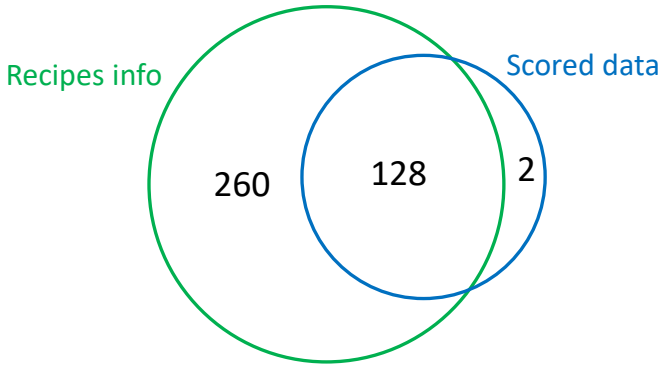
# Recipe Similarity Matrix

# Hybrid item-item similarity

Produce a hybrid similarity matrix for **all recipes** using both content based information and item-item similarity scores.

### Recipe ID crossover




Other data notes:

- Duplication
- Inconsistent missing field and labels
- Mixed type fields

### Important context

- Unknown proxy users
- Unknow recs engine
- Biased distribution

Score counts

| user_id | |
|---|---|
| 10001 | 139 |
| 10011 | 476 |
| 10123 | 341 |
| 10127 | 210 |
| 10141 | 1835 |
| 10145 | 269 |
| 10163 | 1068 |

Mean score/user

| user_id | |
|---|---|
| 10001 | 1.446043 |
| 10011 | 1.686975 |
| 10123 | 1.791789 |
| 10127 | 1.528571 |
| 10141 | 1.557493 |
| 10145 | 1.572491 |
| 10163 | 1.575843 |

### System considerations

- Unknown users
- Unknow recs engine
- Biased scores
- Data contracts are enforced
- Unit tests in place

# CB-CF with weighted switching

Unable to perform pure collaborative filtering or content based similarity, we need an alternative strategy to produce recommendations for all recipes.

| Method | Description | Pros | Cons |
|---|---|---|---|
| CF-X | Combine CF with another method i.e. clustering of recipes or user data to produce average similarities or some other similar method. | Well established<br>Scales well | Cold start<br>Sparsity |
| CB-X | Combine CB with another (non CF) method i.e. user data | Limited cold start<br>Controllable | Labelling/ content representation |
| **CB-CF** | **Commonly uses a switching or weighting methodology when data from either is not available** | **Simple** | **Calibration of both similarities**<br>**Introduces bias** |
| CB-CF-X | Utilises both forms of recommendations and combines with a third method i.e. matrix factorisation, cascading | Sparsity<br>Can limit cold starts | More complex |

# Weighted hybrid similarity score

The code provided is capable of producing multiple similarity matrices

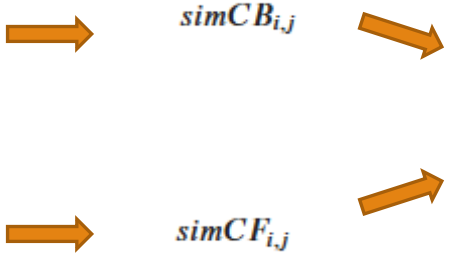
### Base similarities

**CB methods**

- Cosine Similarity – tfidf vectorisation (2 x tokenisation methods)
- Cosine Similarity – count vectorisation (2 x tokenisation methods
- Jaccard Coefficient

**CF methods**

- Cosine Similarity – Mean Score




### Hybridisation

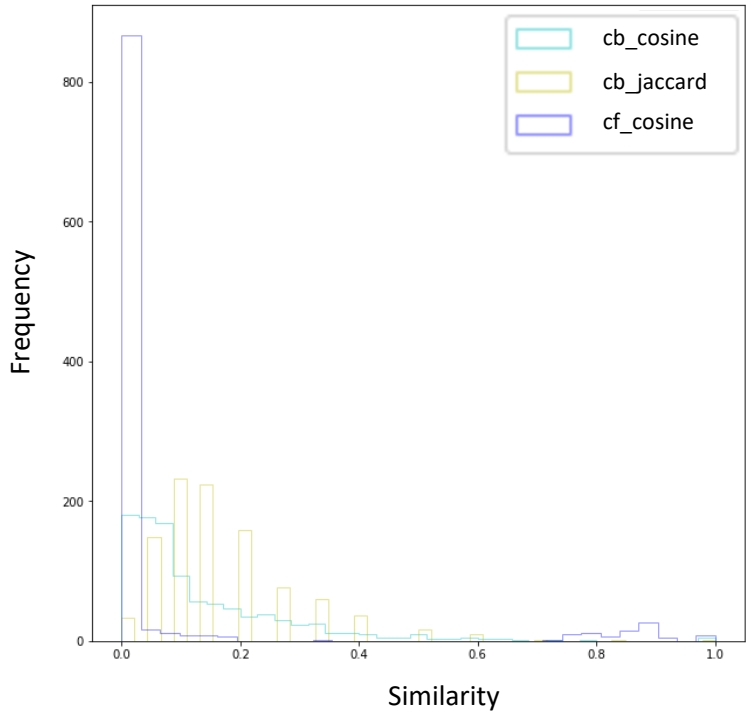
$$hybridSim = \alpha \times simCF_{i,j} + (1 - \alpha) \times simCB_{i,j}$$

The method implemented uses alpha to weight similarities when both scores occur and switches when there is only one score. Allows switching to one method only, if required.

