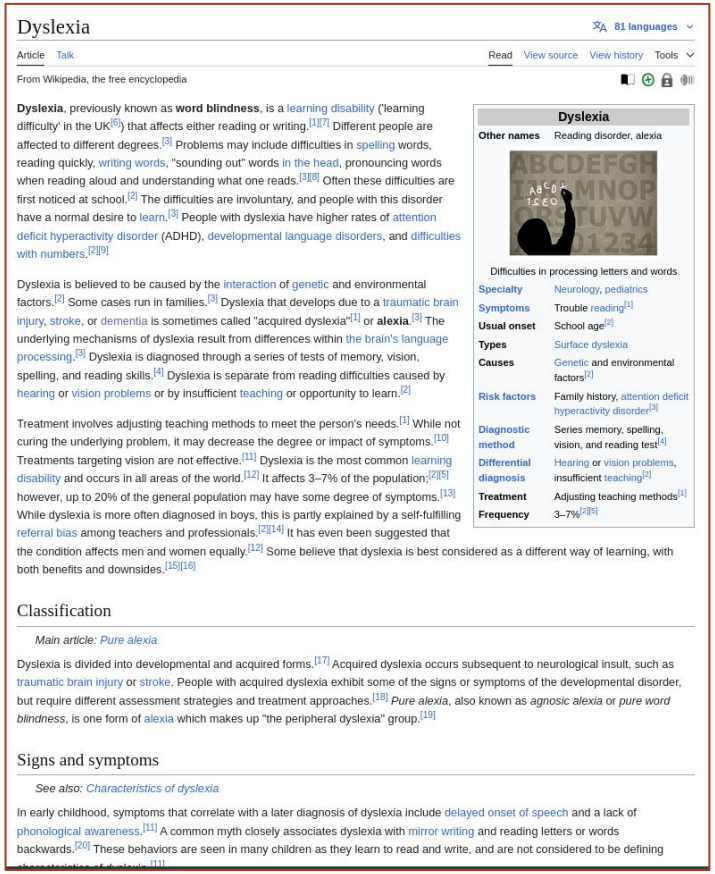


Collecting Information - Process

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

For gathering data, we follow a process consisting on the following topics:

1. Fill content and description of the merged JSON list of disorders.
2. Associate new information (sections organized) to the list of disorders.
3. Associate Wikidata link to a JSON list of disorders and get the number of revisions.
4. Edit the content, scrape again the wikipedia, but this time select only a part of content.
5. Add the infobox (Scrape from Wikipedia) as a new field containing the infobox data in JSON format.
6. Add the wikipedia number of edits and page views over the last 30 days to the JSON.
7. Remove redundant text from text fields, example [1], [2], [1], etc.



Dyslexia

81 languages

Article Talk Read View source View history Tools

From Wikipedia, the free encyclopedia

Dyslexia, previously known as **word blindness**, is a learning disability ('learning difficulty' in the UK[6]) that affects either reading or writing.[1][7] Different people are affected to different degrees.[3] Problems may include difficulties in spelling words, reading quickly, writing words, "sounding out" words in the head, pronouncing words when reading aloud and understanding what one reads.[3][8] Often these difficulties are first noticed at school.[2] The difficulties are involuntary, and people with this disorder have a normal desire to learn.[3] People with dyslexia have higher rates of attention deficit hyperactivity disorder (ADHD), developmental language disorders, and difficulties with numbers.[2][9]

Dyslexia is believed to be caused by the interaction of genetic and environmental factors.[2] Some cases run in families.[3] Dyslexia that develops due to a traumatic brain injury, stroke, or dementia is sometimes called "acquired dyslexia"[1] or **alexia**.[3] The underlying mechanisms of dyslexia result from differences within the brain's language processing.[3] Dyslexia is diagnosed through a series of tests of memory, vision, spelling, and reading skills.[4] Dyslexia is separate from reading difficulties caused by hearing or vision problems or by insufficient teaching or opportunity to learn.[2]

