Project Charter

Date Modified: 10/20/2025
Team Name: Solid Home
Team Members: Zhen Liang, Renjie Zhang, Yiwen Kang
Project Topic Title: Automating Anti-Displacement Qualitative Research

Problem Statement

The City of Seattle recently established an Anti-Displacement Workgroup to better align income levels with rental prices and reduce the risk of residents being priced out of housing.

However, the current policy and planning documents are mostly in lengthy PDFs, which are time-consuming to review and inconsistent across chapters. This makes it difficult for users to quickly grasp policy content and hinders cross-department collaboration.

At the same time, the City faces similar challenges in other long-term initiatives—such as climate and transportation—so the client hopes to develop reusable and auditable text-analysis tools that can support policy decisions and external communication.

Goal Statement

The goal of this project is to develop an automated survey analysis model leveraging Large Language Models. By defining and extracting key tokens, the model will streamline the process of reading. Ultimately, the project aims to improve the efficiency and accuracy of survey analysis, enabling stakeholders to make more informed, data-driven decisions.

Specific: Build an end-to-end pipeline: PDF → structured text → semantic chunking → embeddings → BERTopic clustering → LLM topic naming/summaries → visualization.

Measurable:

- Parsing coverage ≥ 95% (readable pages / total pages); if OCR is triggered, OCR F1 ≥ 0.90.
- Topic coherence ≥ 0.45
- Executive summary hit-rate ≥ 80%

Achievable: Implemented with PyPDF/Unstructured + UMAP/HDBSCAN/BERTopic + GPT-series LLMs; technically feasible.

Relevant: Focus on comprehensive-plan themes (housing/anti-displacement/growth) aligned to city governance needs.

Time-Bound

- Week 6: Data review
- Week 7: Model Design
- Week 10: V1 finish
- Week 11: Client feedback
- Week 12: Modify
- Week 13: Final version

Scope

In-Scope:

1.Data Ingestion & Preprocessing

- Batch PDF ingestion; layout parsing (text & tables as text); optional OCR for scanned pages.
- Layout noise cleanup: TOC, headers/footers, footnotes; hyphenation/line-break repair; table rule noise suppression.
- Semantic chunking (title/section-aware + length/overlap windows); near-duplicate merging (MinHash/SimHash).

2.Modeling & topic engineering

- Embeddings; dimensionality reduction (UMAP); density clustering (HDBSCAN); BERTopic keywords.
- Topic quality evaluation (coherence; sampled fidelity review).

3.Outputs & visualization

- CSV/Parquet exports: topic inventory, keywords, exemplar snippets, document–topic soft distributions.
- Visuals: top-N topics, document coverage heatmap, importance bars/scatter, topic relationship graph.
- One-Pager executive summary + appendix (methods, metrics, reproducibility notes).

Out-of-Scope:

- Deep image/table number extraction or figure semantics
- External fact-checking
- Multi-tenant online platform

Stakeholders

- City of Seattle Program Staff
- Industry Mentor: Janis Jordan (she/her)
- Team member: Zhen Liang, Renjie Zhang, Yiwen Kang

Key Milestones

- Week 6: Cleaning , chunking, parsing.
- Week 7: Making a demo and meeting with the client.
- Week 8: Embedding/clustering/BERTopic baseline.
- Week 9: Implementing the topic merge strategy.
- Week 10: Visualization and testing.
- Week 11: Reviewing by the client and getting feedback.
- Week 12: The final version is done.

Team Members

Yiwen Kang – Ingest & Prepare Text: Build the input pipeline from PDF to clean paragraph chunks. Batch-read PDFs, strip headers/footers/TOC/footnotes, fix line breaks/hyphenation, and split into 300–800-character chunks with slight overlap

Zhen Liang – Embed, Cluster & Extract Keywords: Convert chunks to sentence embeddings, run a stable clustering method to assign each chunk a `topic_id`, and aggregate by topic. Print quick stats and add a simple check that sampled items within a topic look semantically related.

Renjie Zhang – Name Topics, Summarize & Deliver Outputs: Generate short, human-readable topic names, write concise per-topic summaries and a one-page document overview, and create a few essential charts, like topic sizes and topic × section heatmap).

Success Criteria

- Parsing coverage ≥ 95%.
- Topic coherence (c_v/c_npmi) ≥ 0.45.
- Content duplication ≤ 10%.
- Improving the qualitative analysis process by at least 20%.