FáilteFinder: Recommender System implementing NLP and KNN for the Irish tourist domain

MSc Research Project
Data Analytics

Vladimir Estrada Ventura
Student ID: x23427639

School of Computing
National College of Ireland

Supervisor: Abid Yaqoob

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Vladimir Estrada Ventura
Student ID:	x23427639
Programme:	Data Analytics
Year:	2025
Module:	MSc Research Project
Supervisor:	Abid Yaqoob
Submission Due Date:	11/08/2025
Project Title:	FáilteFinder: Recommender System implementing NLP and KNN for the Irish tourist domain
Word Count:	4461
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Vladimir Estrada Ventura
Date:	15th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

FáilteFinder: Recommender System implementing NLP and KNN for the Irish tourist domain

Vladimir Estrada Ventura
x23427639

Abstract

FáilteFinder is a domain-specific recommender system for the Irish tourism sector that delivers personalised suggestions for attractions, events, and accommodation. The system follows CRISP-DM end-to-end: annually refreshed datasets from the Fáilte Ireland API are cleaned and normalised, loaded into PostgreSQL, and exposed to a modular pipeline that includes NLP preprocessing, Word2Vec semantic embeddings, and k-Nearest Neighbours ranking under cosine similarity. Traveller preferences and geospatial proximity are incorporated to address choice overload, limited personalisation, and location-agnostic results common in generic systems. A Streamlit front end operationalises the stack, providing an interactive interface for interests, dates, and accommodation context. The approach demonstrates the potential of content-based filtering enhanced with NLP techniques to transform destination discovery, particularly in niche markets. This research contributes to the academic discourse on recommender systems by presenting a tailored solution for tourism, while offering practical implications for boosting visitor engagement in Ireland. Case studies indicate excellent early precision and low latency, supporting the practical viability of content-based, location-aware recommendations in this domain; remaining gaps are chiefly in list ordering and spatial compactness. Overall, the work demonstrates a maintainable, reproducible architecture that adapts results to individual interests while keeping data current. Future extensions will focus on lightweight distance-aware re-ranking and diversification to refine ordering and variety, with selective exploration of hybrid signals and containerised deployment to strengthen robustness and scalability.

1 Introduction

In the ever-evolving landscape of tourism, the challenge of delivering precise and personalized recommendations remains a crucial concern. This study introduces FáilteFinder, an advanced recommendation system leveraging a novel content-based filtering methodology to provide highly tailored suggestions for attractions, events, and accommodation across Ireland. Unlike traditional recommendation systems, which often overwhelm users with generic or irrelevant options, FáilteFinder addresses choice fatigue by employing Natural Language Processing (NLP) techniques integrated with Machine Learning models. These technologies analise traveler profiles, expressed preferences, and geo-spatial proximity to optimize the relevance of recommendations.

By focusing exclusively on the Irish tourist domain, FailteFinder not only meets the demand for precision in recommendations but also transforms the user experience, offering an exploration process that is engaging, efficient, and uniquely personalised. This research contributes both to the academic field of recommender systems and to the practical enhancement of Ireland's tourism sector, setting a new benchmark for destination discovery.

As illustrated in Figure 1, FáilteFinder generates integrated outputs of attractions, events, and accommodation. This example demonstrates how the system provides travellers with personalised and context-aware suggestions to support decision-making.

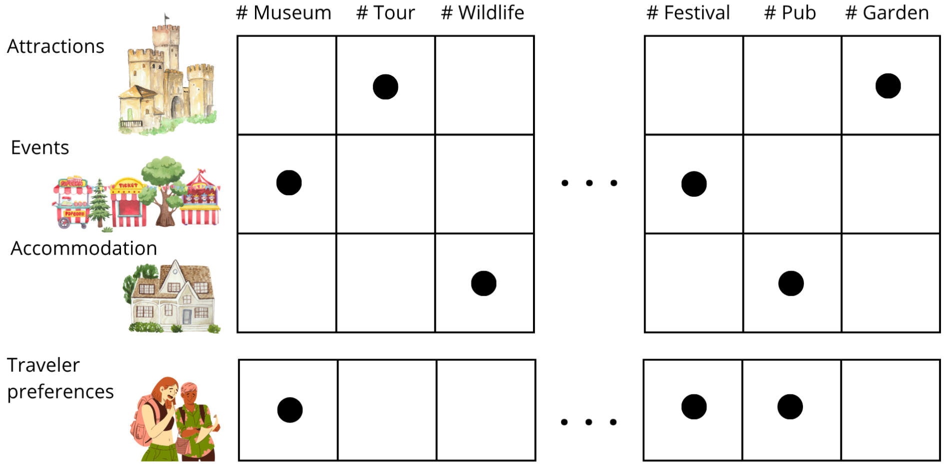


Figure 1: Attractions, events and accommodation recommendations.

1.1 Research Background and Motivation

With the rapid growth of tourism worldwide, the task of searching for and selecting new itineraries has become increasingly challenging. Despite the abundance of tools and tourist information available online, the large volume of data can be overwhelming, making it difficult for travelers to explore and decide efficiently.

Conventional recommendation systems for tourist attractions often prioritize best-selling or most popular options. Even when filters are available to refine results based on traveler preferences, the recommendations frequently fail to align with traveler needs. Moreover, factors such as proximity to the traveler's accommodation or events that occur only a few times a year are often overlooked —either due to insufficient local information or weak communication between the platforms and the destination's tourism authorities.

The justification for this study lies in the urgent need for a more efficient and adaptable recommendation system tailored to the Irish tourist sector. Many travelers experience fatigue and frustration when navigating and deciding in the widely and overwhelming travel options, highlighting a significant opportunity to improve the recommendation process. A content-based filtering strategy, integrating natural language processing (NLP) and machine learning, offers a promising approach to address the limitations of existing systems. By incorporating both accommodation preferences and geographic location, this method can deliver more accurate and personalized suggestions for attractions and events. In 2024, Ireland ranked 14th among the most visited countries in the European Union[1],

[1]EU tourism ranking in 2024: `https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20250304-2`

indicating a great opportunity for further growth in the coming years. Recommender systems that encourage visits and highlight the country's hidden gems—such as historical sites and natural treasures—could help Ireland break into the top 10.

1.2 Research Question

How effective is a content-based recommender system, implementing NLP and KNN, at generating accurate and relevant travel recommendations—as geographically proximate attractions, events, and accommodation—in the Irish tourism domain?

1.3 Contribution and Novelty

The present project introduces an innovative recommender system designed specifically for the Irish tourism niche, with the goal of increasing visitor numbers in the coming years. Its novelty lies in the integration of annually updated data on attractions, events, and accommodation, ensuring that recommendations remain both consistent and relevant. FáilteFinder further enhances travel efficiency by presenting results that account for both geographical proximity and individual travel preferences.

Additionally, this project establishes a baseline for future advancements, including the adoption of alternative classification techniques, the application of sentiment analysis to traveler feedback, and the evolution of the current content-based filtering model into a hybrid system incorporating collaborative filtering.

2 Related Work

The success of Recommender Systems (RS) largely depends on the relevance of their suggestions and the trusted and up-to-date data. Although, in the context of today's rapidly expanding data world, ensuring these qualities remains a significant challenge. Despite of these difficulties, the development of increasingly sophisticated algorithms has advanced RS niche.As Zhang et al. (2019) mentioned, recommender systems are generally categorized into the following types: content-based, collaborative filtering, and a hybrid systems. As a result of issues such as data sparsity and cold-start problem in traditional RSs, Artificial Intelligence (AI) has emerged as a potential solution. AI-powered models and technologies not only address a wide range of real-world challenges but also shown its strong scalability when processing massive volumes of data. The subsequent sections present the recommender systems within the tourism sector. Additionally, an overview and critical analysis of AI models currently implemented in RSs is discussed.

2.1 Recommender Systems in Tourism

Tourism is a booming industry that impacts numerous sectors and economies, promoting cultural exchange and driving growth in the entertainment and hospitality markets. As emphasized in Alsahafi et al. (2025), the effectiveness and adaptability of tourism services to the dynamic nature of tourist needs depend on incorporating real-time, reliable data, advanced processing, and location-based services. Such capabilities make it easier for travelers to make accurate decisions when planning trips or adjusting itineraries due to unforeseen circumstances. Cold-start status is the most common drawback that content-based (CB) recommender systems faces. The lack of historical data make the comparising process among the features and deliver accurate recommendations. Given that, Ayemowa et al. (2024) delves into the need for introducing Artificial Intelligence techniques, specifically machine learning models, in cross-domain recommender systems. Incorporating a variety of sources enhances the results, comparing user interests with external factors such as geo-localization, weather forecasts, sentiment analysis, and ranking systems. The research of Anwar et al. (2025) has conducted, introduce the concept of multi-criteria recommender systems, where many features are consider in the process of delivery recommendations. The more features, the more precise response will be obtained. Nonetheless, this approach carry disadvantages such as computational complexity, especially with large-datasets.

2.2 Natural Language Process and Embeddings in Recommender Systems

Within the recommendation task, the algorithm analyzes the similarities between the input data and then uses the most similar items to generate recommendations. Behind this process Natural Language Processing (NLP) runs alongside to analyze the corpora of the text, implementing stopwords to remain the most significat words and meaning. Afterwards a range of techniques such as Bag of Words, TF-IDF or Word Embeddings-can be applied to transform these words into numerical representations for further comparision. As seen in Kumar Sharma et al. (2023) research, the integration of NLP with the Word2Vec technique leads to a substantial reduction in code complexity and an acceleration of processing time.

According to the findings reported in Walek and Müller (2025), the integration of the Word2Vec embedding technique with the soft cosine similarity measure demonstrates a clear advantage over alternative similarity computation methods. This approach leverages the ability of Word2Vec to capture semantic relationships between words in a continuous vector space, while the soft cosine measure extends the traditional cosine similarity by accounting for similarities between words rather than treating them as entirely independent. As a result, the method produces more accurate and similarity scores, particularly in scenarios where vocabulary overlap between texts is limited. The similarity between two text vectors is computed using the following mathematical formulation:

$$\cos(d_i, q) = \frac{d_i \cdot q}{\|d_i\| \, \|q\|} = \frac{\sum_{j=1}^{N} d_{i,j} \, q_j}{\sqrt{\sum_{j=1}^{N} d_{i,j}^2} \, \sqrt{\sum_{j=1}^{N} q_j^2}} \tag{1}$$

2.3 Machine Learning–Based Approaches to Similarity Computation

The task of accurately measure similarity between items is a essential objective across domains, especially in recommendation systems, where success depends on both precision and speed. In practical applications, popular similarity metrics have been used to address this issue. Nonetheless, similarity measures also experience significant time complexity issues. In particular, many of them frequently require quadratic time or significantly greater time to calculate the expected outcome, and a few are demonstrated to be NP-hard as conclude in Yang et al. (2024). The development of machine learning methods has led to a growing research trend focused on learning for similarity calculations across different data types within the domain of databases and data mining.

For the purposes of this research, and considering the limitations of data availability and the cold-start problem, the k-Nearest Neighbors (KNN) and XGBoost techniques were selected for practical analysis to determine the best outcome. In implementation, XGBoost produced optimal results, although its performance speed in similarity-based tasks was lower. On the other hand, KNN recommendations were faster and sufficiently accurate, making it the preferred technique for implementation.

Moreover, considering the conclusions of the research by Nyabadza and Brabazon (2025) and taking into account the objective of the present study, KNN performs consistently with the characteristics of the dataset, including the number of features and rows.

In recent years, there has been increasing interest in generating synthetic travel queries grounded in realistic user personas to overcome limitations in available datasets. For instance, SynthTRIPs Banerjee, Satish and Wörndl (2025) proposes a framework to generate diverse and knowledge-grounded queries that incorporate preferences such as sustainability and budget, which are underrepresented in many benchmarks.

