DOCUMENTATION:

1. I used two different approaches for finding names: NLTK and Spacy.
2. On cleaning the document, I fed it into NLTK Chunker to extract entities.
3. Separated out PERSON and ORGANIZATIONS entities. On separation, I removed the repetitive names from both the entities.
4. I repeated the same process with Spacy model.
5. I cross-checked the names of the persons recognized by both the models, by comparing them. It is expected that the common names recognized by both of them are actual names. The first name is considered to be the name of the EMPLOYEE as the letter is written to him/her and the that name generally appears first. The last name is considered to be the employer name (the one who signed), as that name is expected to end the letter. In this way we get both the EMPLOYEE and EMPLOYER names.
6. Since the set of documents are mostly the agreement letters for the post of EXECUTIVES of different employees, so some common executive posts are listed in python. This list is cross-checked with every sentence of the document to obtained the mentioned role of the employee. It is seen to have good accuracy.
7. Salary – Since the base salary starts with $ sign, I looked for the greatest $ value as this may denote the yearly base salary of the employee. I didn’t use MONEY entity of SPACY, as the model recognized even some words with the money.
8. Finally, I checked if the document mentions of AMENDMENT. If so, The document type is AMENDMENT LETTER else EMPLOYEMENT AGREEMENT.