NLP - Description for Students

November 7, 2024

1 Natural Language Processing

This project will give you practical experience using Natural Language Processing techniques. This project is in three parts: - in part 1) you will use a traditional dataset in a CSV file - in part 2) you will use the Wikipedia API to directly access content on Wikipedia. - in part 3) you will make your notebook interactive

1.0.1 Part 1)

- The CSV file is available at https://ddc-datascience.s3.amazonaws.com/Projects/Project.5-NLP/Data/NLP.csv
- The file contains a list of famous people and a brief overview.
- The goal of part 1) is provide the capability to
  - Take one person from the list as input and output the 10 other people who's overview are "closest" to the person in a Natural Language Processing sense
  - Also output the sentiment of the overview of the person

1.0.2 Part 2)

- For the same person from step 1), use the Wikipedia API to access the whole content of that person's Wikipedia page.
- The goal of part 2) is to produce the capability to:
  1. For that Wikipedia page determine the sentiment of the entire page
  2. Print out the Wikipedia article
  3. Collect the Wikipedia pages from the 10 nearest neighbors in Step 1)
  4. Determine the nearness ranking of these 10 to your main subject based on their entire Wikipedia page
  5. Compare the nearest ranking from Step 1) with the Wikipedia page nearness ranking

1.0.3 Part 3)

Make an interactive notebook.

In addition to presenting the project slides, at the end of the presentation each student will demonstrate their code using a famous person suggested by the other students that exists in the DBpedia set.

[3]:
```
!curl -s https://ddc-datascience.s3.amazonaws.com/Projects/Project.5-NLP/Data/
↪NLP.csv | wc -l
```

42786

[ ]: