Master BeNeFri in Computer Science

Course: Digital Humanities
Fall 2022

Exercise #7: Vector space model (part two)

Instructions

Download from the ILIAS website the French Theater data (path: “/french-theater/”). This folder contains french theater plays (stored in XML format).
This exercise series consists of 6 practical questions. Upload your answers and the source code used for your computations to the ILIAS website. You can submit either a .pdf file or comment the source .py file. IPython/Jupyter notebook files (.ipynb) are allowed as well.

Practical Questions

1) For each genre, generate a “profile” in the form of a single vector representing the entire set of plays corresponding to this genre. Build such a profile for each of the three genres (Comedy, Tragedy and Tragicomedy).

2) Which are the three plays for each text genre (or group) that are the “closest” to the profile?

3) Usually, we generate a profile by averaging over all term frequencies of plays belonging to a certain group. Do you know another way to generate a profile from a set of documents (or vectors)?

4) Do you think that the profile must include all words appearing at least once in a play of the group? If no, how can we select a subset of the terms that must appear in a profile? Justify your choice.