1. **Intro**

   1.1. **Motivation**

   - Like Movies and videos as entertainment, but annoyed by with too obvious and irrelevant recommendations. Time to fix.
   - Diversity and Serendipity – often ignored, but very crucial pillars of Rec sys,
   - Just going diverse, has bigger chances to recommend irrelevant stuff.
   - Collaborative filtering is good, but doesn't really understand the user as better as content based does, hence, content-based.
   - Lack of Diversity and Serendipity, previous work based on old datasets that squizzed out info from the sparse data, techniques aren't suitable to tackle data of this new era.
   - Tag-genome is the powerful seed for the future of Rec Sys. - as it combines user ratings, IMDB reviews, movie tags etc. Such powerful dataset doesn't have enough research around it since 2012, as most audience have been busy trying to predict the user ratings accurately as possible using numerous techniques, despite the fact that user's actual rating in real system may vary significantly based on their moode, day, time, life events and many other seasonal and psychological aspects which cannot be judged in offline setting with the available data. But for same reason, achieving higher accuracies through predicting the ratings in offline settings is only has the tiniest contribution to the real world Recommender Systems problem.
   - The real challenge is to address the Diversity and Serendipity and Novelty of Recommendations. Key is to broaden user's preferences and still ensure user likes and enjoys the item. It's also highly challenging, yet most important to ensure not to go too much away from the user's usual preferences, hence, Diversity, Serendipity and Novelty go hand-in-hand.

   1.2. **Executive Summary**

   - **what your project is about, 3-4 para**
   - **main features and functional requirements**
   - **Para 1:**
     - Recommending Diverse, Serendipitous and Novel movie recommendations by using Tag-Genome data. Novel ranking algorithms, understanding user's preferences through clustering.
   - **Para 2:**
     - What is tag-genome, it potential to understand user through content, future scope and potential of tag-genome like systems.
   - **Para 3:**
     - Keeping trade-off between relevant items – user's usual preferences and diverse and serendipitous items.
   - **Para 4:**
     - Features and Functional Requirements
       - Novel recommendations which keeps trade-off between user rating the item high enough, brodens users preferences, user wouldn't have explored the item on their own but it's still relevant to them.
       - Creates appropriate number of logical categories about user's preferences using clustering. These categories are different for each user based on the movies they've already watched.
       - Benefit for the business to attract more movie makers to associate with them as well as the entertainment avid customers.
       - User watches 3 or more movies which has tag-genome score and rates them above the like threshold – 3 or 3.5

- Tag-genomes are continuously updated for upcoming movies so that the recency can be researched around thoroughly.

1.3. Contribution

- Explain how – novel reranking algorithms – both based on user's average preferences and user's all movies, very high level notion can be understood by business people.
- Lemmatization, Clustering to understand user's logical preferences, user profile inference.

1.4. Structure of this document

- Briefly describe the structure of this document, enumerating what does each chapter and section stands for.

2. Background

2.1. Thematic Area Within Computer Science

- refer guidelines in the thesis doc

2.2. Project Scope

- How serendipity positively impacts users choices, retains users and makes them like the recommendation system more, increasing both user's visit frequency to the system and the amount of time spent in the system;
- This also helps adding up to the total revenue, also diversity improves the chances for the cold items being recommended, which means, more popularity and revenue for all movie makers associated with the company who owns this system. Increasing the chances for the company to attract more movie makers to associate with them, and the users who's seeking the entertainment as well.
- Support above claim with strong citations if possible.

2.3. A Review of the Thematic Area

- YouTube channels about Recommender Systems
- Important professors at the University to subscribe to on Google Scholar – Derek Bridge, Adamopolous, someone from the serendipity and diversity related work.
- Important topics to watch -
- Movielens dataset sources and the website Movielens.com
- Google's research on Recommender Systems and websites ranking.
- Research from StackOverflow and Quora like systems about recommending appropriate questions to their users.
- Research articles and papers from IMDB, Netflix, Amazon Prime Video, HBO, Zalando, SoundCloud, FoodRecommender about Recommender Systems.
- Describe in detail on how Div and Serendipity would help attract more entertainment making companies and entertainment seeking users to use this system – explain in terms of cold items being recommended, revenue and popularity for movie makers, time spent by user in the system, and improvement in users visit frequencies to the system.
- Books, PhD thesis
- Popular GitHub repositories, courses like Andrew Ng's courses on Coursera, at CIT, at UCC and more.