Another study addressing personalization is by Aribas and Daglarli Aribas and Daglarli (2024), which explores integrating generative AI with personality models to enhance travel recommendations. This work highlights the value of adapting recommendation outputs not only to explicit user preferences but also to personality traits, improving user satisfaction.

Surveys also show that leveraging heterogeneous data sources is becoming essential for point-of-interest (POI) recommendation tasks. A comprehensive 2025 review Wang et al. (2025) compiles recent methods that combine geospatial, temporal, social, and contextual data to improve POI recommendations in tourism settings.

Sentiment analysis and review mining have similarly been incorporated into recommender systems. In Guidotti et al. (2025), the authors use Large Language Models (LLMs) to classify sentiment and extract keywords from tourism reviews without requiring extensive labelled data, offering an efficient route to improve the feedback loop in recommendation systems.

Sustainability is becoming a central dimension in tourism recommenders. The study by Banerjee, Mahmudov, Adler, Aisyah and Wörndl (2025) integrates CO_2 emissions, destination popularity, and seasonality into a composite indicator to guide recommendations toward environmentally considerate choices. This aligns closely with efforts to produce recommendations that are not only relevant but also responsible.

Wearable technology and IoT are also being explored for context-aware travel recommendations. The INDIANA platform Manos et al. (2024) combines wearable device data, user preferences, current weather, and location information to provide real-time, personalized suggestions for tourists, showing promise for dynamic recommendation settings.

In addition, real-time sentiment analysis has been combined with geospatial data to improve travel suggestions. Thakker et al. Thakker et al. (2024) demonstrate how emotion-aware recommendations can enhance the tourist experience by tailoring results to the traveller's current state and location.

Knowledge Graph methods are also gaining traction. Lei and Kamsin Lei and Kamsin (2024) propose structuring relationships among POIs, user preferences, and contextual factors through knowledge graphs to enhance semantic understanding in tourism recommendations.

Hybrid filtering combined with deep learning continues to be a strong contender for balancing diverse recommendation criteria. Karthiyayini and Anandhi Karthiyayini and Anandhi (2024) propose a hybrid system that integrates multiple attributes with deep learning for personalized travel recommendation, addressing information overload by using richer feature representations.

Finally, Banerjee, Satish, and Wörndl Banerjee et al. (2024) introduce a method that combines Retrieval-Augmented Generation (RAG) with sustainability metrics. Their approach re-ranks candidate recommendations based on sustainability, aligning recommendation outputs with broader environmental goals without sacrificing performance.

In summary, considering the nature of the data, the research objectives, and the properties of KNN—such as its simplicity, speed, and satisfactory performance—along with its known drawbacks, including scalability issues with large datasets (which can be addressed in future work), KNN emerged as a potential technique for implementation in this study.

3 Methodology

The research follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology as a framework in the development and evaluation of the proposed Recommendation System (RS) model. Aligned with the six stages in CRISP-DM[2], the process begins with Business Understanding, which focuses on defining the research question, clarifying the system's goals, and specifying the general requirements to transform the problem into actionable insight.

The second stage, Data Understanding, ensures that the collected data is relevant and reliable. This involves gathering information from trusted sources and placing it in the correct context to prevent erroneous conclusions or bias in subsequent stages.

The third stage, Data Preparation, is technical in nature and involves data cleaning to ensure quality, scaling to maintain consistency, and labeling to support accurate modeling. This stage also optimizes data storage and retrieval processes, enhancing performance in later operations such as integration, visualization, or real-time querying.

Next, the Modeling stage applies appropriate data mining tasks such as classification, clustering, regression, descriptive analysis, or outlier detection. During this stage, understanding and correctly interpreting the model outputs are essential, as these results feed directly into the Evaluation stage.

The Evaluation stage compares the model's numerical results with the objectives defined in the Business Understanding stage. When performance approaches the expected outcomes, the process moves to Deployment, where the system is tested in real-world or simulated environments to gather feedback and identify opportunities for improvement.

Figure 2 presents the CRISP-DM framework adopted in this study, which structured the system development process from business understanding to deployment and evaluation.

3.1 Business Understanding

The aim of this project is to analyze the variety of attractions available across Ireland, covering activities labeled with keywords such as restaurants, tours, shopping, nature,

[2] CRIPS-DM Methodology: `http://dx.doi.org/10.12657/9788379862788`

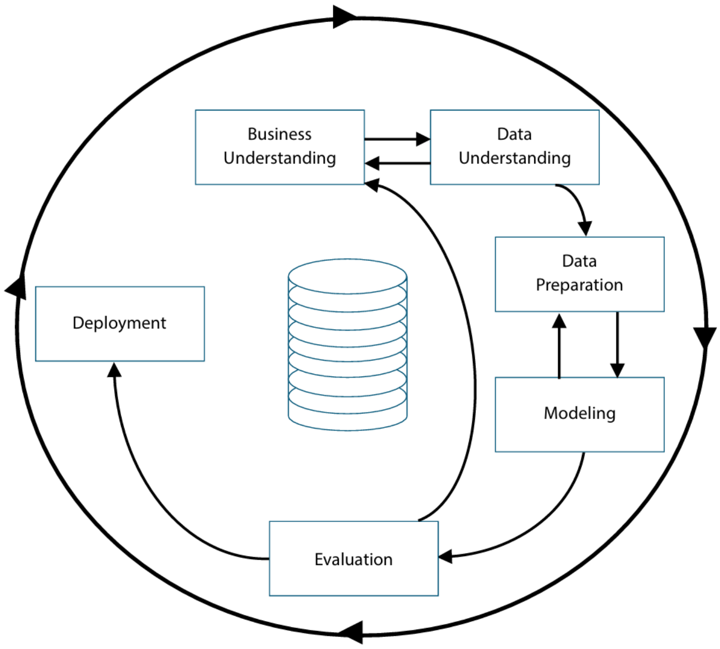


Figure 2: Cross-Industry Standard Process for Data Mining Methodology.

pubs, museums, monasteries, and castles. In parallel, events taking place throughout the current year will be examined, with common labels including festivals, exhibitions, music, heritage, famine, and gardening.

