ITE Project Report

1. Overview & Objectives

This report summarizes the analysis of visitor and exhibitor data from a corporate event. Our goal was to understand attendee profiles, identify networking patterns, and develop a system to recommend relevant connections between visitors and exhibitors. Key objectives included data cleaning, participant analysis, and building a practical matchmaking tool.

---

2. Data & Methodology

- **Data Sources:** We utilized visitor registration details (including JSON answers), exhibitor profiles, and category information provided in separate CSV files.
- **Processing:** Raw data required significant cleaning and structuring.
    - Visitor answers were extracted from JSON and linked to questions.
    - Exhibitor categories (initially pipe-separated) were normalized.
    - Consistent text processing (lowercase, removing stopwords/special characters, lemmatization using NLTK) was applied to both visitor answers and exhibitor categories (utils.py) to enable effective comparison. The processed data is available in processed_visitors_answers.csv and processed_exhibitors_categories.csv.
- **Analysis:** We performed Exploratory Data Analysis (EDA) using Pandas to understand distributions and relationships (e.g., visitor demographics, purchasing roles, popular exhibitor categories). Key findings are visualized in the accompanying notebooks.
- **Matchmaking:** A content-based recommendation model was developed using keyword matching.

---

3. Key Visitor Insights

- **Attendance:** A diverse group attended, primarily motivated by sourcing products/services, gathering information, or promoting their own offerings. [Reference: Reason for Attending Chart]

- **Primary Business:** The majority of visitors identified as "Travel Agents" (54), followed by "Tour Operators" (19). [Reference: Company's Main Area of Business Chart]
- **Purchasing Power:** Many visitors (especially Travel Agents) have 'No influence' or 'Advisory' roles in purchasing. Tour Operators and Event Managers, though fewer, tend to have more direct purchasing responsibility and higher budgets. [Reference: Purchasing Decision Role / Budget Charts & Cross-plots]
- **Job Functions:** Common roles included "Visa support," "Sales," and "Media," aligning with the travel/tourism focus. [Reference: Job Function Chart]

---

4. Key Exhibitor Insights

- **Focus Areas:** Exhibitors strongly represented core travel sectors like Accommodation Providers, Tour Operators, Museums & Parks, and Tourist Boards. [Reference: Top Exhibitor Categories Chart]
- **Category Breadth:** While the average exhibitor listed ~4-5 parent categories, a notable number (12/35) listed 6 or more, with a maximum of 10. This suggests some exhibitors have very broad service offerings. [Reference: Exhibitor Category Distribution / Top Exhibitors by Category Count Charts] But this could also mean that some exhibitors just tell that they provide all these services just to boast.
- **Implication:** The broad category selection by some exhibitors highlighted the need for a mechanism to prioritize more focused matches in the recommendation model.

---

5. Matchmaking Model: Connecting Visitors & Exhibitors

- **Approach:** The model calculates a relevance score based on the overlap (keyword intersection) between a visitor's interests (derived from their answers to relevant questions like 'main business area' and 'job function') and an exhibitor's listed categories.
- **Keyword Matching:** Uses the cleaned and lemmatized terms generated during data processing.
- **Penalization Strategy:** To improve recommendation quality, exhibitors with a high number of unique parent categories (above a threshold of 6 [since the 75^{th} percentile of the data was at 6], based on EDA) receive a score penalty. This slightly down-ranks overly

broad exhibitors, favoring more specialized matches. The penalty strength (PENALTY_ALPHA=0.5) is adjustable. (calculate_match_score in utils.py).

- **Functionality:**
    - **Visitor -> Exhibitor:** Recommends top 7 exhibitors for a given visitor email (recommend_exhibitors_by_visitor_email.ipynb).
    - **Exhibitor -> Visitor:** Recommends top 7 visitors for a given exhibitor ID (recommend_visitors_by_exhibitorID.ipynb).
    - **Interest -> Exhibitor:** Recommends exhibitors based on a custom list of interest terms (recommend_exhibitors_by_answers.ipynb).
    - *(Examples of recommendations are available in the respective notebooks.)*

---

6. Potential Model Enhancements

- **Limitations of Keyword Matching:** The current model relies on exact (or lemmatized) keyword overlap. It may miss relevant connections where different wording is used for similar concepts (e.g., "lodging" vs. "accommodation").
- **Semantic Understanding:** Future iterations could significantly improve accuracy by incorporating **semantic matching using word/sentence embeddings** (e.g., Word2Vec, Sentence-BERT). This would allow the model to understand the *meaning* behind the text, not just the words themselves.
- **Data Requirement:** The effectiveness of semantic models heavily relies on having **richer, more descriptive text data**. If visitors provided more detailed free-text answers about their interests, or if exhibitors had detailed descriptions of their services beyond just category names, embedding-based models would yield much more nuanced and accurate recommendations.

---

7. Addressing Advanced Insights (Bonus Tasks)

- **Predictive Analysis (New Categories):**
    - **Approach:** Analyze visitor interest trends (keywords from answers). Identify frequently mentioned interests that *don't* strongly map to existing exhibitor categories. Techniques like topic modeling (NMF) on visitor answers (if richer text

were available) could reveal emerging themes. Correlating visitor interests with high engagement metrics (if tracked) could also highlight unmet needs.

    - **Current Data Limitation:** With primarily categorical answers, predicting entirely *new* categories is challenging. The analysis would be limited to identifying potential *sub-categories* or highlighting existing categories that are underserved by current exhibitors based on visitor profiles.

- **Cluster Analysis (Visitor Segmentation):**
    - **Approach:** Apply clustering algorithms (e.g., K-Means on numerically encoded answers, or more advanced methods like K-Prototypes for mixed data types) to group visitors based on their answer patterns across multiple questions. Analyze the characteristics (demographics, interests, budget) of each resulting cluster.
    - **Current Data Limitation:** While clustering is feasible with the current data, the resulting segments might be somewhat broad due to the limited number of distinguishing features (questions asked). More granular or free-text questions would allow for the identification of more distinct and insightful visitor personas.

---

8. Summary & Model Value

- **Key Finding:** The event attracts distinct visitor groups, particularly information-seeking Travel Agents and purchase-focused Tour Operators/Event Managers. Exhibitor profiles vary in focus, with some being very broad.
- **Model Value:** The matchmaking system provides a data-driven way to suggest relevant connections, enhanced by a penalty for overly broad exhibitors.

---

9. Conclusion

This analysis demonstrates the value of leveraging registration data to understand event participants and facilitate meaningful interactions. The developed matchmaking model, while based on keyword matching, provides a significant improvement over random networking and includes a mechanism to handle exhibitor diversity. By implementing the recommendations and considering future enhancements, particularly around richer data collection and semantic analysis, the event can further increase value for both visitors and exhibitors.