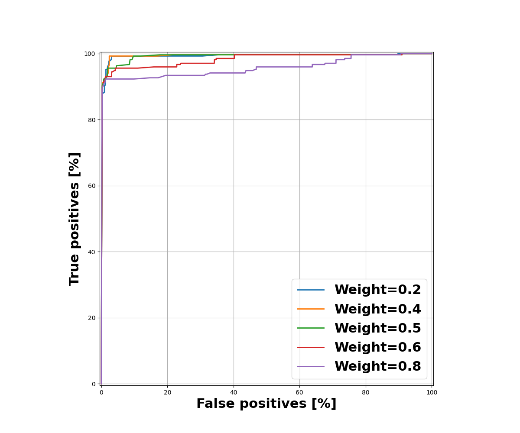
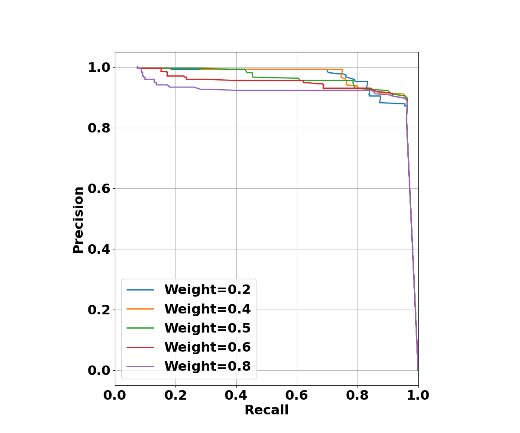
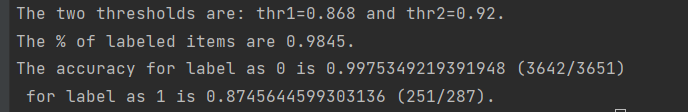
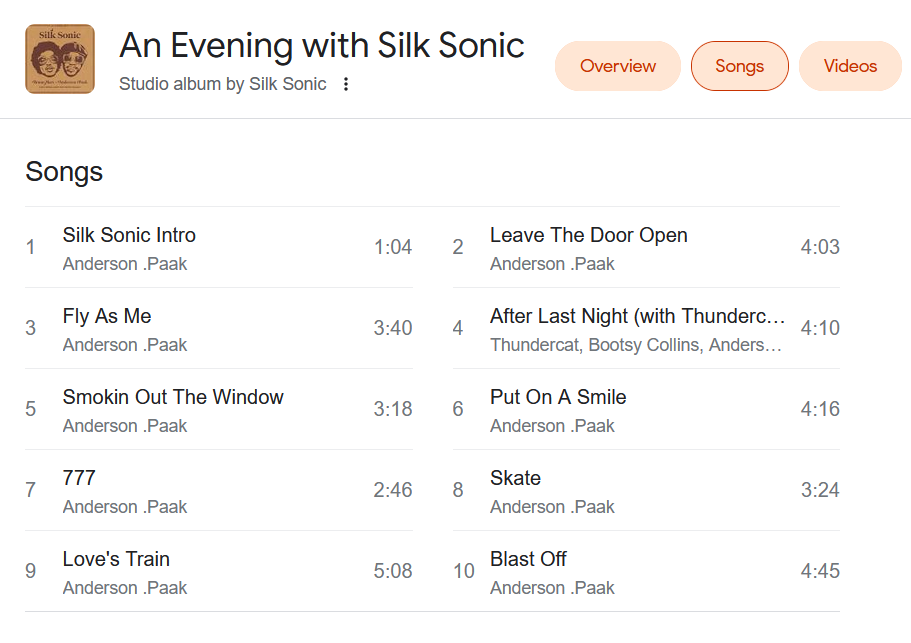
**Entity Matching Annotation Project**

1. **Introduction**:  
    Wavo data is a data set that contains the entries of campaign names from user input, and artist&track names in the database. The key task is to extract entity from the campaign name to match the entity of artist&track name in the database. The extraction of entity requires the annotation to the data. In our study, the fuzzy matching model is applied to establish a baseline model to automatically annotate the data. Fuzzy matching model evaluate the content in artist&track name and provide evaluation score compare to the campaign name.  
    In order to verify the result of fuzzy matching model, the manual annotation tool Doccano is used to create the ground truth label for the entries of campaign name and artist&track names. Then the fuzzy matching result is compared to the manually labeled result to understand how fuzzy matching model performs.
2. **PRC and ROC analysis**: (here takes 1000-5000 data for example)  
   
