NLP Assignment

You are given a JSON file (tweets.json) that contains tweets (sentences) along with the name of the author.

Objective 1: Get the most frequent entities from the tweets.
Objective 2: Find out the sentiment/polarity of each author towards each of the entities.

Sample Input:
Assume we have only 4 tweets:

Tweet1 by Author1: Pink Pearl Apples are tasty but Empire Apples are not.
Tweet2 by Author2: Empire Apples are very tasty.
Tweet3 by Author3: Pink Pearl Apples are not tasty.
Tweet4 by Author1: Pink Pearl Apples smells really good.

Sample output:
Entities in the topics extracted: Share a CSV with extracted entities and the frequency of the extracted entity from all the tweets in the following format

objective1.csv

entity	frequency
Pink Pearl Apples	2
Empire Apples	2

Sentiment/polarity of Authors: Share a CSV file with predicted sentiment values with extracted entities as columns and unique authors as rows. See the example CSV below.

objective2.csv

entity	author_name	overall_polarity
Pink Pearl Apples	Author1	Positive
Empire Apples	Author1	Negative
Empire Apples	Author2	Positive
Pink Pearl Apples	Author3	Negative

Downloading and reading the JSON file:

Get the JSON file here: http://bit.ly/akaiketech_cll_tweets_json

Python code for reading the JSON file:

```
import json
with open('tweets.json') as jfile:
    d = json.load(jfile)
```

d would be a dictionary with tweet_id as key and another dictionary as a value. The inner dictionary contains the information tweet_text and tweet_author. See the sample below.

{"1374140386071961602":

- **{tweet_author:**"Hematopoiesis News"
- **tweet_text:**"⚕ Scientists conducted a Phase II study of acalabrutinib in patients with relapsed/refractory #CLL who were ibrutinib-intolerant, and found an overall response rate of 73%. https://t.co/eJ6m4QpC5P https://t.co/kuZz6ZO47r"}

...
}

Instructions:
Make sure to discuss the following aspects in a text document:

- Document your approach to solve the problem, discussing the difficulties and how your proposed solution tackles them.
- Discuss the technique used and the reason why you have chosen it.
- Discuss the shortcomings or mistakes of your proposed solution with a few examples.
- If there are any shortcomings or mistakes, discuss how you would go about tackling them given more resources and time.

Keep this in mind while implementing your solution:

- The programming language used should be Python.
- You can use any NLP technique, library or code. For some ideas, explore notable NLP toolkits like NLTK, Spacy, StanfordNLP, AllenNLP, HuggingFace Transformers, and so on. (Note that the entire solution to this assignment won't be directly available anywhere).
- Open source implementations of some research papers related to text mining and different types of sentiment analysis could be of help too.

● Treat this as an open ended problem and solve as much as you could. ●
Any open source code used should be credited to the author or the source.

Submission (share below files in a zipped folder):

1. Share a text document discussing the above-mentioned items.
2. Share all the code or notebooks that computed the output. (See sample output section for required output).
3. Use the following file names for sharing the final results and upload to the appropriate slot in the google forms

   Objective 1 - objective1.csv

   Objective 2 - objective2.csv

   Use the column names and format exactly as described in the sample output section
4. Zip your code and documentation with the following naming convection and upload to the appropriate slot in the google forms

   “Yourname_nlp_assignment.zip”

Note:-

A. Submissions are expected to be in the order of hundreds, please comply with guidelines.
B. Not following the guidelines will be subjected to automatic disqualification.

WE WISH YOU GOOD LUCK