Features

Attribute	**Features**	**Weight** (how much should the features for this attribute count towards our total similarity score?)
Product Name	Jaccard (word based) Jaccard (Qgram-based) TF/IDF Soft TF/IDF	High
Brand	TF/IDF Soft TF/IDF Levenshtein edit distance	High
Segment	Levenshtein edit distance	Low
Product Type	Levenshtein edit distance *Note: if this attribute and Segment are categorical, using exact match for them may be more useful. We can try it both ways!*	Low
UPC	Exact Match: 1 or 0	High (if matches, or both products have a value but it does not match) Low (if unavailable for one/both product)
GTIN	Exact Match: 1 or 0 *Note: want to verify that this one is actually useful.*	Not sure

For Reference - these are the string similarity measures available in AnHai's package:

Tokenizers

- Delimiter-based
- Qgram-based
- Word-based

String Similarity Measures

- Levenshtein
- Hamming
- Jaro
- Jaro Winkler
- Needleman Wunsch
- Smith Waterman
- Affine
- Jaccard
- Overlap Coefficient
- Cosine
- Monge Elkan
- TF/IDF
- Soft TF/IDF