**Project Report**

# PUBMED DATA CURATION & EXPERIMENTS

**Abhay Kumar**
University of Wisconsin-Madison
Madison, WI, 53715
abhay.kumar@wisc.edu

December 17, 2020

**ABSTRACT**

This project aims at text mining of published research articles related to "Alzheimer's disease" and generation of new hypothesis/abstract from the mined text, exploiting the extracted knowledge graph and its causal relations. This report contains the details about the PubMed articles data curation and few experiments (Text Classification, BERT, Text Completion, Knowledge Graph, Named Entity Recognition(NER)) with the dataset. **CODE at** https://github.com/abhayk1201/799_Project/tree/main/code_notebook

## 1 Introduction

This project aims at text mining of published research articles related to "Alzheimer's disease" and generation of new hypothesis/abstract from the mined text, exploiting the extracted knowledge graph and its causal relations. To explore the Knowledge Graph(KG) based approach, we need to identify few entities to be represented as nodes of the KG. We identified the following entities- Here is our preliminary list (more to come): Metabolites, Neurotransmitters, Brain regions (RadLex), Neurodegenerative diseases, Other diseases, Biomarkers, Neuroimaging methods, Human Genes, Animal models, Drugs, Environmental pollutants/toxins – environmental agencies, Geographical regions/cities/countries, Hormones, Proteins, mitochondrial proteins, Proteases, Nucleic acids, carbohydrates, Lipids, Gut Microbes/ human microbiome, Chemical etc.

## 2 Data Collection

Complete data can be downloaded from-
https://drive.google.com/file/d/1A-yhN_7RghK5ji88MqoCfyBZwW9sDMxv/view?usp=sharing

I have downloaded all the pubMed articles related to *Alzheimer's Disease*

### 2.1 PubTator

The details of PubTator API is here- https://www.ncbi.nlm.nih.gov/research/pubtator/api.html
The codes for PubTator data curation and NER annotation are available at- PubTator (annoting pubmed articles)
https://academic.oup.com/nar/article/47/W1/W587/5494727
https://www.ncbi.nlm.nih.gov/research/pubtator/

sample annotated abstract- https://www.ncbi.nlm.nih.gov/research/pubtator-api/publications/export/pubtator?pmids=26739349,28483577

Sample annotated pubmed full text- https://www.ncbi.nlm.nih.gov/research/pubtator-api/publications/export/biocxml?pmcids=PMC4743391

Table 1: Entity/Attribute Souce

| Attribute Name | source |
|---|---|
| chemicals | http://ctdbase.org/downloads/ |
| disease | http://ctdbase.org/downloads/ |
| Genes | http://ctdbase.org/downloads/ |
| Proteins | http://www.hmdb.ca/downloads |
| Metabolites | http://www.hmdb.ca/downloads/ |
| pathways | http://ctdbase.org/downloads/ |
| Brain Regions | https://en.wikipedia.org/wiki/List_of_regions_in_the_human_brain |
| Mutation | https://www.ncbi.nlm.nih.gov/research/pubtator/api.html |
| species | https://www.ncbi.nlm.nih.gov/research/pubtator/api.html |

Entities relations http://ctdbase.org/about/

### 2.2 Node/Entity Classification

-Clin Term COOC : a medical term-term co-occurrence graph from (Finlayson et al., 2014)
-node2vec PPI: a PPI graph with functional annotations used in node2vec (Grover and Leskovec, 2016)
-Mashup PPI: a experimental PPI graph with functional annotations used in Mashup (Cho et al., 2016)

### 2.3 Link/Edge Prediction

-CTD DDA : a drug-disease association graph extracted from Comparative Toxicogenomics Database
-NDFRT DDA : a drug-disease association graph extracted from UMLS National Drug File
-DrugBank DDi : a drug-drug interaction graph extracted from DrugBank database
-STRING PPI : a protein-protein interaction graph extracted from STRING database

## 3 Experiments

### 3.1 Text Classification

BERT and CNN based text classification code is here- https://github.com/abhayk1201/799_Project/blob/main/code_notebook/Alz_Classification_with_BERT_%26_DNN.ipynb

### 3.2 Text Completions

We trained three GPT text completion models on three sets of pubMed articles ( (a) all papers till 2000, (b) all papers till 2010 (c) all papers till 2020) to observe if the text completion is able to give different completions based on the newer papers it's trained with. The maximum word count for text completion is set to 50. The results for 100 word length completion are in https://github.com/abhayk1201/799_Project/blob/main/code_notebook/Text_Completion_Results.txt.

Here are the comments of the domain expert about these completions (attached below for reference). Although those completions were as par, we noticed few discrepancies.

- *Alzheimer's disease should be treated with antipiracy, as it has been successfully shown without it to worsen the symptoms of Alzheimer's disease. The treatment of Alzheimer's disease is more complicated when the disease is not treated with antipiracy and when the disease is treated with antipiracy*
  —> Here, somehow "anti piracy" came into although not present in training data at all. Probably, the pre-trained model have similar sentence structure with our completed sentence structure. Q: Is there a way to differentiate between structural bias vs semantic bias. (ON THE EQUIVALENCE BETWEEN POSITIONAL NODE EMBEDDINGS AND STRUCTURAL GRAPH REPRESENTATIONS —> This paper discusses about unifying theoretical framework for node (positional) embeddings and structural graph representations)
- Role Mismatch
  *Alzheimer's disease should be treated with the apolipoprotein E (APOE) epsilon4 allele.*
  —> is not used for treatment, its a protein/gene name. This completion is factually wrong.

- Missing logic/random completions
  *Alzheimer's disease should be treated with caution, because it is believed to be a progressive disease that is likely to progress as slowly as Alzheimer's disease. Alzheimer's disease should be treated with a VPAGA monoclonal antibody.*
  —> Nothing as such as VPAGA.
  Alzheimer's disease should be treated with an agent of the "silent killer," Alzheimer's disease should be treated with an anti-Alzheimer's drug (AVA) as well as the proven anti-Alzheimer's drug (AVA-T).
  —> Nothing exists as "AVA" or "AVA-T"
- Many other completions don't make sense as per Alzheimer's literature. —> biased to pre-trained model

Some examples of GPT-2 completions on the collected dataset.

**Prompt:** ***Alzheimer's disease is caused***

PAPERS TILL 2000:

- Alzheimer's disease is caused by a genetic mutation that results in a mutation that causes the presenilin-1 (PS1) gene to encode an APP fragment which is expressed in the brains of Alzheimer's disease patients. In this study, we demonstrate that the PS1 gene ..
- Alzheimer's disease is caused by mutations of the genes encode for the beta-amyloid precursor protein (APP) gene located on chromosome 21. The mutations are also responsible for the common form of the disease. The neuropathologic features of the Alzheimer's disease (AD ..
- Alzheimer's disease is caused by the accumulation of beta-amyloid peptide in the brains of individuals with Alzheimer's disease. A cytokine response to protect against this beta-amyloid peptide accumulation may be required for the onset of the disease.
- Alzheimer's disease is caused by a genetic defect in the amyloid precursor protein (APP). APP is thought to be defective in humans because of its early activity. To date, the APP gene has been reported to have been used as a locus for early onset familial Alzheimer's ..
- Alzheimer's disease is caused by the excessive production of beta A peptide. These two peptides are "beta A" peptides present as soluble, adenoids and, in addition, beta A peptides which are insoluble, and beta A peptides which are ..

PAPERS TILL 2010:

- Alzheimer's disease is caused by a combination of genetic, environmental, and medical factors. Here, we review the current state of the knowledge regarding the etiology and pathogenesis of AD. Alzheimer's disease (AD) is a progressive disorder that is characterized by progressive and ..
- Alzheimer's disease is caused by vitamin B12 deficiency and vitamin B-12 deficiency exacerbated many neurodegenerative disorders including Alzheimer's disease. The vitamin B-12 deficiency is an integral part of the production of oxidative stress and the maintenance of the brain. Vitamin B- ..
- Alzheimer's disease is caused by abnormal processing of amyloid precursor protein (APP), a key component of the amyloid plaques found in the brain, in which the beta-amyloid (Abeta) peptide has recently been detected to be the primary pathological ..
- Alzheimer's disease is caused by a complex neurodegenerative process that includes the conversion of beta-amyloid (Abeta) to soluble Abeta (Abeta(1-40)]. The two peptide fragments undergo normal synthesis by degradation, which is the consequence of ..
- Alzheimer's disease is caused by oxidative damage to the cell surface. Vitamin E, a major component of Vitamin E supplements, also lowers the levels of total cholesterol and the levels of total amyloid-beta peptide (Abeta) (42). These findings suggest that ..

