## metrics for determining similarities between two strings.

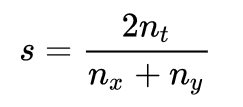
String metric

<https://en.wikipedia.org/wiki/String_metric>

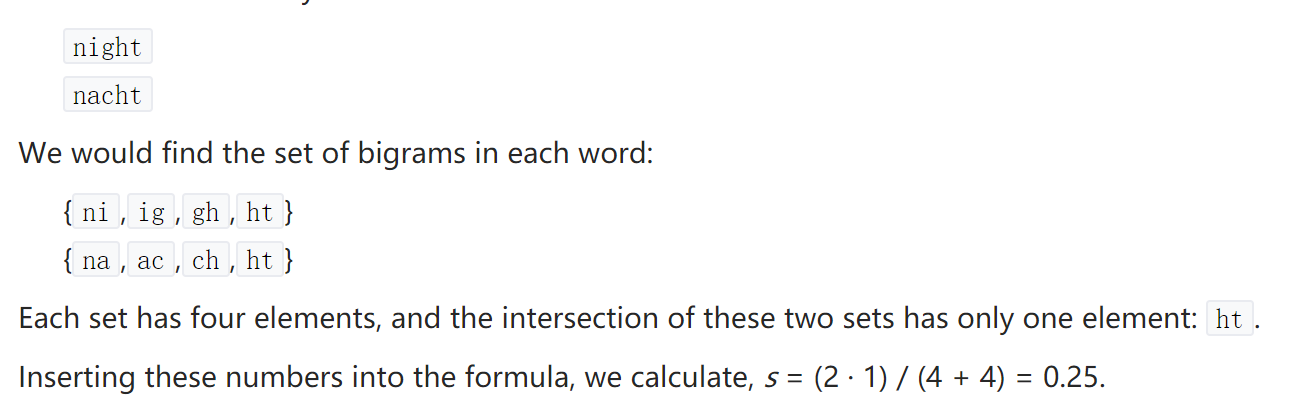
recommended: without building a corpus

* [Sørensen–Dice coefficient](https://en.wikipedia.org/wiki/S%C3%B8rensen%E2%80%93Dice_coefficient)

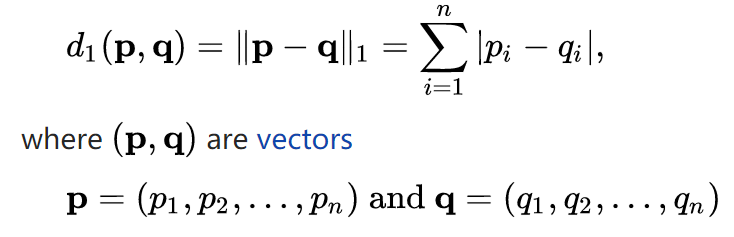
twice the shared information (intersection) over the sum of cardinalities.



