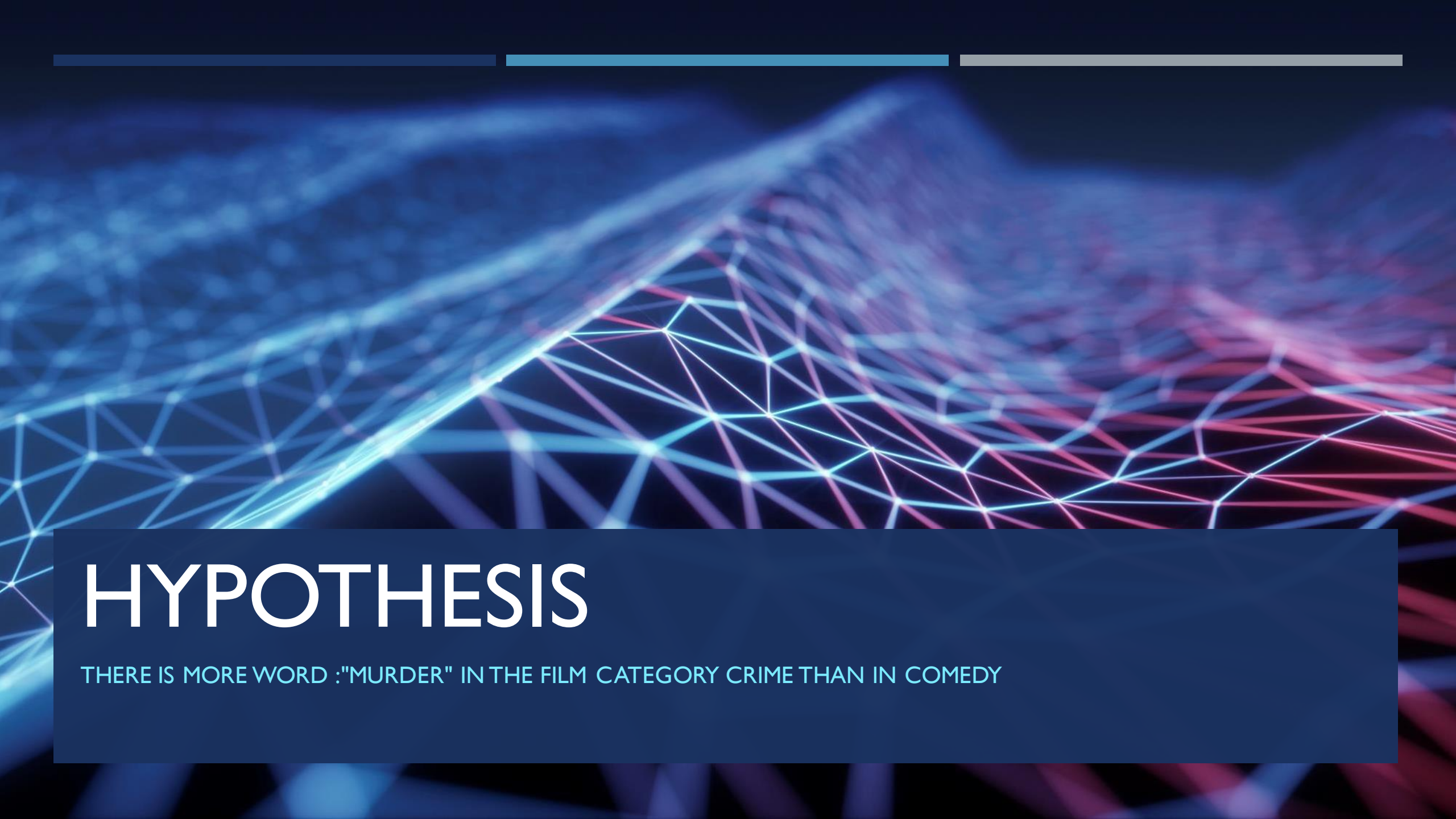
ASSIGNEMENT 1



HYPOTHESIS

THERE IS MORE WORD :"MURDER" IN THE FILM CATEGORY CRIME THAN IN COMEDY

- I took about 100 scripts of each category and indexed all documents with an ID.
- Then, I took all words of each scripts and clean the stopwords (link words, lowercase, regex, determinant).
- I also cleaned the punctuation by using a regex for example: "and about the crime...". The string "crime..." is considered as a word so I used the regex on it and that became "crime"
- I did not take the numbers in the clean list of words.
- Moreover, I have counted all the "murder" contained in each category

TECH REQUIEMENTS

EXECUTION

- As you can see, there is more word murder in Crime category than in Comedy category.
- I also did the pourcentage of the word in each category so:
- ~ 0.025% in Crime
- ~ 0.007% in Comedy