Treatment involves adjusting teaching methods to meet the person's needs.[1] While not curing the underlying problem, it may decrease the degree or impact of symptoms.[10] Treatments targeting vision are not effective.[11] Dyslexia is the most common learning disability and occurs in all areas of the world.[12] It affects 3–7% of the population;[2][5] however, up to 20% of the general population may have some degree of symptoms.[13] While dyslexia is more often diagnosed in boys, this is partly explained by a self-fulfilling referral bias among teachers and professionals.[2][14] It has even been suggested that the condition affects men and women equally.[12] Some believe that dyslexia is best considered as a different way of learning, with both benefits and downsides.[15][16]

Dyslexia	
Other names	Reading disorder, alexia
	Difficulties in processing letters and words
Specialty	Neurology, pediatrics
Symptoms	Trouble reading[1]
Usual onset	School age[2]
Types	Surface dyslexia
Causes	Genetic and environmental factors[2]
Risk factors	Family history, attention deficit hyperactivity disorder[9]
Diagnostic method	Series memory, spelling, vision, and reading test[4]
Differential diagnosis	Hearing or vision problems, insufficient teaching[2]
Treatment	Adjusting teaching methods[1]
Frequency	3–7%[2][5]

Classification

Main article: Pure alexia

Dyslexia is divided into developmental and acquired forms.[17] Acquired dyslexia occurs subsequent to neurological insult, such as traumatic brain injury or stroke. People with acquired dyslexia exhibit some of the signs or symptoms of the developmental disorder, but require different assessment strategies and treatment approaches.[18] *Pure alexia*, also known as *agnosic alexia* or *pure word blindness*, is one form of alexia which makes up "the peripheral dyslexia" group.[19]

Signs and symptoms

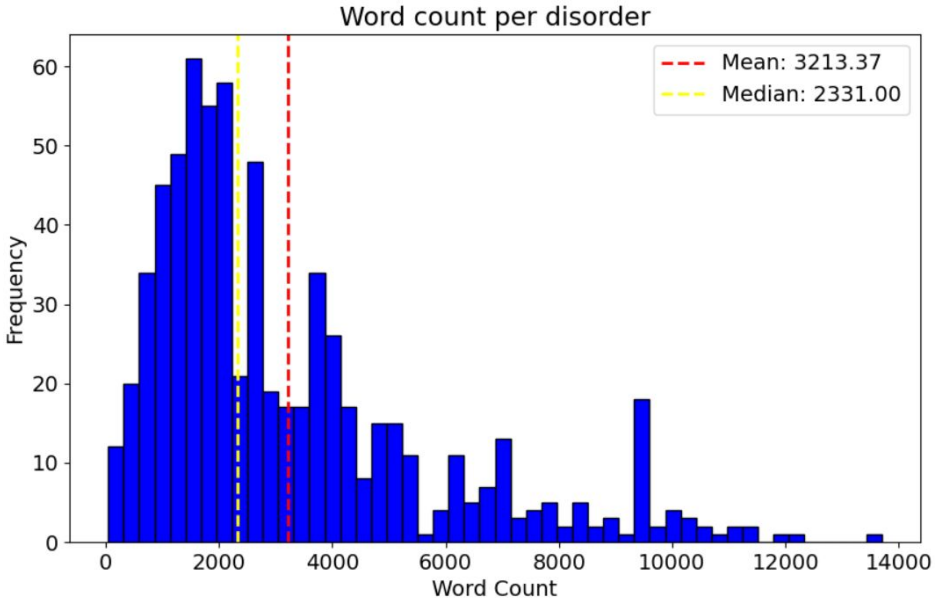
See also: Characteristics of dyslexia

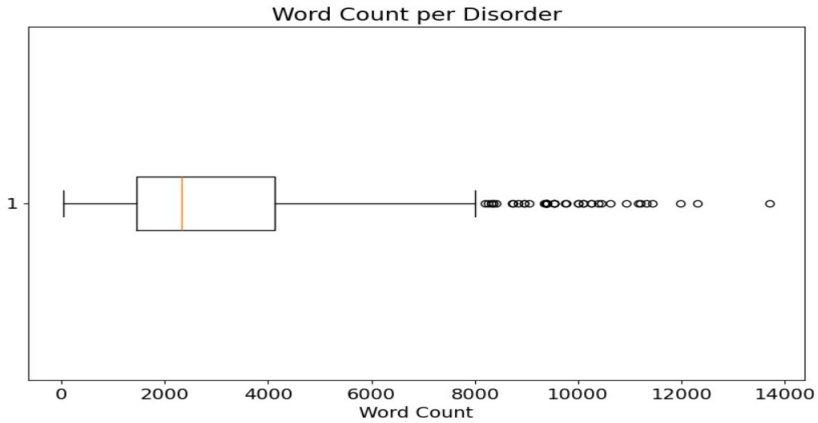
In early childhood, symptoms that correlate with a later diagnosis of dyslexia include delayed onset of speech and a lack of phonological awareness.[11] A common myth closely associates dyslexia with mirror writing and reading letters or words backwards.[20] These behaviors are seen in many children as they learn to read and write, and are not considered to be defining