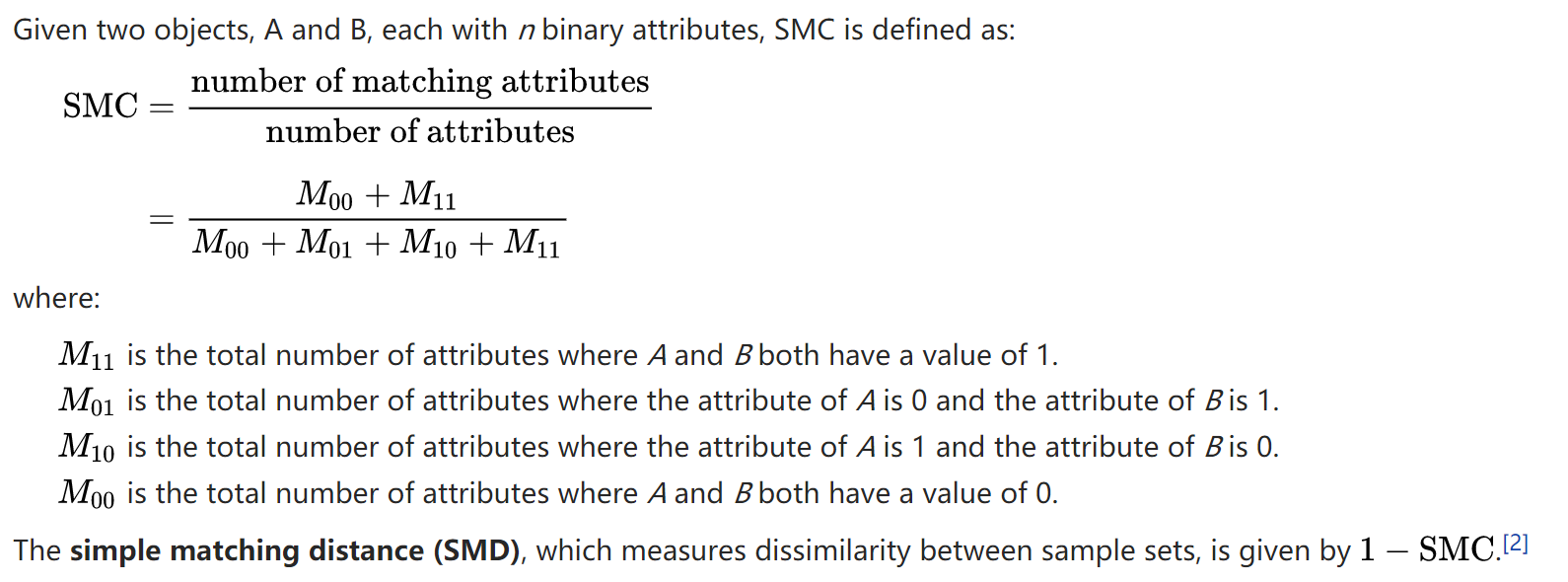
nt is the number of character bigrams found in both strings, nx is the number of bigrams in string x and ny is the number of bigrams in string y



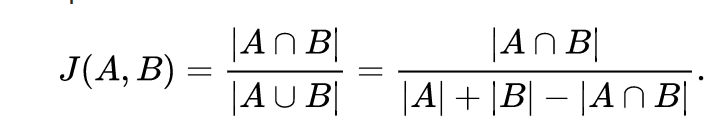
* [Block distance](https://en.wikipedia.org/wiki/Block_distance) or [L1 distance](https://en.wikipedia.org/wiki/L1_distance) or [City block distance](https://en.wikipedia.org/wiki/City_block_distance)

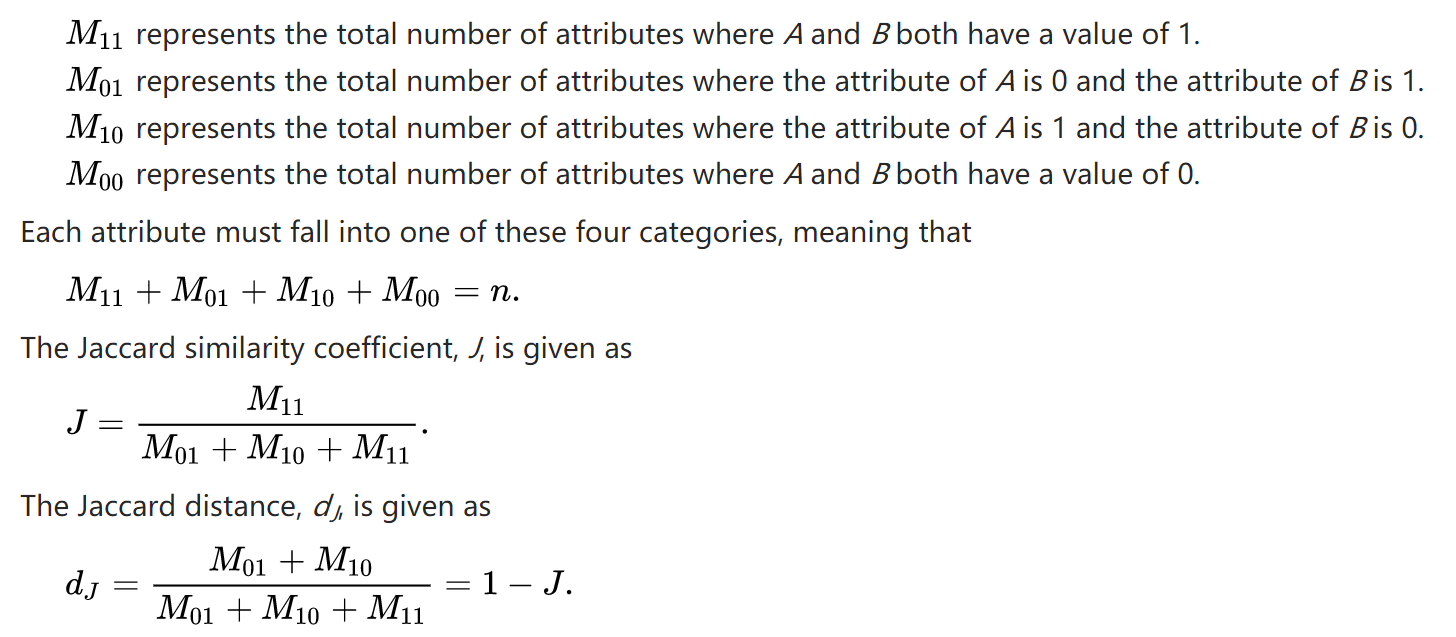


* Simple matching coefficient (SMC)



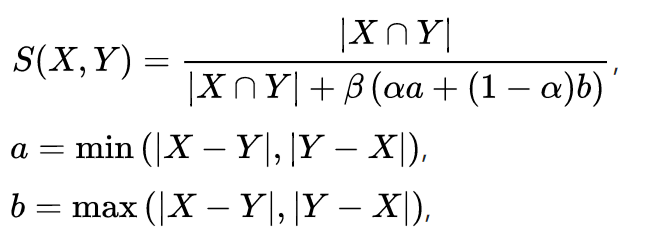
* [Jaccard similarity](https://en.wikipedia.org/wiki/Jaccard_similarity) or [Jaccard coefficient](https://en.wikipedia.org/wiki/Jaccard_coefficient) or [Tanimoto coefficient](https://en.wikipedia.org/wiki/Tanimoto_coefficient" \o "Tanimoto coefficient)



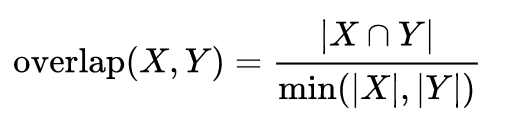


* [Tversky index](https://en.wikipedia.org/wiki/Tversky_index)

a variant of the original formulation:

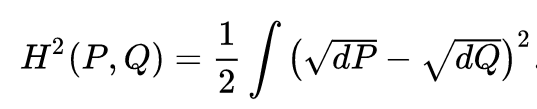


* [Overlap coefficient](https://en.wikipedia.org/wiki/Overlap_coefficient)

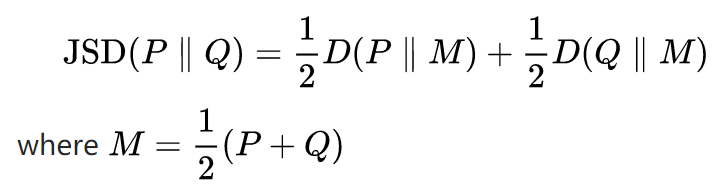


* [Variational distance](https://en.wikipedia.org/w/index.php?title=Variational_distance&action=edit&redlink=1)
* [Hellinger distance](https://en.wikipedia.org/wiki/Hellinger_distance) or [Bhattacharyya distance](https://en.wikipedia.org/wiki/Bhattacharyya_distance)