   1. **Goal1：**determine the weight percentage of evaluation function
      1. Evaluation function is to combine the artist fuzzy score and track fuzzy score with different weights.
      2. **Eval\_score = w\*artistFuzzyScore + (1-w)\*trackFuzzyScore**
      3. Here is to determine the weight **w** for the artist fuzzy score using the Prc and Roc curves.
      4. Area under the curve (AUC) score is applied to determine which weight **w** is better. Typically, the Prc and Roc will reach the same conclusion.
   2. **Goal2:** set 2 the proper **threshold** to categorize all the data into 3 parts:
      1. Label = 0: 0 <= eval\_score <= threshold1
      2. Label = uncertain: threshold1<eval\_score<threshold2
      3. Label = 1: threshold2 <= eval\_score <= 1
      4. \*The **current thresholds (empirical)** are set by:
3. Threshold1: precision = 0.9
4. Threshold2: recall = 0.97
5. **Evaluate result**：
   1. Console outputs:  
      
      1. The **percentage of data that are labeled as 0 or 1** are reported
      2. The **accuracy of label=0 and label=1** are both reported.
   2. In addition, the categorized data are summarized in three tables containing the data with label=0, label=uncertain, and label=1.
   3. In order to facilitate the error analysis, two more tables are generated：
      1. Incorrect label for label=0 (false negative)
      2. Incorrect label for label=1 (false positive)
   4. The following section summarizes all the problems when doing the fuzzy wuzzy and manual entity maching.
6. **Error analysis summary (for 0-8000 data)**:

|  |  |  |  |
| --- | --- | --- | --- |
| Campaign name | Artist name | Track name | Comment |
| WrIe - Silk Sonic - Silk Sonic 鈥?An Evening with Silk Sonic (Album) - P1 - R - 0062L00000WCVqnQAH [Ireland] | Bruno Mars, Anderson .Paak, Silk Sonic | Smoking Out The Window | Album vs Track name |

* 1. **Issue: track name vs album name**
     1. Campaign name here contains the album name, not the track name.
     2. From the search below, the album name is matching the artist name and the track name  
        
     3. \*This situation may have happened to many other inputs.
     4. Current handle: assign these inputs as Not Match, until the business people respond.

|  |  |  |  |
| --- | --- | --- | --- |
| Campaign name | Artist name | Track name | Comment |
| [A] WrNa - Warner Music Nashville - November YouTube Campaigns - P1 - R - 0062L00000WCWyiQAH | gnash | us | both artist and track names accidently match the content in campaign name |

* 1. **False Positive**:  
     1. As shown here, the artist and track names make no sense, but are both short and match in the campaign name.
     2. This suggests that **fuzzy scores with both 100 points are still not good enough to give label = 1.**
  2. **False Negative**:

|  |  |  |  |
| --- | --- | --- | --- |
| Campaign name | Artist name | Track name | Comment |
| Atln - Jasmine Thompson - Love Is Just A Word ft. Calum Scott - P1 - R - 0062L00000WCTxSQAX | Jasmine Thompson and Calum Scott | love is just a word | Artist name contains 2 musicians |
| PLUMITAS | LIKE | Plumas | Like | Possible a typo |

* + 1. Two cases:
       1. The first case is due to 2 musician names appear in the “artist name” entry, but “campaign name” only contains 1 musician name
       2. The second case is possibly due to a typo.

1. **Summary**:  
    Overall speaking, fuzzy matching model is very successful in doing the automatic annotation. Among the 8000 data that has been studies here, more than 90% of the data can be successfully categorized as either matching or not matching for the entries of campaign name and artist&track names. The accuracy is also very high (~97%) for the 8000 data. The fuzzy matching model, however, suffers from certain cases that leads to false positive and false negative results. These cases are mostly due to the accidental matching of text from the artist&track name to the campaign name (false positive) and due to the additional information provided in the artist&track name (false negative).  
    For the future study, a machine learning model should be established to further achieve the automatic annotation of the data entry. The machine learning model will be able to annotate a higher percentage of data with a lower error rate.