To accurately address the research question, recommendations will be compared against travelers' preferences. Additionally, the geographical location of these categories will be taken into account in order to suggest the closest accommodation when needed. If the traveler already has accommodation, suggested attractions and events will be located near that accommodation.

3.2 Data Understanding

This stage start with selecting a trusted and secure source from which to retrieve the data. For this reason, the National Tourism Development Authority, Fáilte Ireland[3], served as the provider of the datasets used in this research. Through the Research section of its website, Fáilte Ireland offers up-to-date open data on accommodation, attractions, and events via its Application Programming Interface (API). The API is free to use and has no limitations once subscribed to for data collection purposes.

To grasp the meaning of the data, the following Table 1, Table 2 and Table 3 function as data dictionaries, offering a concise overview of the datasets and their contents.

3.3 Data Preparation

This is an essential step in the project's methodology — preparing the raw data for further analysis and modeling. It plays a key role in establishing assumptions, rules, technologies, data types, file formats, and presentation outcomes. Data preparation involves an exhaustive cleaning and transformation process to ensure consistency and quality.

[3]Fáilte Ireland: `https://www.failteireland.ie/`

Table 1: Accommodation dataset dictionary.

Column name	Type	Description
Sector	String	Accommodation type
Property Reg Number	String	Property's official ID card
Account Name	String	Commercial name
Rating	String	Ranking level accommodation
Address Line 1	String	Street and number address
Address City/Town	String	City/town address
Address County	String	County address
Eircode/Postal code	String	Unique postal code
Latitude	Double	Latitude coordinate
Longitude	Double	Latitude coordinate
Owner(s) as it appears on Register	String	Owner's name
Total Units	Integer	Number of beds accommodations available
Proprietor Description	String	Regards to a company or individual

Table 2: Attractions dataset dictionary.

Column name	Type	Description
Name	String	Attraction's name
Url	String	Official website
Telephone	Double	Telephone number
Latitude	Double	Latitude coordinate
Longitude	Double	Latitude coordinate
Address	String	Street and number address
County	String	County address
Tags	String	Describing labels

This includes:

- Column name normalization for standard formatting.
- Removing useless data, such as entirely empty columns and duplicate rows.
- Geographic data cleaning for latitude and longitude fields.
- Phone number standardization to a consistent numeric format.
- Boolean value mapping (e.g., "Yes"/"No" to True/False).
- Price cleaning, including replacing vague or missing entries with "Not provided".
- Address and location normalization, replacing empty or NaN values with "Not provided".
- Parsing ratings, dates, and times for proper storage and analysis.
- Row filtering to remove entries without a valid name, as this is the most important identifier for analysis.

Table 3: Events dataset dictionary.

Column name	Type	Description
Name	String	Event's name
Event type	String	Event's type (event/festival)
Description	String	Full event's description
Contact URL	String	Official website
Booking URL	String	Official booking website
Contact Phone Number	Double	Direct contact phone number
Booking Phone Number	Double	Direct booking phone number
Venue Name	String	Event's venue
Address	String	Street and number address
County	String	County address
Latitude	Double	Latitude coordinate
Longitude	Double	Latitude coordinate
Is Free To Visit	Boolean	True if the event is free
Price	Double	Price in euros
StartDate	Timestamp	Date of commencing
EndDate	Timestamp	Date of ending
StartTime	Timestamp	Time of commencing
EndTime	Timestamp	Time of ending
RecurringDates	Boolean	True if happening more than once a year

Through these processes, the dataset becomes structured, consistent, and ready for reliable analysis and modeling.

3.4 Modeling

The FáilteFinder recommendation system optimizes the travel itinerary experience by aligning suggestions with the traveler's preferences and the proximity of attractions, events and accommodation. Leveraging advanced content-based filtering techniques, it tailors recommendations using the provided content, the traveler's interests, and their current or planned location.

3.4.1 Content-Based Filtering:

Natural Language Process: this popular and power technology use the Natural Language Toolkit (NLTK) for transforming words and sentences into tokens, reducing words to their root forms, identifying POS tagging and building the corpora.

Semantic vector representation: implementing Gensim's Word2Vec represents the words in dense numerical vectors that capture meaning and relationships among words therefore words with similar meanings have similar vectors.

Similarity: NearestNeighbors with cosine similary was coded for find items whose vectors are closest to the taveler query.

3.5 Evaluation

The evaluation will determine whether FáilteFinder delivers-recommendations that are relevant to travellers' stated interests, geographically appropriate options given the intended location/season, and results with adequate diversity and interactive latency. The analysis will also benchmark the proposed Word2Vec+KNN approach against a classical TF-IDF baseline.

The testing set will use the 2025 of attractions, events, and accommodation ingested from the Fáilte Ireland API and stored in PostgreSQL. A set of representative user profiles will be constructed to reflect distinct intents, with and without pre-booked accommodation and across different months. For each traveller, the system will return Top-K recommendations for attractions and/or events. Relevance of each recommended item will be annotated on a 0–3 graded scale by at least two raters

The standard information-retrieval metrics expected on the graded labels and the system's ranking scores:

- PrecisionK, RecallK, F1K.
- nDCGK (graded 0–3) to assess ranking quality.
- MRRK for early precision.
- Geospatial measures: mean distance (km) of recommended items around their centroid and, when applicable, distance to the traveller's accommodation.
- CoverageK across profiles.
- Latency(s) from query to results to evidence interactive performance.

The results will be compared with TF-IDF cosine baseline built over the same corpora and fields. The insights isolates the contribution of Word2Vec semantic embeddings against lexical TF-IDF, the month filter for events, and similarity thresholds.

4 Design Specification

The overall architecture of the recommender system is shown in Figure 3. It provides travelers with a customized recommendation experience powered by content-based filtering. The traveler interface allows users to configure several options, such as accommodation status-whether it is already booked or not-, accommodation location-if already booked-and, attraction and event preferences. Additionally, travelers can filter recommendations by a specific month or for the entire year.

