ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837
ATHENA Research & Innovation Information Technologies
ΙΕΛ LSP

M901
Προγραμματισμός για Γλωσσική Τεχνολογία I

Final Project

Κυλάφη Χριστίνα-Θεανώ

LT1200012

January, 2021

Overview

The purpose of this readme file is to briefly present some notes on the Final Project.

Submission files

In the submitted .zip file, the following are also included:

1. final_project.py
2. final_project_test.py
3. corpus_stats.txt
4. total_frequencies.txt

Implementation

Text Cleaning

The **text-cleaning** method that was used, was based on **regular expressions**. First, some basic **substitutions** were made in order to optimise the process and then the text was **split** into tokens. The recognised **pattern** contains only **alphanumeric** characters, the **underscore** and the **dash**.

Alternative approaches could have been followed to clean the text, depending on the task at hand, keeping in mind that there is no generic and optimal manner to do so. For example, **emails** could have been kept, **special case handling** could have been performed to take into account specific **formats**, etc. However, in order for the aforementioned procedure to be successful, prior **knowledge** of the **dataset** is required.

The method that was finally kept and used, was affected by the **observation** of the **output** each time and the **comparisons** made between the resulted **data**.

Output

Corpus Statistics & Dictionary

The final **dictionary** of the whole given **corpus**, the respective **document dictionary sizes** and some **information** on the **corpus**, are written to files “corpus_stats.txt” and “total_frequencies.txt”, included in the submitted .zip file as mentioned above.

In total, the corpus was split into **95421 tokens**. The **document** with the largest number of tokens (**13916**) was “**35557.pdf.txt**”. Finally, the 3 **lowest frequencies 1**, **2** and **3**, counted **44432**, **14032** and **7235** tokens respectively.