

○ ○ ○

```
def create_T_dict_list(folder): # helped with collecting frequencies
    T_dict_list = []

    for file in ann:
        T_dict = {}
        with open(f"{folder}{file}") as file_encoded:
            for line in file_encoded.readlines():
                if (line.startswith("T")) and (any(char.isdigit() for char in line)) == True:
                    T_line = line.split("\t")
                    T_dict.update({T_line[0]:T_line[1]})
            T_dict_list.append(T_dict)

    return T_dict_list

#-----#

def find_drugs(folder): # helped with identifying misannotated files
    drugs_files = []

    for file in ann:
        with open(f"{folder}{file}") as file_encoded:
            for line in file_encoded.readlines():
                if "Drugs" in line:
                    drugs_files.append(file)

    drugs_files = sorted(list(set(drugs_files)))
    return drugs_files

#-----#

def find_multis(folder): # helped with identifying misannotated files
    multis_files = []

    for file in ann:
        smoker_counter = 0; drinker_counter = 0; duser_counter = 0
        with open(f"{folder}{file}") as file_encoded:
            for line in file_encoded.readlines():
                smoker_counter += line.count("Smoker")
                drinker_counter += line.count("Drinker")
                duser_counter += line.count("DUser")
                if smoker_counter > 1:
                    multis_files.append(file)
                if drinker_counter > 1:
                    multis_files.append(file)
                if duser_counter > 1:
                    multis_files.append(file)

    multis_files = sorted(list(set(multis_files)))
    return multis_files
```

○ ○ ○

```
def create_label_dict(folder): # helped with labeling files
    # ----- #

    smoker_dict = {}; drinker_dict = {}; duser_dict = {}

    for file in ann:
        with open(f"{folder}{file}") as file_encoded:
            for line in file_encoded.readlines():
                if "Smoker" in line:
                    smoker_ant = [string for string in line.split(" ") if "Smoker" in string]
                    smoker_ant = "".join([str(elem) for elem in smoker_ant])
                    smoker_ant = smoker_ant.split("\t")[-1]
                    smoker_dict.update({file:smoker_ant})
                if "Drinker" in line:
                    drinker_ant = [string for string in line.split(" ") if "Drinker" in string]
                    drinker_ant = "".join([str(elem) for elem in drinker_ant])
                    drinker_ant = drinker_ant.split("\t")[-1]
                    drinker_dict.update({file:drinker_ant})
                if "DUser" in line:
                    duser_ant = [string for string in line.split(" ") if "DUser" in string]
                    duser_ant = "".join([str(elem) for elem in duser_ant])
                    duser_ant = duser_ant.split("\t")[-1]
                    duser_dict.update({file:duser_ant})

    print(len(smoker_dict), len(drinker_dict), len(duser_dict))

    # ----- #

    files_dict = {}
    for file in ann:
        files_dict.update({file:""})

    dd = defaultdict(list)

    for d in (files_dict, smoker_dict, drinker_dict, duser_dict):
        for key, value in d.items():
            dd[key].append(value)

    for i in dd:
        dd[i].pop(0)
        dd[i] = " | ".join([str(elem) for elem in dd[i]])

    substance_ants = ["Smoker", "Drinker", "DUser"]

    for key in dd.keys():
        for ant in substance_ants:
            if (ant not in dd[key]):
                dd[key] = dd[key] + f" | Unknown{ant}"

    # ----- #

    return dd
```