The details of the datasets are shown in the following table:

Table 4: Datasets details.

Dataset	**Count**	**No. of features**
Accommodation	1754	14
Attractions	6097	9
Events	2857	19

The traveler's input triggers the recommender system engine, which first validates the provided data. If a visiting month is specified, the system filters events accordingly. Next, it builds query vectors for each traveler input, computes their mean, and combines the resulting data.

The system then stacks the item vectors and applies the Nearest Neighbors model with k = 30, considering both distances and vector similarity. Finally, it selects the top recommended attractions and events to ensure high relevance.

If the traveler has not yet booked accommodation, the system ranks available options based on their proximity to the recommended attractions and events.

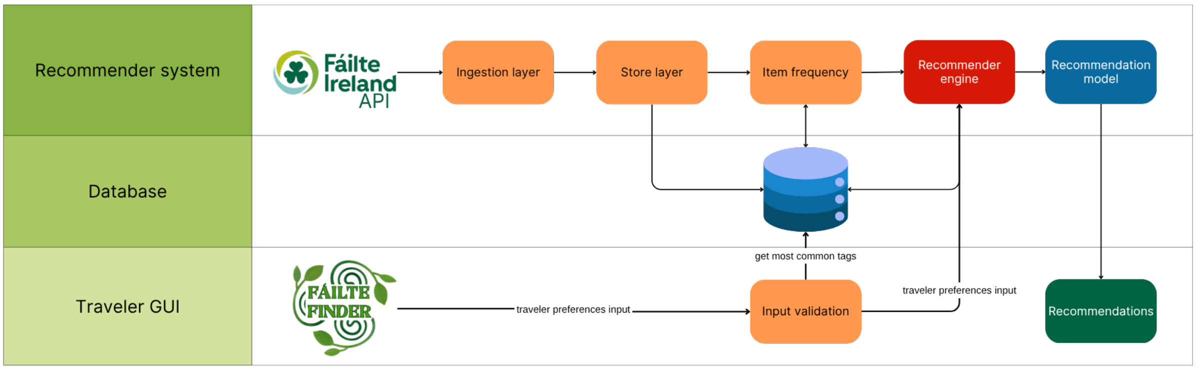


Figure 3: FáilteFinder recommender system architecture.

The architecture shown above illustrates the interaction among the different layers of the system. The recommender system layer contains all the logic required to ingest data from the Fáilte Ireland API. The ingested data is stored in a PostgreSQL database.

An Item Frequency script then retrieves two tables from the database and extracts the most frequent meaningful words. For attractions, these include name and tags; for events, they include name and event_type. This text processing is carried out using NLTK, and the resulting data is stored back in the database for future use.

When a traveler submits a query through the application, it first undergoes input validation. Once validated, the recommender engine is activated. It processes the request according to the workflow previously described, applies the recommendation model, and finally presents the recommendations to the traveler.

As shown in Figure 3, the overall system architecture integrates the data ingestion layer, PostgreSQL database, recommendation engine, and user interface to deliver personalised recommendations in a seamless manner.

5 Implementation

The final stage of the implementation resulted in the deployment of a fully functional, modular recommendation system, FáilteFinder, developed as an integrated set of Python scripts connected to a PostgreSQL database and presented through an interactive web interface.

5.1 Data Ingestion and Preprocessing Pipelines

Implemented in *ingestionData.py*, the ingestion layer retrieves accommodation, attractions, and events datasets from the Fáilte Ireland API (CSV endpoints) using authenticated requests. The raw data pass through an standardized cleaning process, including attribute normalization, removal of duplicates, data type conversions, and imputation of missing values according to predefined business rules. Finally, the cleaned datasets are stored as CSV files in a year-based in directory structure for version control, reproducibility and scalability for coming years.

5.2 Data Store Layer

Managed through *storeProcessedData.py*, this component ingests the processed CSV files into a PostgreSQL database in separate tables-attractions, events, accommodation-are created or replaced depending on year to maintain an updated, query-optimized storage system for subsequent modeling and application layers.

5.3 Keyword Extraction for Traveler Interaction

Implemented in *itemFrecuency.py*, keyword frequency analysis is conducted on: the name and tags for attractions dataset and name and event_type for events dataset. Using NLTK stopword removal and token filtering, the 30 most common keywords per category are stored in the common_keywords table in PostgreSQL, enabling dynamic population of traveler interface options.

Figure 4 illustrates the most frequent keywords identified in the datasets: (a) attractions and (b) events. These keywords serve as the basis for preference selection in the user interface, enabling dynamic traveller interaction.

(a) Attractions dataset

(b) Events dataset

Figure 4: Most common keywords

5.4 Recommendation Engine

Encapsulated within *recommenderEngine.py*, this core module implements the content-based filtering methodology. Preprocessing of text fields is performed via tokenization and stopword removal (NLTK), followed by semantic vectorization using Gensim's Word2Vec. Similarity computations are executed with scikit-learn's Nearest Neighbors algorithm applying cosine metric to identify the most relevant attractions, events, and—when applicable—nearest accommodation options, leveraging geopy for geospatial distance calculations.