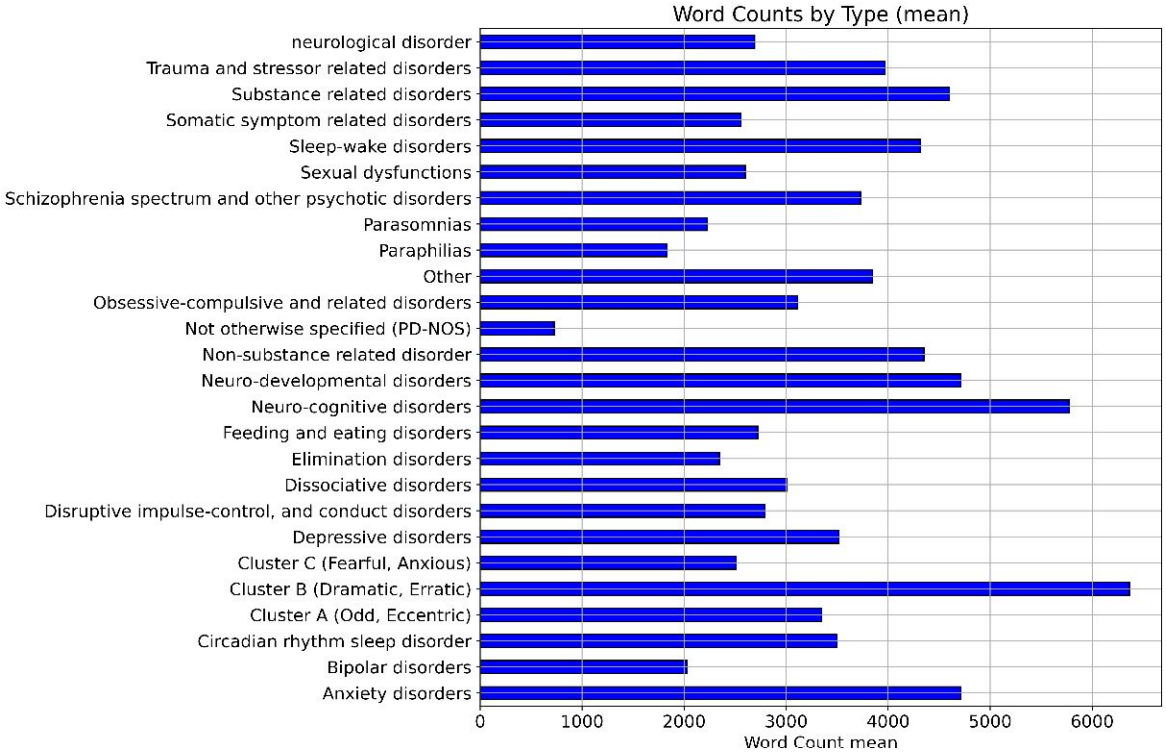
Final Document - Example

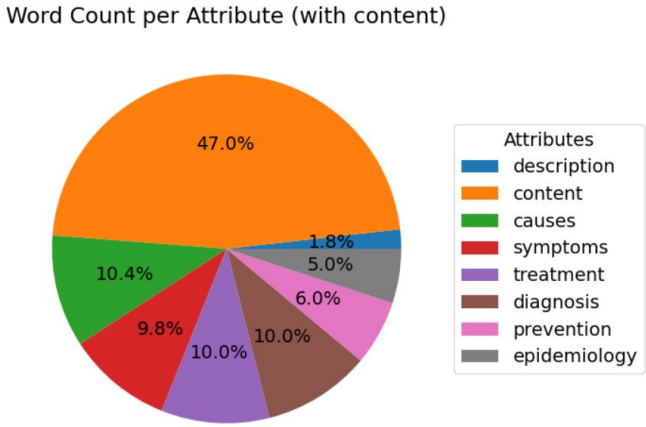

```
name:                "Agoraphobia"
type:                "Anxiety disorders"
link:                "https://en.wikipedia.org/wiki/Agoraphobia"
description:         "\n\nAgoraphobia is a mental and behavioral disorder, specifically an anxiety disorder (
                     public transit, shopping centers, crowds and queues, or simply being outside their home
                     become completely unable to leave their homes."
content:             `/ˌægərəˈfoʊbiə, əˌgɔːrə-…r agitation\nStereotypy`
causes:              "Agoraphobia is believed …panic attacks occurred."
symptoms:            `Agoraphobia is a conditi…g control of behaviors.`
treatment:           "Therapy\nSystematic dese…they should be avoided."
diagnosis:           'Most people who present … be diagnosed together.'
prevention:          ""
epidemiology:        "Agoraphobia occurs about twice as commonly among women as it does in men. It can develo
                     agoraphobia affects roughly 5.1% of Americans, and about 1/3 of this population with pa
                     disorders as well."
wikidata_id:         "Q174589"
wikidata_url:        "https://www.wikidata.org/wiki/Q174589"
wikidata_url_json:   "https://www.wikidata.org/wiki/Special:EntityData/Q174589.json"
number_of_revisions: "1249439475"
infobox:
  Pronunciation:     "/ ˌ æ g ər ə ˈ f oʊ b i ə ˌ ə ˌ g ɔːr ə -/"
  Specialty:         "Psychiatry , clinical psychology"
  Symptoms:          "Anxiety in situations perceived to be unsafe, panic attacks [ 1 ] [ 2 ]"
  Complications:     "Depression , substance use disorder [ 1 ]"
  Duration:          "> 6 months [ 1 ]"
  Causes:            "Genetic and environmental factors [ 1 ]"
  Risk factors:      "Family history, stressful event [ 1 ]"
  Differential diagnosis: "Separation anxiety , post-traumatic stress disorder , major depressive disorder [ 1 ]"
  Treatment:         "Cognitive behavioral therapy [ 3 ]"
  Prognosis:         "Resolution in half with treatment [ 4 ]"
  Frequency:         "1.9% of adults [ 1 ]"
number_of_edits:     2492
page_views:          73198
```

