Project plan

Project name: Moodify

Project members:
András Borbély, Olena Ushenko, Kateryna Manoilova, Darius Kosk

Problem statement: Initially, our project is for learning purposes. Valence is usually calculated mainly via other methods, such as the song's loudness, tempo, timbre. It is unclear how reliable it would be to calculate it from text processing. We will predict the emotional value out of lyrics of the Spotify dataset.

Objectives:

1. Predict continuous valence scores (0–1) from lyrics using regression models.
2. Categorize songs into three emotion classes — Sad, Neutral, Happy — based on valence thresholds.
3. We would be satisfied with a RMSE of under 0.3, 65%+ accuracy.
4. Visualize artists grouped by average valence
5. Calculating the valence of each sentence in lyrics -> Find if there are sentences that substantially increase/decrease the song's valence
6. Calculating the valence of a song based on its title -> Find out if the title has any correlation with the song's valence
7. Calculating the valence of a song based on its lyrics -> Predicting a song's valence based on only the lyrics.

Data: The data is from Kaggle. The data consists of 135 991 songs
Link: https://www.kaggle.com/datasets/edenbd/150k-lyrics-labeled-with-spotify-valence
This dataframe has four columns:

- artist = artist name. (words)
- seq = song's lyrics. (words)
- song = song title. (words)
- label = Spotify valence feature attribute for this song. (float)

Steps:

1. Split words into tokens(splitting the lyrics to separate words)
2. Cleaning data -> Remove words that shouldn't affect valence, such as prepositions, stop words etc. Stem and lemmatize the words.
3. Split the data into training and testing sets.
4. Apply PCA to reduce features. (???)
5. Using the Cross Validation algorithm, find the most appropriate model.
6. Evaluate the final model on the test set.
7. Visualization

Methodology:
Text preprocessing: tokenization, stopword removal, lemmatization.

Modeling: Linear Regression, Random Forest, Logistic Regression (for classification), LSTM or BERT fine-tuning if time allows.

Evaluation: Compare calculated valence scores for songs to the test data labels. Since the valence score is a float, we cannot expect matching numbers, thus we will evaluate the accuracy with MSE.

Expected challenges: Both high and low valence songs can have the same words, the valence depends on many different aspects of a song, including the context of the words. Songs can contain ambiguous language (sarcasm, metaphors).

Resources and tools:
Meeting schedule: https://www.when2meet.com/?32785818-Mdzfy
Google Docs
Google Collaboratory
Libraries:

- Scikit learn
- Numpy
- Pandas
- Matplotlib
- Natural Language Toolkit (ntlk)

Questions for further guidance:
Can we use Vader or do we need to implement the model from scratch?
Are there other limitations on our implementations?
Can we adjust objectives of the project later on?

Milestones and timeline:

WEEK	**DATE**	**MILESTONES OR PLANNED ACTIVITIES**
Week 3	September 28	Forming teams, communication, project idea
Week 4	October 5	Finding relevant data. Writing the project plan
Week 5	To be chosen. Online/Delta	Summarize answers to our questions. Figure out possible uncertainties.
Week 5	October 12	Finalizing and submitting the project plan. Domain exploration. Choosing the collaborational tools.Research possible algorithms. Agree on which algorithms we will use
Week 6	October 18	Start coding. Tokenization + cleaning
Week 7	October 26	Tokenization + cleaning should be ready.
Week 8	November 2	Split the data, PCA(?), preliminary (simpler) model selection and experimentation.
Week 9	November 9	Initial visualization of the data for basic evaluation of progress.
Week 10	November 16	Continue prior work. Prepare for presentation.
Week 11	November 23 (17-19)	Giving the intermediate presentation. Getting feedback and if necessary, adjusting the objectives/methods.
Week 12	November 30	Cross-validation, evaluation
Week 13	December 7	Cross-validation, evaluation. Visualization of results.
Week 14	December 14	Final small steps of the project to be finalized. Prepare for the final presentation.
Week 15	December 21 (15-17)	Final presentation.