AMI23B – Business Intelligence Lab 4

Objectives:

- To understand the basics of natural language processing (NLP).
- Analyse the frequency distribution of words in The Original Trilogy.
- Create a frequency distribution plot of the most used words in The Original Trilogy.
- Perform text-mining operations to pre-process the dataset for further analysis using NLTK.
- Reanalyse the frequency distribution of words after pre-processing.
- Create Word Clouds to represent the most used words for Darth Vader and Yoda.
- Identify the most relevant words in The Original Trilogy script using the TF-IDF model.
- Perform sentiment analysis on the movie scripts to determine the overall sentiment of the trilogy.

Text Mining and NLP: Star Wars Movies Scripts

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

Here, you will demonstrate your text mining and linguistic skills to deduce insights about the Star Wars movies scripts.

You are provided with a collection of script dialogues between characters for the first three movies (episodes 4-6), also known as *The Original Trilogy*. You are also provided with some word cloud masks for you to use.

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

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

Your tasks:

1. Find the characters with the most dialogues in each episode of The Original Trilogy (Episodes IV, V, VI).
2. Plot the number of dialogues according to the character for each episode (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 (lemmatization), etc.
    b. Add the resulting array list to the dataset as a new column, "new_script".
7. Repeat steps 4 & 5 but check the frequency distribution of the "new_script" this time.
8. Use Word Clouds to visually represent the most repeated words for Darth Vader and Yoda. (Use the provided word cloud masks, make a single word cloud for each character.)
9. Discover the most relevant words in The Original Trilogy script. The TF-IDF model contains information on the more important and less important words (relevance).
10. Perform sentiment analysis on the movie scripts.

In Python, you will find that the most common way to perform sentiment analysis is employing a Naïve Bayes Classifier, where you build the model (but you are not necessarily required to build one here). You could make use of libraries to perform your sentiment analysis. Check out "Sentic" or the quite ironically named library, "VADER" (or any other library you prefer).

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. Do you notice differences between the Dark Side characters and the Light Side characters according to your previous sentiment analysis? Explain your insights!

Submission:

A python file **Yourusername_Lab4.ipynb**.

"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/
- Sentiment Analysis Made Easy Using VADER
  https://analyticsindiamag.com/sentiment-analysis-made-easy-using-vader/