1. Privacy and Legal Frameworks (GDPR)

The source file terspegelt.json contains personal data of guests, such as the displayName (e.g., "Anita Weenink" or "Jan Aarts"). In accordance with the General Data Protection Regulation (GDPR), specifically **Article 5(1)(c)** regarding "Data Minimisation", it is essential that this data is limited to what is strictly necessary for the intended purpose.

- **Anonymity & Data Minimization:** The data pipeline in clean_reviews.py operationalizes the GDPR principle of "Data Minimisation". The script processes raw reviews into a structured format where personal identifiers are stripped or simplified. By converting string ratings like "FIVE" to integers, the focus shifts from the *individual* to the *metric*, ensuring that the final dataset (final_data_for_powerbi.json) aggregates trends rather than exposing individual guest identities.
- **Transparency:** The project utilizes publicly available reviews from platforms such as Google Maps. This aligns with legal standards for web scraping where data is used exclusively for internal quality improvement, respecting the balance between public data access and individual privacy.

2. Ethical AI and Algorithmic Integrity

- **Advanced Dutch Sentiment Analysis:** The sentiment analysis in analyse_sentiment.py utilizes the **RobBERT-v2** model. As detailed in the official model card, this model was pre-trained on the Dutch part of the OSCAR corpus and fine-tuned on book reviews to achieve a 93.3% accuracy. Unlike generic multilingual models, RobBERT is specifically optimized for Dutch syntax and cultural nuance, as described in the paper "RobBERT: a Dutch RoBERTa-based Language Model".
- **Granular Analysis:** To further mitigate bias, the analysis is performed at the sentence level using nltk.sent_tokenize. This granular approach prevents a single negative sentence from skewing the interpretation of an entire review.
- **Objective Topic Modeling:** The project utilizes **BERTopic**, a neural topic modeling technique. According to the research paper "BERTopic: Neural topic modeling with a class-based TF-IDF procedure", this method leverages transformers and c-TF-IDF to create dense clusters that are easily interpretable. By using this automated approach in analyse_topics.py, the project avoids the human bias inherent in manual tagging, identifying issues purely based on data frequency and relevance.
- **Contextual Objectivity:** To ground these subjective insights in objective reality, merge_with_weather.py integrates historical weather records from weather_data.csv. This prevents management from misinterpreting negative sentiment caused by external factors (like rain) as service failures.

3. Professional Integrity and Intellectual Property

- **Licensing:** The repository is provided under an **MIT License**. As defined by the Open Source Initiative, this permissive license grants permissions for private and commercial use, distribution, and modification, provided the license and copyright

notice are preserved. This choice reflects a commitment to open standards and collaboration.

- **Accountability & Reproducibility:** The automation of the pipeline via run_pipeline.py ensures a reproducible workflow. By sequentially running cleaning, sentiment analysis, and topic modeling, and managing data hygiene by removing intermediate files, the project minimizes the risk of human error or manipulation, ensuring the integrity of the insights delivered to TerSpegelt.

References

Delobelle, P., Winters, T., & Berendt, B. (2020). *RobBERT: a Dutch RoBERTa-based Language Model*. Findings of the Association for Computational Linguistics: EMNLP 2020. https://huggingface.co/DTAI-KULeuven/robbert-v2-dutch-sentiment

European Parliament & Council of the European Union. (2016). *Regulation (EU) 2016/679 (General Data Protection Regulation)*. https://gdpr-info.eu/art-5-gdpr/

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv preprint arXiv:2203.05794. https://arxiv.org/abs/2203.05794

Open Source Initiative. (n.d.). *The MIT License*. https://opensource.org/license/mit