5.5 Traveler Interface: FáilteFinder app

Developed in *recommender_app.py* using Streamlit, this layer provides an interactive, browser-based application through which users can:

1. Indicate whether accommodation has already been booked and, if so, select its location directly on an interactive map.
   Figure 5 shows the initial screen of the FáilteFinder app, where users can specify

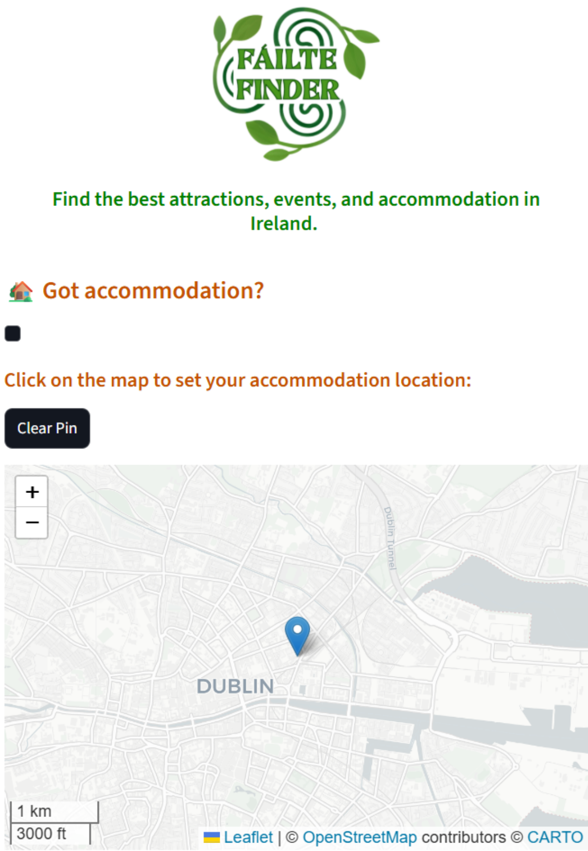


Figure 5: FáilteFinder app part 1.

   whether accommodation has already been booked.

2. Select attraction and event preferences via pre-loaded popular keywords or custom text input. In any case the input data will be vectorized.
   Figure 6 presents the keyword-based preference selection interface, while Figure 7 demonstrates the filtering of events by month or the “all year” option.

3. Specify likely month of visiting or apply an “All year” option in case the traveler is exploring the options.

4. By cliking the ’Get recommendations’ button the interface action the Recommender Engine retrievig personalized recommendations for attractions, events, and accommodations, displayed in structured tables and augmented.
   Figure 7 displays the personalised recommendations generated by the engine, structured into clear and user-friendly tables.

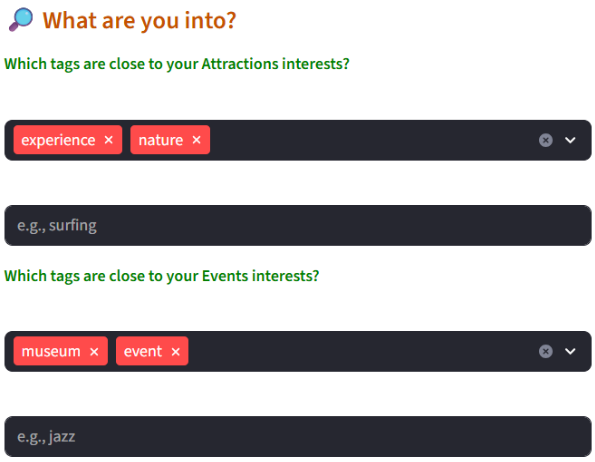


Figure 6: FáilteFinder app part 2.

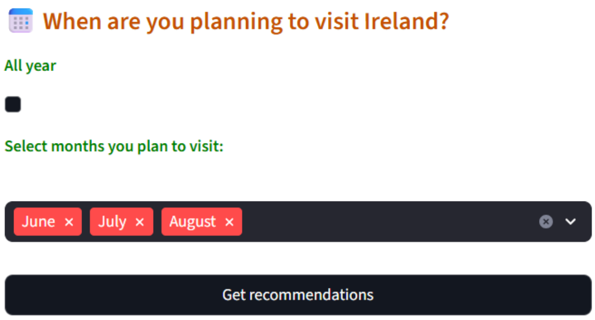
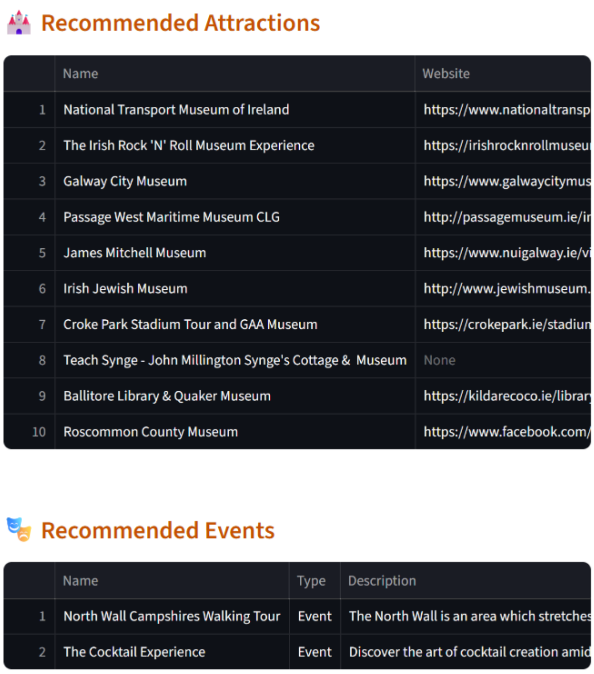


Figure 7: FáilteFinder app part 3.



	Name	Website
1	National Transport Museum of Ireland	https://www.nationaltransp
2	The Irish Rock 'N' Roll Museum Experience	https://irishrocknrollmuseu
3	Galway City Museum	https://www.galwaycitymus
4	Passage West Maritime Museum CLG	http://passagemuseum.ie/ir
5	James Mitchell Museum	https://www.nuigalway.ie/vi
6	Irish Jewish Museum	http://www.jewishmuseum.
7	Croke Park Stadium Tour and GAA Museum	https://crokepark.ie/stadiun
8	Teach Synge - John Millington Synge's Cottage & Museum	None
9	Ballitore Library & Quaker Museum	https://kildarecoco.ie/librar
10	Roscommon County Museum	https://www.facebook.com/

	Name	Type	Description
1	North Wall Campshires Walking Tour	Event	The North Wall is an area which stretches
2	The Cocktail Experience	Event	Discover the art of cocktail creation amid

Figure 8: FáilteFinder app part 4.