PAPERS TILL 2020:

- Alzheimer's disease is caused by an accumulation of amyloid-beta protein. A chaperone identified by the first gene (SC18B) is required for the clearance of amyloid-beta, and is now the preferred chaperone for amyloid-beta ..
- Alzheimer's disease is caused by mutations in the APP gene. We present the results of a systematic review and an analysis of the literature on the pathomechanism of Alzheimer's disease, focusing on the role of the APP gene in the pathogenesis of the disease. We ..
- Alzheimer's disease is caused by the growing burden of neurodegenerative and inflammatory diseases. This article discusses the current state of knowledge in the field of neuroinflammation and its potential therapeutic implications. BACKGROUND: The mechanisms that regulate the polymerization of amyloid beta ..
- Alzheimer's disease is caused by mutations in the amyloid precursor protein (APP) gene. The disease can be classified into two main types: sporadic, and rarer. The sporadic cases are those with a high risk of Alzheimer's disease, no known mutations and no known risk ..
- Alzheimer's disease is caused by oxidative stress and is hyper-oxidative. This is the first proposed mechanism of the disease. A recent study showed that the following oxidative stress factors are also risk factors for the development of Alzheimer's disease. It seems that oxidative stress factors are ..
- Alzheimer's disease is caused by a combination of a mutation in the gene for the gamma-secretase gene and a mutation in the gene for the presenilin 2 gene. This mutation is associated with a higher risk of developing type 2 diabetes, and yet, the risk ..
- Alzheimer's disease is caused by a misfolded protein (mAb) function, tau, and a misfolded protein (mAb) other than tau, termed amyloid beta protein (Abeta). The tau mutation of the presenilin 1 ..
- Alzheimer's disease is caused by the accumulation of beta-amyloid and the subsequent proteolytic degradation of the amyloid precursor protein. Thus, the mechanisms that promote the production of beta-amyloid are not yet fully understood. We have developed a system of
- Alzheimer's disease is caused by the sub-cortical degeneration of neurons of the central nervous system, and the sub-cortical degeneration of neurons of the brain undergoes a compensatory aging process. To date, the most studied and cleavages in the ..
- Alzheimer's disease is caused by the presence of amyloid beta (Abeta) in the brain. Since the well-known cholinergic neurotransmitter, serotonin, has an important role in the regulation of behavior, a variety of cholinergic abnormalities in Alzheimer's disease have ..

**Prompt:** ***Alzheimer's disease should be treated with***

PAPERS TILL 2000:

- Alzheimer's disease should be treated with low doses of tacrine, but at higher doses. We report a case of A beta-pleated sheet-like structures in the anterior part of the hippocampus of a patient with Alzheimer's disease, who had a severe memory disturbance and memory loss ..
- Alzheimer's disease should be treated with antipiracy, as it has been successfully shown without it to worsen the symptoms of Alzheimer's disease. The treatment of Alzheimer's disease is more complicated when the disease is not treated with antipiracy and when the disease is treated with antipiracy ..
- Alzheimer's disease should be treated with the drug ABT, which has been shown to affect cognition, memory and intellectual function. Abnormal gene expression in Alzheimer's disease has been found to be marked in the cerebellum, which is characterized by a high level of the apol ..
- Alzheimer's disease should be treated with a combination of the multisystemic approach and the multidisciplinary approach. The beta-amyloid precursor protein (beta AP) is a major component of the amyloid deposits found in the senile plaques of Alzheimer's ..
- Alzheimer's disease should be treated with a selective, non-drug therapy. New strategies, including selective and non-selective cathepsin B inhibitors, may be developed in the near future. We report the results of a study of the association between memory and learning and

PAPERS TILL 2010:

- Alzheimer's disease should be treated with cerebrovascular administration of purer acetylcholine. The clinical and biochemical characteristics of Alzheimer's disease should be improved, for various indications in Alzheimer's disease (AD) the action of acetylcholine is mediated by activation of cholin ..
- Alzheimer's disease should be treated with caution when the drug is being used for Alzheimer's disease. We examined the association between the Apolipoprotein E genotype and cerebrovascular disease in a large prospective study of elderly Chinese. A total of 400 participants participated in the study ..
- Alzheimer's disease should be treated with caution, because it is believed to be a progressive disease that is likely to progress as slowly as Alzheimer's disease. This review discusses the main problems first reported by the authors and summarizes the potential use of specific therapeutic agents for Alzheimer's disease.
- Alzheimer's disease should be treated with a combination of antipsychotic medication, tacrine, and memantine (equivalent to mg/day) and with the use of new drugs. Although attention has been focussed on Alzheimer's disease (AD) as a disease with a major ..
- Alzheimer's disease should be treated with caution when serious cognitive impairment is associated with dementia. We performed a comparison of the severity of Alzheimer disease (AD) and vascular dementia (VaD) with those of Alzheimer disease (AD) and vascular dementia (VaD). The three major neuro ..

PAPERS TILL 2020:

- Alzheimer's disease should be treated with a VPAGA monoclonal antibody. The two next steps are: (1) to determine the clinical efficacy of a monoclonal antibody to a monoclonal antibody of 4,5-dihydroxyvitamin D
- Alzheimer's disease should be treated with a combination of symptomatic anti-Alzheimer's disease drugs and a non-Alzheimer's disease drug combined with a single anti-Alzheimer's drug. BACKGROUND: In this retrospective study, we aimed to determine the prevalence
- Alzheimer's disease should be treated with an agent of the "silent killer," and the first step of the treatment should be the treatment of the disease. Over the past few years, research has explored a variety of neurodegenerative diseases. In the 1960s, the ..
- Alzheimer's disease should be treated with haloperidol with no side effects. The loss of white matter (WM) in the hippocampus is a prominent but not necessarily well-understood aspect of Alzheimer's disease (AD). We have investigated the relationship between the volume of WM in ..
- Alzheimer's disease should be treated with a beta-amyloid precursor protein inhibitor. Tyracyclines are the most abundant intracellular protein of the serine/threonine chain in the mammalian brain. The only known compound that can inhibit the expression of these neurot ..
- Alzheimer's disease should be treated with a combination of beta-amyloid and pioglitazone. Alzheimer's disease is the most widely inherited disease in the United States, and there are many causes of the disease. The most widely known treatment for the disease is ..
- Alzheimer's disease should be treated with the apolipoprotein E (APOE) epsilon4 allele. There is a strong genetic risk for Alzheimer's disease. The APOE epsilon4 allele is a risk factor for AD. The APOE genotype can be ..
- Alzheimer's disease should be treated with an anti-Alzheimer's drug (AVA) as well as the proven anti-Alzheimer's drug (AVA-T). Protease cleaving enzyme 2 (CREB2) is a key regulator of the ..
- Alzheimer's disease should be treated with an anti-diabetic drug. The role of amino acid substitution in the pathogenesis of Alzheimer's disease is not well defined, but the amino acid sequence of the amyloid precursor protein (APP) was identified in the brains of two ..
- Alzheimer's disease should be treated with a combination of acetylcholinesterase inhibitors and cholinesterase inhibitors. The purpose of this study was to identify the major molecular events associated with the accumulation of amyloid-beta (Abeta) in the brain ..

**Prompt:** ***Alzheimer's disease is characterized by***

PAPERS TILL 2000:

- Alzheimer's disease is characterized by progressive dementia of the Alzheimer type, and the occurrence of a cortical subcortical Lewy body disease. The ventricular size of the cerebral cortex is reduced in Alzheimer's disease, and the cerebellar density is reduced by an inordinate magnitude ..
- Alzheimer's disease is characterized by the presence of multiple neurodegenerative processes, the presence of which are accentuated in the AD brain. In this study, we have compared the levels of cAMP and Ginkgo biloba extract, the enzyme that cleaves phosphat
- Alzheimer's disease is characterized by a progressive neurodegenerative process, including progressive as well as severe neurodegenerative changes. A progressive memory deficit is also present in many demented patients. Using the NINCDS-ADRDA criteria, the present case report provides the
- Alzheimer's disease is characterized by neuronal degeneration and synaptic loss. In addition, the gene encoding APP is expressed by up to 40% of all neurons in the nervous system. For example, the expression of the APP gene is increased in the hippocampal region of the human brain.
- Alzheimer's disease is characterized by the presence of senile plaques, and Alzheimer's disease is associated with formic acid plaques. The present study explored the role of an electrical current in the neurodegeneration of Alzheimer's disease (AD) patients. The current was ..