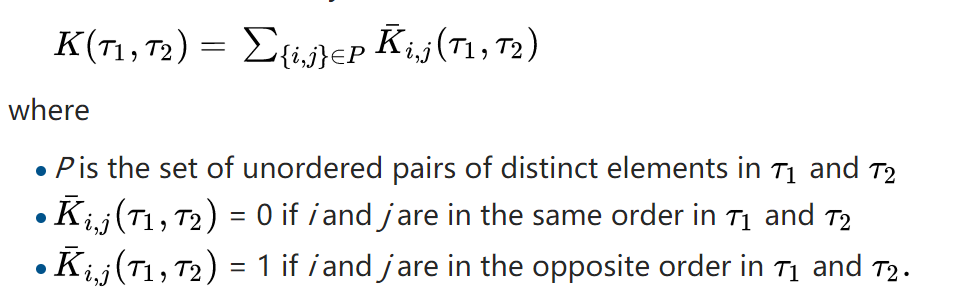
P and Q denote two probability measures that are absolutely continuous



* [Information radius](https://en.wikipedia.org/wiki/Information_radius) ([Jensen–Shannon divergence](https://en.wikipedia.org/wiki/Jensen%E2%80%93Shannon_divergence))



* [Skew divergence](https://en.wikipedia.org/w/index.php?title=Skew_divergence&action=edit&redlink=1)
* [Confusion probability](https://en.wikipedia.org/w/index.php?title=Confusion_probability&action=edit&redlink=1)
* [Tau metric](https://en.wikipedia.org/wiki/Kendall_tau_distance), an approximation of the [Kullback–Leibler divergence](https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence" \o "Kullback–Leibler divergence)

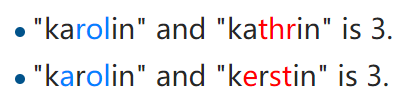


* [Fellegi and Sunters metric](https://en.wikipedia.org/w/index.php?title=Fellegi_and_Sunters_metric&action=edit&redlink=1) (SFS)
* [Maximal matches](https://en.wikipedia.org/w/index.php?title=Maximal_matches&action=edit&redlink=1)
* [Grammar-based distance](https://en.wikipedia.org/w/index.php?title=Grammar-based_distance&action=edit&redlink=1)
* [TFIDF](https://en.wikipedia.org/wiki/Tf%E2%80%93idf) distance metric
* Hamming distance

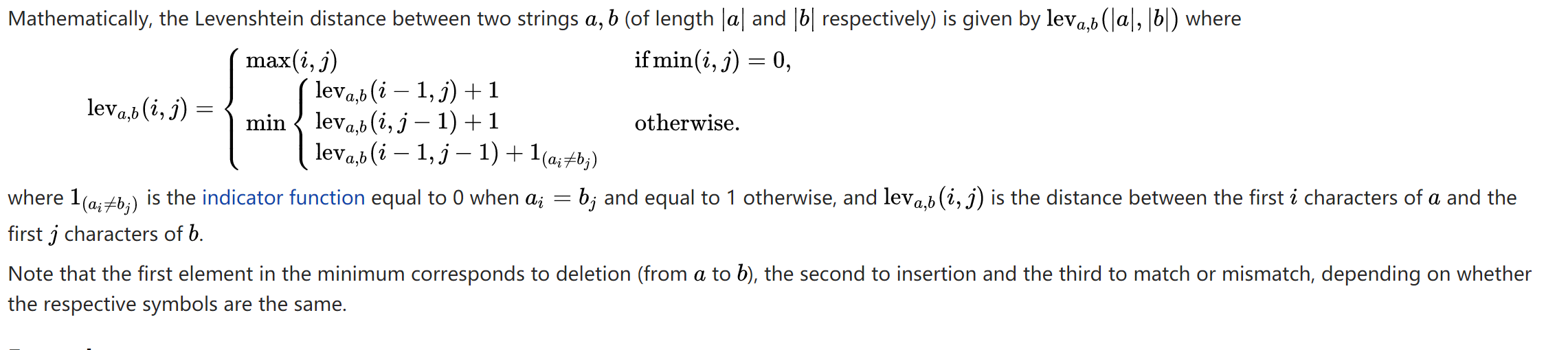
between two strings of equal length is the number of positions at which the corresponding symbols are different.