# Some results

CF and CB similarities are similarly distributed and could be improved

**Distribution of similarity measures**




**Example Random Top 4 (alpha = 0.95)**

Ref Recipe

| | | | | | |
|---|---|---|---|---|---|
| **recipe_id** | 57 | 58 | 553 | 498 | 288 |
| **country** | japan | japan | japan | united kingdom | china |
| **country_secondary** | japan | japan | japan | united kingdom | china |
| **dish_category** | soups | soups | stove top / bowl food | stove top / bowl food | protein&veg |
| **dish_type** | ramen | ramen | stir fry | stir fry | fish & side veg |
| **diet_type** | meat | vegan | meat | meat | fish |
| **carbohydrate_base** | wholewheat noodle nests | wholewheat noodle nests | wholewheat noodle nests | thai rice noodles | brown rice |
| **carbohydrate_category** | noodles | noodles | noodles | noodles | rice |
| **protein** | chicken | tofu | chicken | pork | whitefish |
| **protein_cut** | skin off thigh chicken | NaN | breast chicken | steak pork | basa |
| **protein_type** | poultry & meat | vegetarian | poultry & meat | poultry & meat | fish & seafood |
| **family_friendly** | yes | no | no | no | no |
| **spice_level** | no spice | no spice | no spice | mild | mild |
| **prep_time** | 35 | 30 | 30 | 30 | 35 |
| **hybridSim** | 1 | 0.871405 | 0.861165 | 0.859478 | 0.859071 |

# Other examples of alpha

**Example Top 4 (alpha = 0, pure CB)**

Ref Recipe

| | | | | | |
|---|---|---|---|---|---|
| **recipe_id** | 57 | 285 | 553 | 1173 | 395 |
| **country** | japan | japan | japan | japan | japan |
| **country_secondary** | japan | japan | japan | japan | japan |
| **dish_category** | soups | soups | stove top / bowl food | soups | stove top / bowl food |
| **dish_type** | ramen | ramen | stir fry | ramen | stir fry |
| **diet_type** | meat | meat | meat | meat | meat |
| **carbohydrate_base** | wholewheat noodle nests | wholewheat noodle nests | wholewheat noodle nests | thai rice noodles | wholewheat noodle nests |
| **carbohydrate_category** | noodles | noodles | noodles | noodles | noodles |
| **protein** | chicken | chicken | chicken | pork | chicken |
| **protein_cut** | skin off thigh chicken | breast chicken | breast chicken | mince pork | breast chicken |
| **protein_type** | poultry & meat | poultry & meat | poultry & meat | poultry & meat | poultry & meat |
| **family_friendly** | yes | no | no | no | no |
| **spice_level** | no spice | mild | no spice | spicy | mild |
| **prep_time** | 35 | 25 | 30 | 35 | 35 |
| **hybridSim** | 1 | 0.6 | 0.5 | 0.411765 | 0.411765 |

**Example Top 4 (alpha = 1, pure CB)**

| | | | | | |
|---|---|---|---|---|---|
| **recipe_id** | 57 | 288 | 58 | 498 | 561 |
| **country** | japan | china | japan | united kingdom | china |
| **country_secondary** | japan | china | japan | united kingdom | china |
| **dish_category** | soups | protein&veg | soups | stove top / bowl food | protein&veg |
| **dish_type** | ramen | fish & side veg | ramen | stir fry | fish & side veg |
| **diet_type** | meat | fish | vegan | meat | fish |
| **carbohydrate_base** | wholewheat noodle nests | brown rice | wholewheat noodle nests | thai rice noodles | basmati |
| **carbohydrate_category** | noodles | rice | noodles | noodles | rice |
| **protein** | chicken | whitefish | tofu | pork | oily fish |
| **protein_cut** | skin off thigh chicken | basa | NaN | steak pork | salmon |
| **protein_type** | poultry & meat | fish & seafood | vegetarian | poultry & meat | fish & seafood |
| **family_friendly** | yes | no | no | no | no |
| **spice_level** | no spice | mild | no spice | mild | mild |
| **prep_time** | 35 | 35 | 30 | 30 | 25 |
| **hybridSim** | 1 | 0.901996 | 0.895597 | 0.894187 | 0.888435 |

# Closing thoughts

- Explore additional similarity measures and calibrations of  2+ similarities
- Further understanding of recommendation engine objective and experimentation capabilities to create a feedback loop from online recs
- Explore removal of user bias and/or use real user data
- Introduce a cascade method for latent taste similarities when more data is available
- Increase test coverage

# Thanks