# AMI23B – Business Intelligence Lab4

## Text Mining and NLP: Star Wars Movie Scripts

This task is based on a Kaggle competition that was launched a few years back as a tribute to the Star Wars day on the 4th of May.

In this task you will demonstrate your text mining and linguistic skills to deduce insights about the Star Wars movie scripts.

You are provided with a collection of script dialogue between characters for the first three movies (episodes 4-6) also known as *The Original Trilogy*. You are also provided with Word Cloud Masks for you to make use of.

Check the list of libraries and tutorial articles at the end of this document.

## Your tasks:

*"Do. Or do not. There is no try."* — Yoda

1. Find the characters with most dialogue in each of The Original Trilogy (episodes IV, V, VI).
2. Plot the number of dialogues according to character for each of the episodes (i.e. plot the above findings).
3. Add a new column "episode" to the three datasets (to distinguish between the three episodes) and concatenate them into one dataset.
4. Discover the Frequency Distribution of words in The Original Trilogy.
5. Create a Frequency Distribution plot of the most repeated words in The Original Trilogy.
6. Perform text-mining operations to prepare your dataset for further text-analysis. (Use the NLTK library)
   a. Convert to lower case, word tokenization, removing stopwords, lexicon normalization (lematization)…etc.
   b. Add the resulting array list to the dataset as a new column "new_script".
7. Repeat steps 4 & 5 but this time check the frequency distribution of the "new_script".
8. Use Word Clouds to visually represent the most repeated words for Darth Vader and Yoda. (use the provided wordcloud masks, make one word cloud for each character)
9. Discover the **most repeated** words and the **most relevant** words in The Original Trilogy script.
   Remember: Bag of Words just creates a set of vectors containing the

count of word occurrences in the document, while the TF-IDF model contains information on the more important words and the less important ones as well.

10. Perform a sentiment analysis on the movie script.
    In python, sentiment analysis libraries for quick solutions are rare while in R, there are many such libraries. Check out https://pypi.org/project/sentic/ which provides a solution compatible with Python (or any other library you prefer). In Python you will find that the most common way to perform sentiment analysis is done by means of a Naïve Bayes Classifier, where you build the model (but you are not necessarily required to build one, you choose the way you want to do it).

In the Star Wars universe, the Sith (like Darth Vader or Emperor Palpatine) are associated with negative feelings such as anger, fear, hate, etc. Conversely, the Jedi (like Luke Skywalker or Yoda) teach its followers to not give in to feelings of anger toward other lifeforms, which would help them resist fear and prevent them from falling to the Dark Side of the Force. According to your sentiment analysis done previously, do you notice differences between the Dark Side characters and the Light Side characters? Explain your insights!

## Submission Instructions:

To pass this lab you need to hand in a short report in which you present and motivate with the help of text and diagrams, the insights you gathered from The Original Trilogy movie scripts (the report does not need to include code). The hand-in must also contain the code file. Submit your solutions no later than Thursday the 4th of June at 23:59.

“The world is one big data problem.” ~ Andrew McAfee

Libraries and Tutorial Articles: ( + all libraries from previous labs)

- NLTK 3.5 documentation
  https://www.nltk.org/
- Text Analytics for Beginners using NLTK
  https://www.datacamp.com/community/tutorials/text-analytics-beginners-nltk
- Text Mining in Python: Steps and Examples
  https://medium.com/towards-artificial-intelligence/text-mining-in-python-steps-and-examples-78b3f8fd913b
- Word Cloud for Python documentation
  https://amueller.github.io/word_cloud/index.html
- Masked Wordcloud
  https://amueller.github.io/word_cloud/auto_examples/masked.html
- Image Module
  https://pillow.readthedocs.io/en/stable/reference/Image.html
- Getting Started with Chart Studio in Python
  https://plotly.com/python/getting-started-with-chart-studio/
- Bar Charts in Python with Plotly (docs)
  https://plotly.com/python/bar-charts/
- re — Regular expression operations (docs)
  https://docs.python.org/3/library/re.html
- Regular Expressions in Python
  https://www.pythonforbeginners.com/regex/regular-expressions-in-python
- Removing Stop Words from Strings in Python
  https://stackabuse.com/removing-stop-words-from-strings-in-python/
- How to Use Tfidftransformer & Tfidfvectorizer
  https://kavita-ganesan.com/tfidftransformer-tfidfvectorizer-usage-differences/#.XsLFCqgzY2w
- Simplifying Sentiment Analysis in Python
  https://www.datacamp.com/community/tutorials/simplifying-sentiment-analysis-python
- How To Perform Sentiment Analysis in Python 3 Using the Natural Language Toolkit (NLTK)
  https://www.digitalocean.com/community/tutorials/how-to-perform-sentiment-analysis-in-python-3-using-the-natural-language-toolkit-nltk
- Python Interface for Semantic and Sentiment Analysis using Senticnet4 (http://sentic.net/)
  https://pypi.org/project/sentic/