2.4. Current State of the Art

- top 5 conferences, journals, papers, articles, videos. Sort to favour 1) your research idea, 2) Serendipity, 3) Diversity and 4)Novelty. And also some work related to tag-genomes.
- Cover Mesut, Marius Kaminskas, D. Kotkov and others work on Serendipity and Diversity
- Mesut's reranking algorithms and user profiling idea.
- Cover and focus on Novelty and diversity.
- Cover online / offline testing, and related papers like the serendipity-2018

- cover definitions of diversity, novelty and serendipity from papers, esp. Which considered 8 different questions around it, books, journals, Netflix and other companies research etc.not just movies, but also music recommendation, serendipitous research paper recommendation etc.

3. Problem – Recommending Diverse and Serendipitous Movies usign Tag-Genome

3.1. Problem Definition

- **4-6 pages:**
    - Start with describing the problem
        - Para 1: What's recommender system and the trends – from the past era, to the latest ones. State about the challenges, painful areas – diversity and serendipity with Novelty.
        - Para 2: combining all-in-one, ratings, community knowledge, movie tags altogether.
        - Para 3: Understanding user's preferences – averaging methods like mean() , or considering minimum distance for candidate item to every movie already watched by the user.
        - Para 4: Understanding the logical groups/clusters of user's preferences, state that how and why it should be different for each user, why is it better than just clustering all movies together, and justify with silhouette score tests per user vs on all movies.
        - Para 5: per-user cluster sizing problem, how it's done now, how it can be made faster, why agglomerative clustering etc.
        - Para 6: Ranking problem, state all the ranking approaches you would want to try – refer to the commented 4 possibilities from the code for ranking, now describe your novel ranking solution, further describe the possible weight parameters – cluster rank, diversity, similarity to the user's profile and user's rating for similar watched movies, also describe why these 4 parameters matter.
        - Para 7: Justify with results about the different possible values for above parameters – about their overall impact on all users for Diversity, Serendipity and Novelty.
        - Para 8: Describe importance of top-N items in RL – recommendation list belonging to the user's obvious choice. Show with results about how it affects on div, serendipity, novelty and overall quality of recommendations.
        - Para 9: Problem of offline vs online evaluations, Choosing right metric to assess diversity, serendipity, novelty etc.
        - Para 10: Lemmatized vs non-lemmatized
        - Para 11: Thresholded vs non-thresholded and why
        - Para 12: Genre vs Genomes, and why

3.2. Objectives

- To find the impact of serendipity, diversity and novelty offered by the proposed system over the recommendations from the state-of-the-art dataset.
- To analyze whether clustering on all movies works better or clustering per user based on the movies watched by each user works well.
- To explain the impact of Lemmatized variant of tag-genomes over the non-lemmatized full set of tag-genomes.

3.3. Functional Requirements

- **what:**
    - serendipitous recommendations, diversity in bottom n-1
    - high relevance in top-N recommendations

3.4. Non-functional Requirements

- Evaluation using Serendipity metric, diversity of a list metric etc.

4. Implementation Approach

4.1. Architecture

- follow guidelines from the template.

a) Use Case Description

- follow guidelines from the template.

4.2. Risk Assessment

4.3. Methodology

4.4. Implementation Plan Schedule

4.5. Evaluation

- follow guidelines from the template.

4.6. Prototype

- follow guidelines from the template.

5. Conclusions and Future Work

5.1. Discussion

- **Problems**
  - User profile aggregating based on movies watched by user – average method initially, later replaced with min distance from all movies watched by user.
  - Cluster sizing
    - whether Per Movie or per user
    - what size Per User, and assessment criteria as a solution.
  - Thresholding vs non-thresholding of relevance scores for genome-tags
  - Choice of evaluation metric for diversity, novelty, and serendipity.

5.2. Conclusion

- Overall conclusion on quality of recommendations and chances of user engagement, brodening of users preferences and still being useful.
- Conclusions around using right metrics to evaluate diversity, serendipity, novelty etc.
- Conclusions around lemmatization vs using all tag-genomes as it is(full set of genomes).
- Thresholding vs non-thresholding.
- Genre vs tag-genomes.
- Conclusions around understanding users preferences – averaging vs min dist of candidate movie to all watched movies.
- Conclusions around per user clustering vs clustering on all movies.
- Conclusions around best parameters for top-N items in RL.
- Conclusions around best parameter values for weights for cluster ranks, users' rating for similar watched movies, diversity and similarity to the users profile, and similarity to the watched movie from the chosen cluster.

5.3. Future Work

- **Add following as more to the existing:**
- A tradeoff between choosing too many movies out of one highly ranked sparse cluster vs choosing movies similar to movies from all sparse clusters.
- Understanding the best hyper-parameters for different set of user groups, where groups/clusters can be formed based on the number of movies watched, types of movies, rating habits etc.