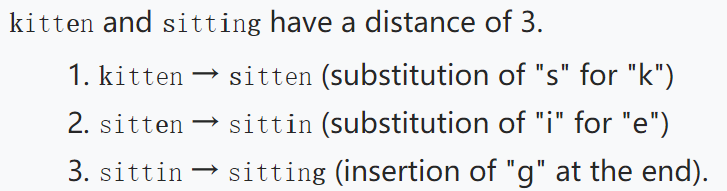
it measures the minimum number of substitutions required to change one string into the other,

or the minimum number of errors that could have transformed one string into the other.

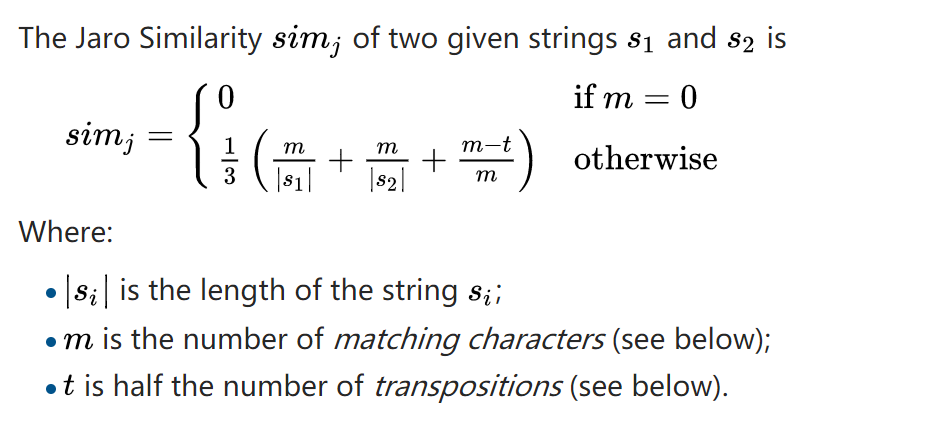


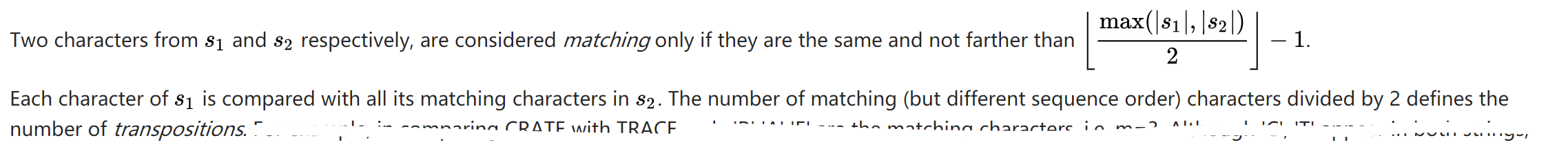
* Levenshtein distance and Damerau–Levenshtein distance