PAPERS TILL 2010:

- Alzheimer's disease is characterized by a progressive loss of cognitive functions. The molecular pathophysome of Alzheimer's disease is based in the formation of amyloid-beta (Abeta) and promotes an inflammatory response. The growing evidence for an etiopathogenesis of Alzheimer's disease ..
- Alzheimer's disease is characterized by the presence of amyloid plaques that are the hallmark of Alzheimer's disease (AD). Amyloid fibrils (A beta, beta-amyloid, and tau) can exhibit concomitant pathological changes, such as ext ..
- Alzheimer's disease is characterized by severe pathological changes in the brain. The aim of this study was to mimic the effects of the amyloid-beta-peptides and the tau-immunoreactive fragment of the presenilin 1 (PS-1)
- Alzheimer's disease is characterized by a progressive loss of synapses and the loss of neurons. This loss provides a new target for the development of new drugs that target synapse loss in Alzheimer's disease, which has been the target of intense interest. In this study, we tested ..
- Alzheimer's disease is characterized by changes in the brain, particularly of the central nervous system, as well as in the blood. A significant amount of the disease is caused by mutations in the amyloid precursor protein, the most important protein found in Alzheimer's disease.

PAPERS TILL 2020:

- Alzheimer's disease is characterized by an increased risk of dementia. We propose that apoE genotype mediate the genotype-phenotype interaction, and that we are able to identify the allele-phenotype interaction in AD. The results indicate that the APOE4/
- Alzheimer's disease is characterized by progressive hyperphosphorylation of tau, and the accumulation of phosphorylated tau is associated with neurodegeneration. In the present study, we measured the expression of the tau and phosphorylated tau by immunohist ..
- Alzheimer's disease is characterized by the aggregation of the amyloid precursor protein (APP) and the metabolic alterations leading to the formation of the amyloid beta-protein (A beta). Both soluble and insoluble APP and A beta are required for the initiation of A beta
- Alzheimer's disease is characterized by severe cognitive impairment, a progressive neurodegeneration, and the development of senile plaques and neurofibrillary tangles. A genetic study was undertaken to characterize the protein kinase A-II (apoA-II), a key enzyme
- Alzheimer's disease is characterized by progressive neurodegeneration linked to neurodegenerative changes. This review summarizes the currently available evidence on the role of amyloid-beta (Abeta) as a potential therapeutic target for Alzheimer's disease. We also discuss the recent advances in
- Alzheimer's disease is characterized by a progressive decline in cognitive function and a poor prognosis which has been attributed to the hyperphosphorylation of the amyloid precursor protein in the brain. In the present study, we assessed the expression of the APOE4 gene in

- Alzheimer's disease is characterized by the presence of amyloid-beta (Abeta) and transmembrane-rich amyloid plaques. Although this is the first case of a neurodegenerative disease, the presence of Abeta in the brains of patients with
- Alzheimer's disease is characterized by the deposition of amyloid-beta (Abeta) peptides, which accumulate in the brains of Alzheimer's disease patients. In this study, the expression of the human Abeta peptide receptor (4-hydroxynonenal (4
- Alzheimer's disease is characterized by the presence of abnormal amyloid-beta (Abeta) deposits and by the accumulation of extracellular plaques. The predominant pathological features in AD are amyloid plaques and neurofibrillary tangles (NFT) and
- Alzheimer's disease is characterized by progressive neurodegeneration, and the amyloid plaques are the major pathological hallmarks of the disease. Recent studies have shown that the underlying role of amyloid precursor protein (APP) is involved in the pathological process of AD. However

## 4 Code

**CODE at** `https://github.com/abhayk1201/799_Project/tree/main/code_notebook`
[1]

## References

[1] `https://github.com/abhayk1201/799_Project/tree/main/code_notebook`