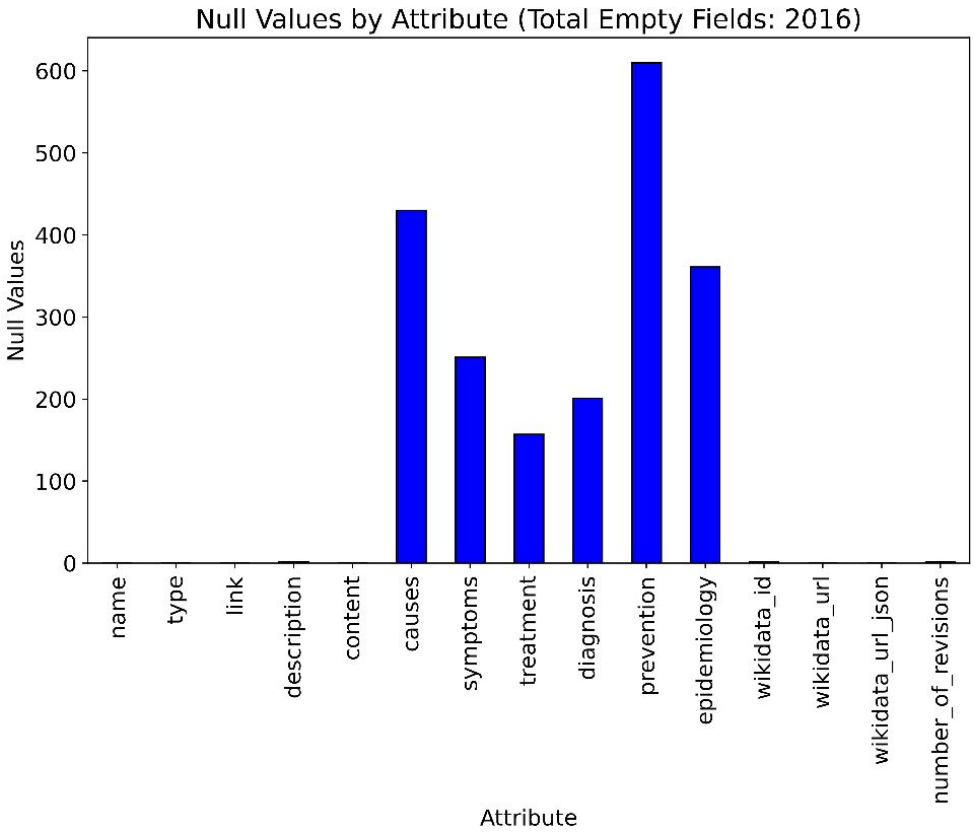
Data Analysis

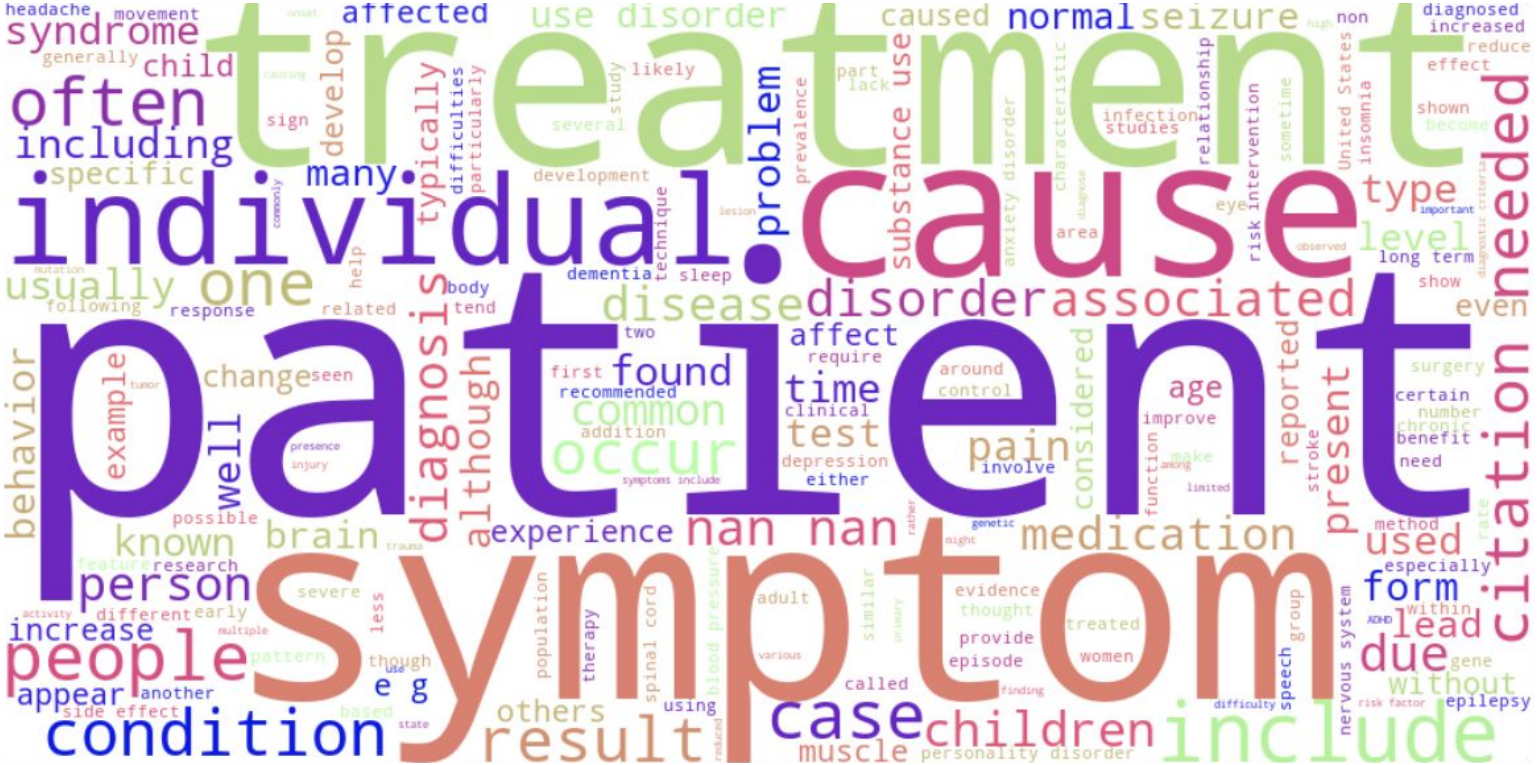

Word count per disorder
Mean: 3213.37
Median: 2331.00
Frequency
Word Count
0
10
20
30
40
50
60
0
2000
4000
6000
8000
10000
12000
14000


Word Count per Disorder
1
0
2000
4000
6000
8000
10000
12000
14000
Word Count


Word Counts by Type (mean)
neurological disorder
Trauma and stressor related disorders
Substance related disorders
Somatic symptom related disorders
Sleep-wake disorders
Sexual dysfunctions
Schizophrenia spectrum and other psychotic disorders
Parasomnias
Paraphilias
Other
Obsessive-compulsive and related disorders
Not otherwise specified (PD-NOS)
Non-substance related disorder
Neuro-developmental disorders
Neuro-cognitive disorders
Feeding and eating disorders
Elimination disorders
Dissociative disorders
Disruptive impulse-control, and conduct disorders
Depressive disorders
Cluster C (Fearful, Anxious)
Cluster B (Dramatic, Erratic)
Cluster A (Odd, Eccentric)
Circadian rhythm sleep disorder
Bipolar disorders
Anxiety disorders
0
1000
2000
3000
4000
5000
6000
Word Count mean


Word Count per Attribute (with content)
47.0%
1.8%
5.0%
6.0%
10.0%
10.0%
9.8%
10.4%
Attributes
description
content
causes
symptoms
treatment
diagnosis
prevention
epidemiology


Null Values by Attribute (Total Empty Fields: 2016)
Null Values
0
100
200
300
400
500
600
name
type
link
description
content
causes
symptoms
treatment
diagnosis
prevention
epidemiology
wikidata_id
wikidata_url
wikidata_url_json
number_of_revisions
Attribute


treatment
patient
symptom
individual
cause
citation
needed
disorder
associated
disease
diagnosis
although
occur
common
found
time
test
