Run this notebook to check that your annotated data is in the proper format. Before running it, there are two things you need to do:

1. Change these files to point to your data

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=Tr
```

```
adjudicated_path="/content/drive/MyDrive/Annotation project/AP2/adjudicated.txt"
individual_annotation_path="/content/drive/MyDrive/Annotation project/AP2/individual_annotations.txt"
```

2. In the set below, enumerate the valid categories that are described in your guidelines. (This helps check that there aren't typos in your labels.)

```
valid_categories=set(["poor", "average", "good", 'excellent'])
```

3. Now execute the rest of the cells below. If this throws any errors, or notes any failures, go back and correct your data to be in the proper format.

```
from collections import Counter
import numpy as np

def check_file(filename, min_count):
    annotator_triples={}
    annos_by_data_id={}
    with open(filename, encoding="utf-8") as file:
        for idx, line in enumerate(file):
            cols=line.rstrip().split("\t")
            assert len(cols) == 4, "%s does not have 4 columns" % cols
            assert len(cols[3]) > 0, "text #%s# in row %s is empty" % (cols[3], idx)
            assert len(cols[2]) > 0, "label #%s# in row %s is empty" % (cols[2], idx)
            annotator_triples[cols[1], cols[0], cols[2]]=1
            annos_by_data_id[cols[0]]=1

            label=cols[2]

            if label not in valid_categories:
                print("\"%s\" is not a valid category" % label)
                print("Check failed.")
                return

        assert len(annos_by_data_id) >= min_count, "You must have at least %s labels; this file only has %s" % (min_count, count)

        print("This file looks to be in the correct format; %s data points" % len(annos_by_data_id))
    return list(annotator_triples.keys())

adjudicated=check_file(adjudicated_path, 500)
```

```
This file looks to be in the correct format; 500 data points
```

```
def check_individual_file(filename):
    annotator_triples={}
    annos_by_data_id={}
    annos_by_annotator={}
    labels={}
    with open(filename, encoding="utf-8") as file:
        count=0
        for idx, line in enumerate(file):
            cols=line.rstrip().split("\t")
            data_id=cols[0]
            anno_id=cols[1]
            label=cols[2]

            if label not in valid_categories:
                print("\"%s\" is not a valid category" % label)
                print("Check failed.")
                return

            assert len(cols) == 4, "%s does not have 4 columns" % cols
            assert len(cols[3]) > 0, "text #%s# in row %s is empty" % (cols[3], idx)
            assert len(label) > 0, "label #%s# in row %s is empty" % (cols[2], idx)
            count+=1

            annotator_triples[anno_id, data_id, label]=1

            if data_id not in annos_by_data_id:
                annos_by_data_id[data_id]={}
            annos_by_data_id[data_id][anno_id]=1

            if anno_id not in annos_by_annotator:
                annos_by_annotator[anno_id]={}
            annos_by_annotator[anno_id][data_id]=1

            if label not in labels:
                labels[label]=0
            labels[label]+=1


    assert len(annos_by_data_id) >= 0, "You must have labels for at least 500 documents; this file only has %s" % (len(annos_by_

    for data_id in annos_by_data_id:
        assert len(annos_by_data_id[data_id]) == 2, "Each data point must have two annotations; data id %s does not" % data_id

    print("Annotators:\n")
    for anno_id in annos_by_annotator:
        print("%s: %s" % (anno_id, len(annos_by_annotator[anno_id])))

    print("\nLabels:\n")
    for label in labels:
        print("%s: %s" % (label, labels[label]))


    if len(annos_by_data_id) < 250:
        print("\nThis file needs to contain annotations for at least 250 data points; this only contains %s." % len(annos_by_dat
        return

    print("\nThis file looks to be in the correct format; %s data points; %s annotations" % (len(annos_by_data_id), len(annotatc
    return list(annotator_triples.keys())


annotation_triples=check_individual_file(individual_annotation_path)
```

```
Annotators:

Lan: 251
Khoa: 251

Labels:

poor: 39
average: 233
excellent: 43
good: 187

This file looks to be in the correct format; 251 data points; 502 annotations
```

Execute the following cell to calculate Fleiss' kappa on your individual annotations.

```
def fleiss(annotation_triples):
    cats={}
    items={}
    uid_counts=Counter()
    uid_id={}
    aid_counts=Counter()

    # get label categories and unique data points
    for aid, uid, label in annotation_triples:
         if label not in cats:
            cats[label]=len(cats)
            if uid not in uid_id:
                uid_id[uid]=len(uid_id)

            uid_counts[uid]+=1

    ncats=len(cats)
    ps=np.zeros(ncats)

    data = []

    for aid, uid, label in annotation_triples:

        if uid not in items:
            items[uid]=np.zeros(ncats)

        items[uid][cats[label]]+=1
        ps[cats[label]]+=1

    ps/=np.sum(ps)

    expected=0.
    for i in range(ncats):
        expected+=ps[i]*ps[i]

    agreements=[]
    for item in items:
        total=np.sum(items[item])
        assert total >= 2, "every data point must have at least two annotations; this one has %s" % (total)
        summ=0

        for i in range(ncats):
            summ+=items[item][i]*(items[item][i]-1)
        summ/=(total*(total-1))

        agreements.append(summ)

    observed=np.mean(agreements)
    print ("Observed: %.3f" % (observed))
    print ("Expected: %.3f" % (expected))
    print ("Fleiss' kappa: %.3f" % ((observed-expected)/(1-expected)))



fleiss(annotation_triples)
```

```
Observed: 0.713
Expected: 0.368
Fleiss' kappa: 0.546
```