NAME: SEPIDEH SHOSSEINZADEH

Report

- **Problem Statement**:
  Wikidata - the Sherlock Holmes investigative AI
  Given personal information about a person determine the probable cause of death.
  Personal information is provided in wikidata format, for example: Abraham Lincoln. Given all other information and entity links you need to build a classifier that would determine the most probable manner of death (you should not use 'manner of death' or 'cause of death' fields as your input data).
  There are many such manners defined in wikidata, but we are interested in the 4 most common ones: natural causes, suicide, accident, homicide.
- **Input**:
  I extracted the person names who their manner of death were each of the 4 common manners by wikipedia query LINK
  Then, I used the wikipedia links provided in their wikidata, and input the document words as the input data. The documents' label will be its associated manner of death.
- **Output**:
  Most probable manner of death.
- **Methodology**:
  First, we construct a list of documents, labeled with the appropriate categories.
  Next, we define a feature extractor for documents, so the classifier will know which aspects of the data it should pay attention to.
  For document topic identification, we can define a feature for each word, indicating whether the document contains that word. To limit the number of features that the classifier needs to process, we begin by constructing a list of the 2000 most frequent words in the overall corpus. We can then define a feature extractor that simply checks whether each of these words is present in a given document.
  At last, Naive Bayes Classifier is trained and tested.
- **Assumptions and observations**:
  We should use more data to increase the accuracy. Relevant data like cause of death should be used to increase the accuracy, but in the problem statement it mentioned not to use it.
- **Codes**:
  - nlp.py is the main code that should be run.
  - pywikibot folder is a python package for working with wikidata database.
- **Note**:
  There is a warning because of using python 2.7.
- **Guide for using the codes**:
  - Dependencies to be installed: urllib2, pandas, nltk, pywikibot (installation page link)
  - Line 5 of the code should be assigned to pywikibot folder in your home directory. Variable *path* in line 16 should be the data path .csv files.
  - Please copy the 3 codes (in codes folder) and data folder into pywikibot, then run the nlp.py.