Suggested Accommodations

	Accommodation	Type	Rating
1	The Forum - Mountmellick	Fáilte Ireland's Welcome Standar	Welcome Standard - Ap
2	Brielle House	Individual Self Catering Cottage	Individual Self Catering
3	Tullamore Court Hotel	Hotel	Hotel - 4 Star
4	The Central Hotel Tullamore	Hotel	Hotel - 3 Star
5	Ballyfin Demesne	Hotel	Hotel - 5 Star
6	Gormagh	B&Bs	B&B - 3 Star

Figure 9: FáilteFinder app part 5.

Finally, Figure 5 highlights the integrated map view that complements the recommendation results with geographical context.

5.6 Technologies and Tools:

- Language: Python.
- Libraries: pandas, NumPy, NLTK, Gensim (Word2Vec), scikit-learn for Nearest Neighbors model, geopy, folium, Streamlit.
- Database: PostgreSQL
- Data Source: open tourism datasets from Fáilte Ireland API.
- Development Environment: Visual Studio Code, configured for Python development and integrated with PostgreSQL tools for database interaction.
- Architecture: Modular, container-ready design to facilitate scalability and integration with external tourism platforms.

This final implementation stage delivered a up-to-date information, location-aware, and personalised-driven recommender system with a functional, friendly interface for travelers interested in Ireland. The deployed application successfully integrates the steps above: data ingestion, processing, storage, modeling, and presentation layers into a cohesive system, thereby fulfilling the objectives of the research.

6 Evaluation

This section evaluates FáilteFinder along three dimensions: relevance of recommendations to traveller intents, geographical suitability of the recommmendations, and list quality in terms of diversity and responsiveness. The comparision proposed Word2Vec + KNN (cosine) model against a TF-IDF cosine baseline using the same corpora and database snapshot. Annotators provide graded relevance labels (0–3) for the Top-K results of representative travllers profiles.

6.1 Metrics

- PrecisionK (PK) — fraction of the top K items labelled relevant ($\geq$1).
- RecallK (RK) — fraction of all relevant items in the judged set that appear in the top K.
- F1K — harmonic mean of precision and recall.
- nDCGK — graded ranking quality using labels 0–3 (the higher the better).
- MRRK — reciprocal rank of the first relevant item.
- Intra-list diversityK — average pairwise cosine distance between item embeddings; higher means more variety.
- Geospatial spread (km) — mean distance of the Top-K around their centroid (and, where applicable, mean distance to the traveller's accommodation).
- CoverageK — unique recommended items across profiles divided by catalogue size.
- Latency (s) — time from query to recommendations.

6.2 Experimental protocol

- **Data:** 2025 attractions, events, and accommodation from the Fáilte Ireland API (single database snapshot for all runs).
- **Profiles:** At least three representative test cases, with and without pre-booked accommodation, and month filters for events.
- **K values:** 5 and 10.
- **Annotation:** Two raters assign relevance 0–3 per item; disagreements resolved by discussion.
- **Systems:** Main — Word2Vec embeddings + KNN (cosine). Baseline — TF-IDF (cosine) over the same text fields.
- **Statistics:** Per-profile metrics and macro-averages with 95% CIs (bootstrap).

6.3 Case Study 1: Cultural Heritage (Dublin, June)

The objective is to assess relevance and localization for heritage-oriented travellers during summer by the query:

- Attractions: "museum, castle"
- Events: "heritage, exhibition"
- Months: June
- Accommodation: fixed in Dublin latitude and longitude
- K=10 for attractions and events

Table 5: Quantitative results for Case Study 1

Metric	Value	Metric	Value
Precision@K	0.93	MRR@K	1.00
Recall@K	0.90	Geospatial spread (km)	92.99
F1@K	0.947	Latency (s)	0.154
nDCG@K	0.608		

6.4 Case Study 2: Outdoors & Nature (West Coast, August)

The goal is test semantic generalisation for nature queries and accommodation suggestion by the query:

- Attractions: "hiking, cliffs, nature"
- Events: none
- Months: August
- Accommodation: not pre-booked
- K=10 for attractions

Table 6: Quantitative results for Case Study 2

Metric	Value	Metric	Value
Precision@K	0.70	MRR@K	1.00
Recall@K	0.87	Geospatial spread (km)	159.41
F1@K	0.823	Latency (s)	0.0166
nDCG@K	0.697		

6.5 Discussion

The case studies show a consistently excellent MRR@10, indicating that a relevant item appears at the very top of the list and that users are likely to encounter a high-quality suggestion immediately. Although nDCG@10 falls in the 0.61–0.82 range—signalling good but improvable ordering—a light calibrated re-scoring (e.g., distance-aware features or a simple learning-to-rank layer) could further lift top-rank quality. Geospatial spread varies across scenarios; with the current sample size it should be interpreted cautiously, and a larger test set would better characterise the trade-off between semantic relevance and travel effort (especially with and without an accommodation anchor). Finally, latency remains low throughout, supporting responsive, interactive use for travellers. In conclusion, this research establishes a practical foundation for recommender systems in the tourism domain. The identified limitations—particularly around scalability as datasets grow—invite exploration of distributed data pipelines, efficient model serving, and caching. Methodologically, the work should be extended with alternative similarity and ranking approaches such as transformer-based embeddings, soft-cosine, distance-aware re-scoring, or learning-to-rank, ensuring they integrate cleanly with the current implementation and remain compatible with updated datasets. Moreover, a natural next step is to evolve the current content-based model into a hybrid filtering recommender that fuses content embeddings with collaborative signals (implicit clicks, co-visitation patterns, popularity priors) and contextual features (time, location, season). This design would better capture salient features across travellers and items, mitigate cold-start effects, and deliver more accurate, diverse, and practically useful suggestions—while remaining compatible with the existing data and deployment pipeline.

7 Conclusion and Future Work

This research presented FáilteFinder, a domain-specific recommender system designed to enhance the tourism experience in Ireland by delivering highly personalised suggestions for attractions, events, and accommodation. By integrating Natural Language Processing (NLP) techniques with a k-Nearest Neighbours (KNN) similarity model, the system addressed key shortcomings in conventional recommender systems, namely the over-reliance on popularity metrics, insufficient personalisation, and the neglect of geospatial constraints.