pain
medication
nan
nan
experience
person
people
condition
result
case
children
include
use disorder
normal
seizure
caused
affected
syndrome
often
including
many
specific
usually
one
behavior
example
well
known
brain
change
affect
age
type
level
form
lead
due
without
used
present
reported
considered
substance use
problem
typically
develop
child
headache
movement
diagnosed
increased
reduce
effect
shown
become
long term
show
even
surgery
certain
number
chronic
benefit
need
method
especially
within
gene
epilepsy
nervous system
speech
group
treated
women
evidence
thought
provide
episode
called
using
others
muscle
personality disorder
adult
population
therapy
spinal cord
blood pressure
less
severe
early
different
increase
appear
another
side effect
based
state
pattern
though
use
e g
feature
research
possible
presence
injury
seen
tumor
following
response
related
help
body
tend
dementia
sleep
technique
lesion
development
several
study
likely
part
lack
prevalence
anxiety disorder
characteristic
relationship
risk intervention
sometime
United States
insomnia
infection
studies
area
eye
important
observed
diagnostic criteria
rate
stroke
function
make
limited
among
involve
depression
either
clinical
control
around
require
two
first
recommended
addition
symptoms include
rather
genetic
night
trauma
multiple
activity
severe
based
state
reduced
difficulty
risk factor
finding
various
primary
similar
adult
non
high
sign
causing
commonly
mutation
tumor
presence
injury
feature
research
multiple
use
difficulties
particularly
difficulties
study
lack
rather
ADHD
difficulty
risk factor

The end of the Powerpoint

Thanks for the attention. Do you have any question?