Data Science Challenge

Methodology: Simulation framework

- The entries in `x_descriptions` of `train data` are converted into a set containing subsets of shingles. A MinHash is computed for set.
- The entries in `y_descriptions` of `train data` are treated in the same way.
- Similarity is computed between the MinHashs generated from `x_descriptions` and `y_descriptions` using Jaccard distance. The distance is a real number [0,1].
- The true labels for similarity are determined by `x_id` and `y_id`, where 0: no-match & 1: match.
- ROC type simulation analysis is performed to determine the Jaccard threshold above which the Jaccard distances are deemed similar.
- The performance of the technique is examined using `test data`, using the jaccard threshold determined from the `train data` set

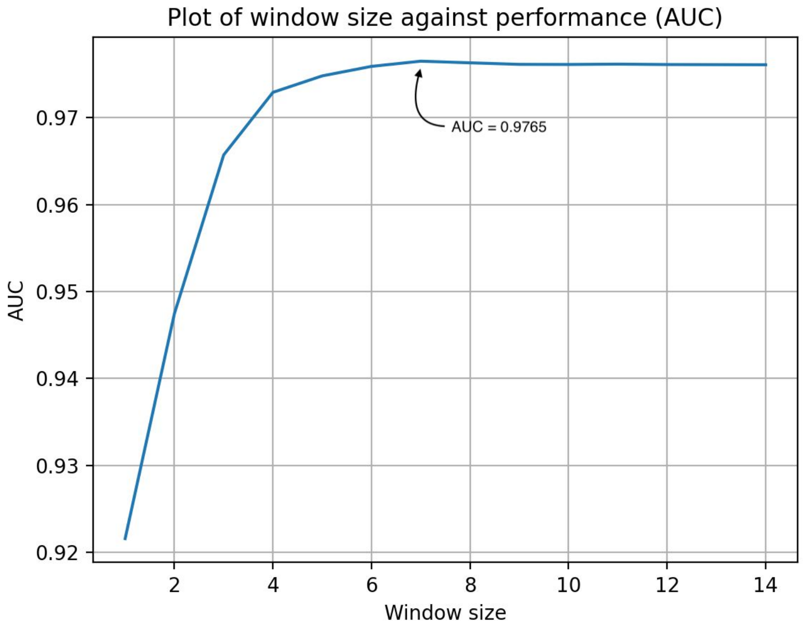
Methodology: Converting descriptions into sets

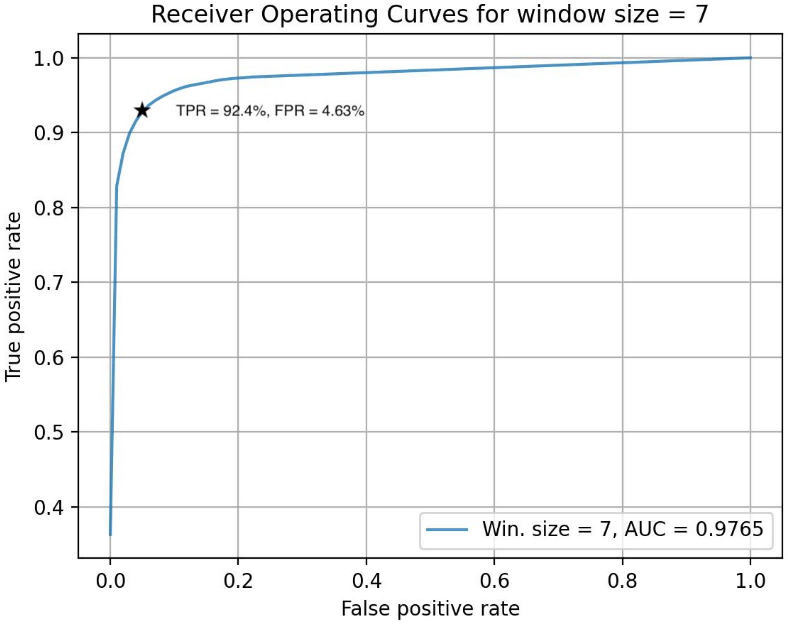
Word-level shingles

- Words are identified using spaces, everything is lower-case, punctuations are removed and a sliding window is applied to create subsets of words.
- Example:
  “If you can meet with Triumph and Disaster And treat those two impostors just the same”

  Window size = 3:
  [[if, you, can], [you, can, meet], [can, meet, with] , . . . , [impostors, just, the], [just, the, same]]
- We examine the performance across different window sizes, and select the window with the highest area under the curve (AUC).

