Master BeNeFri in Computer Science

Course: Digital Humanities
Fall 2022

Exercise #6: Vector space model (part one)

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) Represent each play by a vector with only the *tf* component. You can apply some preprocessing before generating this vector representation.

2) For each genre, it is possible to 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).

3) How many terms with a weight strictly larger than 0 do you have in each text genre profile?

4) Select randomly 10 plays for each text genre. Represent each play by a vector.

5) For each text genre and play, how many terms with a weight strictly larger than 0 do you have in the vector?

6) For each text genre and play, how many terms with a weight strictly equal to 1 do you have in the vector?