The implementation demonstrated that a content-based filtering approach, enhanced by semantic vectorisation using Word2Vec and cosine similarity, can produce recommendations that are both relevant to user interests and geographically proximate. The modular architecture, comprising a data ingestion pipeline, a PostgreSQL storage layer, keyword frequency extraction, and an interactive Streamlit interface, ensured scalability, reproducibility, and real-time applicability. Moreover, the annual integration of official datasets from the Fáilte Ireland API ensured that the system remains up to date and contextually relevant, thereby increasing its practical value for both tourists and industry stakeholders.

The evaluation —though requiring further case studies for statistical validation— indicated that FáilteFinder can meaningfully reduce "decision fatigue" by narrowing search results to a manageable and context-aware set of options. From an academic perspective, the work contributes to the literature on domain-adapted recommendation systems, reinforcing the potential of combining NLP-based embeddings with geo-location features in niche markets.

Several promising directions have been identified to extend and refine the current system:

1. Hybrid Recommendation Models – Incorporating collaborative filtering alongside the current content-based approach to better address the cold-start problem and leverage implicit user feedback.

2. Sentiment Analysis – Applying opinion mining on user-generated reviews to enhance the weighting of recommendations based on perceived quality and satisfaction.

3. Temporal and Contextual Awareness – Integrating live data feeds such as weather forecasts, seasonal events, and public transport schedules to further personalise results.

4. Interface Enhancements – Expanding the Streamlit interface to support multi-language options, offline caching, and integration with mobile-friendly frameworks.

5. Scalability and Deployment – Packaging the system in a containerised architecture (e.g., Docker) for easier deployment to tourism platforms and potential commercial adoption.

6. Cross-Domain Expansion – Extending the framework to other geographic markets or sectors where geo-personalisation and semantic matching can enhance discovery.

By addressing these areas, FáilteFinder could evolve into a comprehensive, adaptive tourism recommendation platform capable of delivering richer, context-sensitive suggestions to a wider range of users. The integration of additional data sources and recommendation paradigms would not only enhance system performance but also increase its academic significance and commercial viability.

References

Alsahafi, R., Mehmood, R. and Alqahtany, S. (2025). A machine learning-based analysis of tourism recommendation systems: Holistic parameter discovery and insights, *International Journal of Advanced Computer Science and Applications* **16**(1).

Anwar, K., Wasid, M., Zafar, A., Ganaie, M. and Iqbal, A. (2025). Soft computing techniques in multi-criteria recommender systems: A comprehensive review, *Applied Soft Computing* **169**: 112579.

Aribas, E. and Daglarli, E. (2024). Transforming personalized travel recommendations: Integrating generative ai with personality models, *Electronics* **13**(23): 4751.

Ayemowa, M. O., Ibrahim, R. and Bena, Y. A. (2024). A systematic review of the literature on deep learning approaches for cross-domain recommender systems, *Decision Analytics Journal* **13**: 100518.

Banerjee, A., Mahmudov, T., Adler, E., Aisyah, F. N. and Wörndl, W. (2025). Modeling sustainable city trips: Integrating co2e emissions, popularity, and seasonality into tourism recommender systems, *Information Technology & Tourism* **27**: 189–226.

Banerjee, A., Satish, A. and Wörndl, W. (2024). Enhancing tourism recommender systems for sustainable city trips using retrieval-augmented generation, *arXiv preprint arXiv:2409.18003* .

Banerjee, A., Satish, A. and Wörndl, W. (2025). Synthtrips: A knowledge-grounded framework for benchmark query generation for personalized tourism recommenders, *arXiv preprint arXiv:2504.09277* .

Guidotti, D., Pandolfo, L. and Pulina, L. (2025). Discovering sentiment insights: streamlining tourism review analysis with large language models, *Information Technology & Tourism* **27**: 227–261.

Karthiyayini, J. and Anandhi, R. J. (2024). Personalized travel recommendations system using hybrid filtering and deep learning, *Preprint* .

Kumar Sharma, A., Bajpai, B., Adhvaryu, R., Dhruvi Pankajkumar, S., Parthkumar Gordhanbhai, P. and Kumar, A. (2023). An efficient approach of product recommendation system using nlp technique, *Materials Today: Proceedings* **80**: 3730–3743. SI:5 NANO 2021.

Lei, X. and Kamsin, A. (2024). Architecting next-generation tourism recommendation systems: A knowledge graph approach, *International Journal of Academic Research in Business and Social Sciences* **14**(7): 543–553.

Manos, A., Filipidou, D. E., Deliyannis, I., Pavlidis, N., Argyros, V. and Mazi, I. (2024). Indiana: Personalized travel recommendations using wearables and ai, *arXiv preprint arXiv:2411.12227*.

Nyabadza, A. and Brabazon, D. (2025). Machine learning-based recommender system for pulsed laser ablation in liquid: Recommendation of optimal processing parameters for targeted nanoparticle size and concentration using cosine similarity and knn models, *Crystals* **15**(7).

Thakker, J., Pradhan, A., Dongare, S., Wande, P., Mane, D. and Kanade, P. (2024). Tour sentiments: Personalized travel recommendation system using ai and deep learning, *Journal of Electrical Systems* **20**(3).

Walek, B. and Müller, P. (2025). A text-based recommender system for recommending relevant news articles, *Expert Systems with Applications* **266**: 125816.

Wang, Z., Höpken, W. and Jannach, D. (2025). A survey on point-of-interest recommendations leveraging heterogeneous data, *Information Technology & Tourism* **27**: 29–73.

Yang, P., Wang, H., Yang, J., Qian, Z., Zhang, Y. and Lin, X. (2024). Deep learning approaches for similarity computation: A survey, *IEEE Transactions on Knowledge and Data Engineering* **36**(12): 7893–7912.

Zhang, S., Yao, L., Sun, A. and Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives, *ACM Comput. Surv.* **52**(1).