Results: Word-level shingles

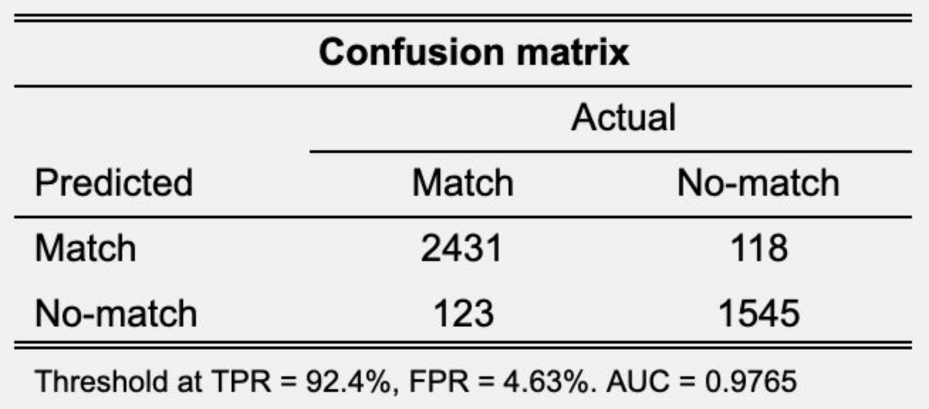




Results: Word-level shingles

- A window size = 7 yielded maximum auc of 0.9765 in train data and was applied on test data.

  The threshold was selected at TPR = 92.4% and FPR = 4.63%.
- The confusion matrix analyzes performance of the technique on test data.



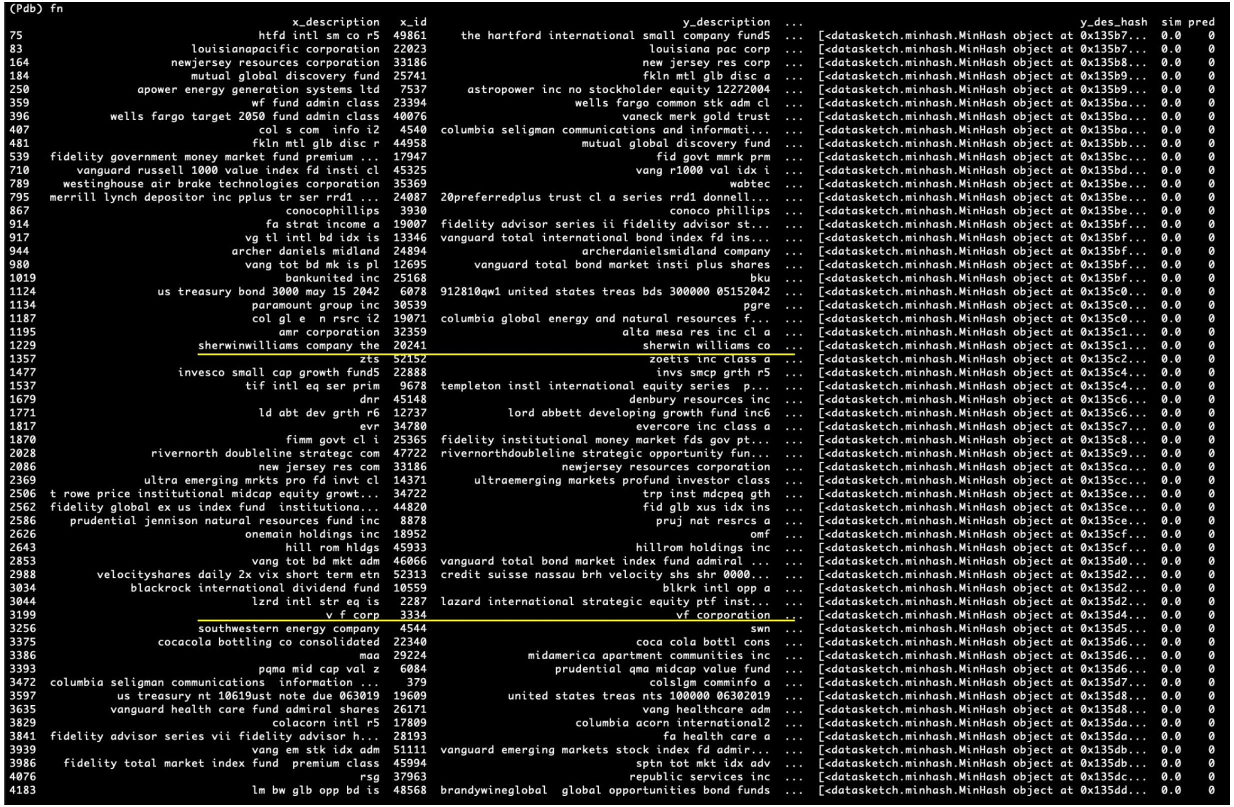
Confusion matrix

	Actual	
Predicted	Match	No-match
Match	2431	118
No-match	123	1545

Threshold at TPR = 92.4%, FPR = 4.63%. AUC = 0.9765

- We observe 241 mis-classifications over 5762 data points

Results: Examining false negatives



```
(Pdb) fn
                                          x_description   x_id                                          y_description  ...                                         y_des_hash  sim pred
75                                   htfd intl sm co r5  49861         the hartford international small company fund5  ...  [<datasketch.minhash.MinHash object at 0x135b7...  0.0     0
83                         louisianapacific corporation  22023                                     louisiana pac corp  ...  [<datasketch.minhash.MinHash object at 0x135b7...  0.0     0
164                     newjersey resources corporation  33186                                    new jersey res corp  ...  [<datasketch.minhash.MinHash object at 0x135b8...  0.0     0
184                        mutual global discovery fund  25741                                    fkln mtl glb disc a  ...  [<datasketch.minhash.MinHash object at 0x135b9...  0.0     0
250                apower energy generation systems ltd   7537          astropower inc no stockholder equity 12272004  ...  [<datasketch.minhash.MinHash object at 0x135b9...  0.0     0
359                                 wf fund admin class  23394                          wells fargo common stk adm cl  ...  [<datasketch.minhash.MinHash object at 0x135ba...  0.0     0
396            wells fargo target 2050 fund admin class  40076                                 vaneck merk gold trust  ...  [<datasketch.minhash.MinHash object at 0x135ba...  0.0     0
407                                  col s com  info i2   4540     columbia seligman communications and informati...  ...  [<datasketch.minhash.MinHash object at 0x135ba...  0.0     0
481                                 fkln mtl glb disc r  44958                           mutual global discovery fund  ...  [<datasketch.minhash.MinHash object at 0x135bb...  0.0     0
539   fidelity government money market fund premium ...  17947                                      fid govt mmrk prm  ...  [<datasketch.minhash.MinHash object at 0x135bc...  0.0     0
710       vanguard russell 1000 value index fd insti cl  45325                                   vang r1000 val idx i  ...  [<datasketch.minhash.MinHash object at 0x135bd...  0.0     0
789     westinghouse air brake technologies corporation  35369                                                 wabtec  ...  [<datasketch.minhash.MinHash object at 0x135be...  0.0     0
795   merrill lynch depositor inc pplus tr ser rrd1 ...  24087     20preferredplus trust cl a series rrd1 donnell...  ...  [<datasketch.minhash.MinHash object at 0x135be...  0.0     0
867                                      conocophillips   3930                                        conoco phillips  ...  [<datasketch.minhash.MinHash object at 0x135be...  0.0     0
914                                   fa strat income a  19007     fidelity advisor series ii fidelity advisor st...  ...  [<datasketch.minhash.MinHash object at 0x135bf...  0.0     0
917                                vg tl intl bd idx is  13346     vanguard total international bond index fd ins...  ...  [<datasketch.minhash.MinHash object at 0x135bf...  0.0     0
944                              archer daniels midland  24894                           archerdanielsmidland company  ...  [<datasketch.minhash.MinHash object at 0x135bf...  0.0     0
980                                vang tot bd mk is pl  12695           vanguard total bond market insti plus shares  ...  [<datasketch.minhash.MinHash object at 0x135bf...  0.0     0
1019                                     bankunited inc  25168                                                    bku  ...  [<datasketch.minhash.MinHash object at 0x135bf...  0.0     0
1124                  us treasury bond 3000 may 15 2042   6078      912810qw1 united states treas bds 300000 05152042  ...  [<datasketch.minhash.MinHash object at 0x135c0...  0.0     0
1134                                paramount group inc  30539                                                   pgre  ...  [<datasketch.minhash.MinHash object at 0x135c0...  0.0     0
1187                                col gl e  n rsrc i2  19071     columbia global energy and natural resources f...  ...  [<datasketch.minhash.MinHash object at 0x135c0...  0.0     0
1195                                    amr corporation  32359                                 alta mesa res inc cl a  ...  [<datasketch.minhash.MinHash object at 0x135c1...  0.0     0
1229                        sherwinwilliams company the  20241                                    sherwin williams co  ...  [<datasketch.minhash.MinHash object at 0x135c1...  0.0     0
1357                                                zts  52152                                     zoetis inc class a  ...  [<datasketch.minhash.MinHash object at 0x135c2...  0.0     0
1477                     invesco small cap growth fund5  22888                                      invs smcp grth r5  ...  [<datasketch.minhash.MinHash object at 0x135c4...  0.0     0
1537                               tif intl eq ser prim   9678     templeton instl international equity series  p...  ...  [<datasketch.minhash.MinHash object at 0x135c4...  0.0     0
1679                                                dnr  45148                                  denbury resources inc  ...  [<datasketch.minhash.MinHash object at 0x135c6...  0.0     0
1771                                 ld abt dev grth r6  12737                lord abbett developing growth fund inc6  ...  [<datasketch.minhash.MinHash object at 0x135c6...  0.0     0
1817                                                evr  34780                                   evercore inc class a  ...  [<datasketch.minhash.MinHash object at 0x135c7...  0.0     0
1870                                     fimm govt cl i  25365     fidelity institutional money market fds gov pt...  ...  [<datasketch.minhash.MinHash object at 0x135c8...  0.0     0
2028                 rivernorth doubleline strategc com  47722     rivernorthdoubleline strategic opportunity fun...  ...  [<datasketch.minhash.MinHash object at 0x135c9...  0.0     0
2086                                 new jersey res com  33186                        newjersey resources corporation  ...  [<datasketch.minhash.MinHash object at 0x135ca...  0.0     0
2369                ultra emerging mrkts pro fd invt cl  14371           ultraemerging markets profund investor class  ...  [<datasketch.minhash.MinHash object at 0x135cc...  0.0     0
2506  t rowe price institutional midcap equity growt...  34722                                    trp inst mdcpeq gth  ...  [<datasketch.minhash.MinHash object at 0x135ce...  0.0     0
2562  fidelity global ex us index fund  institutiona...  44820                                    fid glb xus idx ins  ...  [<datasketch.minhash.MinHash object at 0x135ce...  0.0     0
2586     prudential jennison natural resources fund inc   8878                                      pruj nat resrcs a  ...  [<datasketch.minhash.MinHash object at 0x135ce...  0.0     0
2626                               onemain holdings inc  18952                                                    omf  ...  [<datasketch.minhash.MinHash object at 0x135cf...  0.0     0
2643                                     hill rom hldgs  45933                                   hillrom holdings inc  ...  [<datasketch.minhash.MinHash object at 0x135cf...  0.0     0
2853                                vang tot bd mkt adm  46066     vanguard total bond market index fund admiral ...  ...  [<datasketch.minhash.MinHash object at 0x135d0...  0.0     0
2988         velocityshares daily 2x vix short term etn  52313     credit suisse nassau brh velocity shs shr 0000...  ...  [<datasketch.minhash.MinHash object at 0x135d2...  0.0     0
3034              blackrock international dividend fund  10559                                       blkrk intl opp a  ...  [<datasketch.minhash.MinHash object at 0x135d2...  0.0     0
3044                                lzrd intl str eq is   2287     lazard international strategic equity ptf inst...  ...  [<datasketch.minhash.MinHash object at 0x135d2...  0.0     0
3199                                           v f corp   3334                                         vf corporation  ...  [<datasketch.minhash.MinHash object at 0x135d4...  0.0     0
3256                        southwestern energy company   4544                                                    swn  ...  [<datasketch.minhash.MinHash object at 0x135d5...  0.0     0
3375                  cocacola bottling co consolidated  22340                                   coca cola bottl cons  ...  [<datasketch.minhash.MinHash object at 0x135d6...  0.0     0
3386                                                maa  29224                   midamerica apartment communities inc  ...  [<datasketch.minhash.MinHash object at 0x135d6...  0.0     0
3393                                 pqma mid cap val z   6084                       prudential qma midcap value fund  ...  [<datasketch.minhash.MinHash object at 0x135d6...  0.0     0
3472  columbia seligman communications  information ...    379                                     colslgm comminfo a  ...  [<datasketch.minhash.MinHash object at 0x135d7...  0.0     0
3597            us treasury nt 10619ust note due 063019  19609                united states treas nts 100000 06302019  ...  [<datasketch.minhash.MinHash object at 0x135d8...  0.0     0
3635           vanguard health care fund admiral shares  26171                                    vang healthcare adm  ...  [<datasketch.minhash.MinHash object at 0x135d8...  0.0     0
3829                                   colacorn intl r5  17809                          columbia acorn international2  ...  [<datasketch.minhash.MinHash object at 0x135da...  0.0     0
3841  fidelity advisor series vii fidelity advisor h...  28193                                       fa health care a  ...  [<datasketch.minhash.MinHash object at 0x135da...  0.0     0
3939                                vang em stk idx adm  51111     vanguard emerging markets stock index fd admir...  ...  [<datasketch.minhash.MinHash object at 0x135db...  0.0     0
3986    fidelity total market index fund  premium class  45994                                   sptn tot mkt idx adv  ...  [<datasketch.minhash.MinHash object at 0x135db...  0.0     0
4076                                                rsg  37963                                  republic services inc  ...  [<datasketch.minhash.MinHash object at 0x135dc...  0.0     0
4183                                lm bw glb opp bd is  48568      brandywineglobal  global opportunities bond funds  ...  [<datasketch.minhash.MinHash object at 0x135dd...  0.0     0
```

- We examine the false negatives to understand scenarios that are not being identified as a match.

- We posit another technique of creating subsets from `descriptions`, based on characters (as opposed to words).

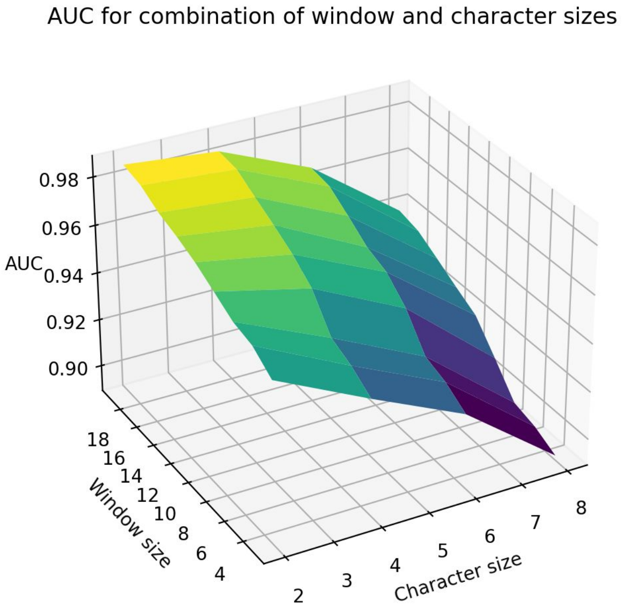
Methodology: Converting descriptions into sets

Character-level shingles

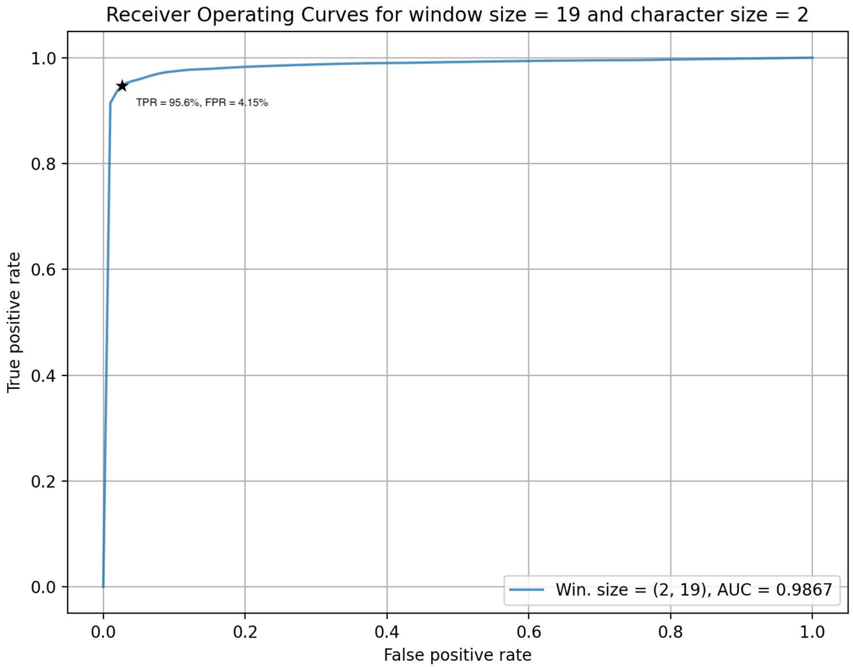
- Spaces are removed, everything is lower-case, punctuations are removed and a sliding window is applied to create subsets of characters.
- Example:
  "If you can meet with Triumph and Disaster And treat those two impostors just the same"

  Character size = 2, Window size = 5:
  [[if, fy, yo, ou, uc], [fy, yo, ou, uc, ca], [yo, ou, uc, ca, an], . . . , [he,es,sa,am,me]]
- We examine the performance across different character sizes and window sizes, and select the combination with the highest area under the curve (AUC).

Results: Character-level shingles



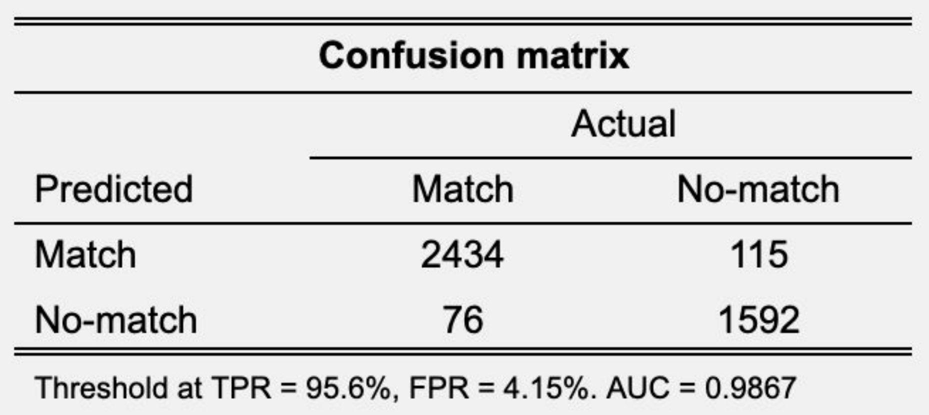
AUC for combination of window and character sizes



Receiver Operating Curves for window size = 19 and character size = 2

Results: Character-level shingles

- A window size = 9 and character size = 2 yielded maximum auc of 0.9867 in train data and was applied on test data. The threshold was selected at TPR = 95.6% and FPR = 4.15%.
- The confusion matrix analyzes performance of the technique on test data.



Confusion matrix

	Actual	
Predicted	Match	No-match
Match	2434	115
No-match	76	1592

Threshold at TPR = 95.6%, FPR = 4.15%. AUC = 0.9867

- We observe that character-level shingles has a better performance than word-level shingles, both in terms of reducing false negatives and false positives.

- A more rigorous search could be performed, for example across a more granular and across a larger span of character and window size combinations, to optimize for AUC.