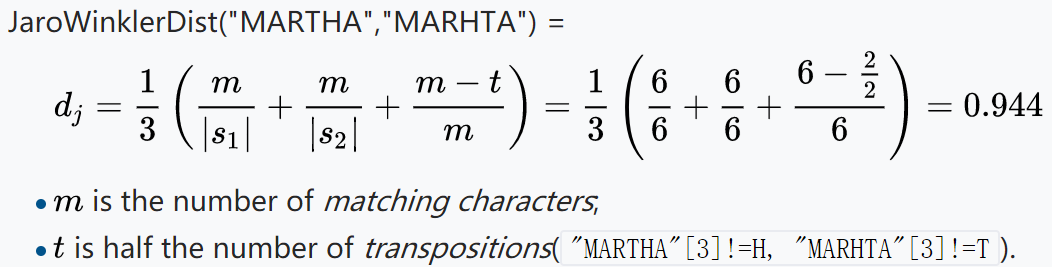




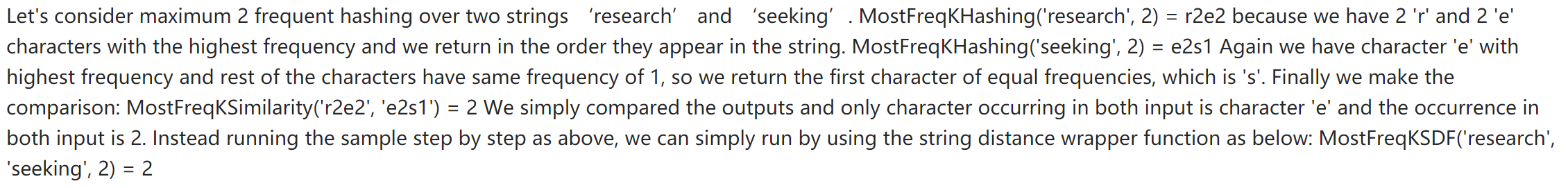
* Jaro–Winkler distance

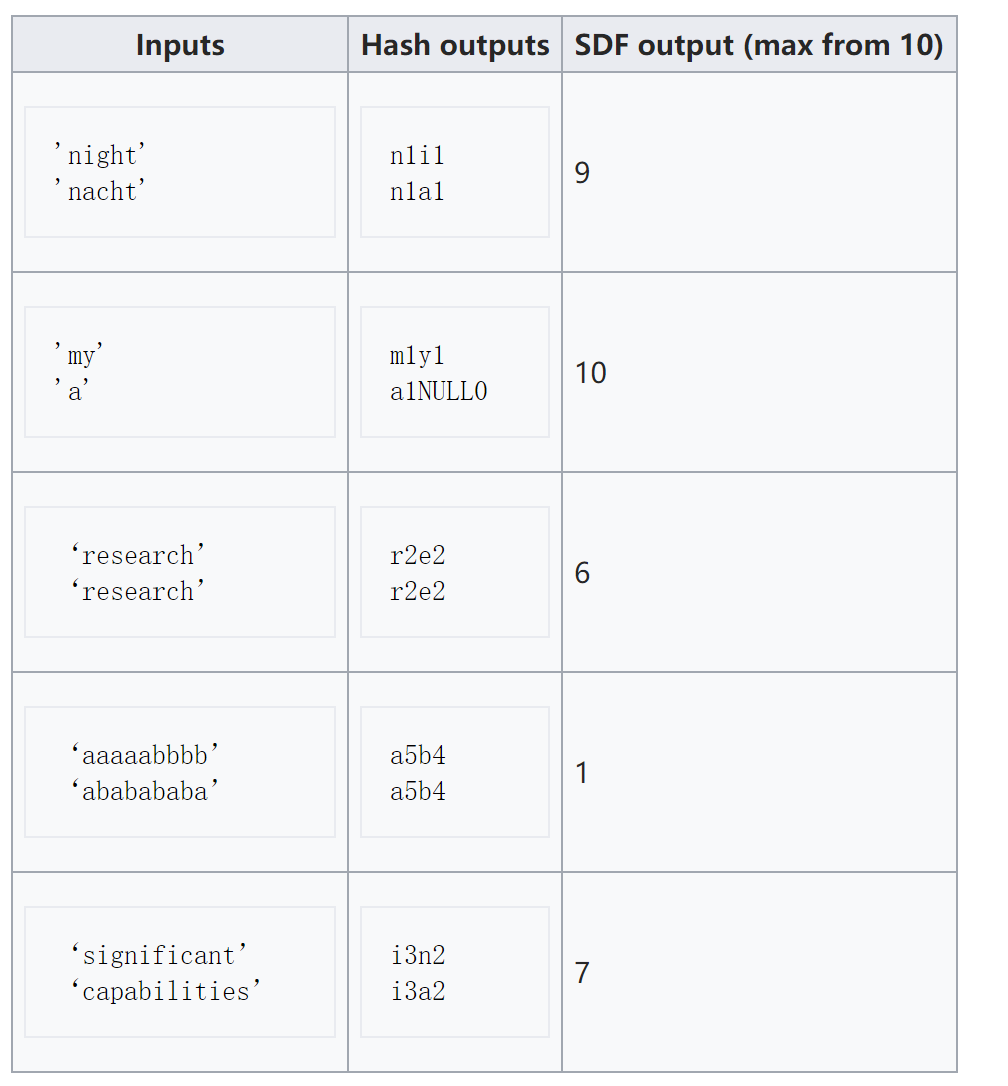






* Most frequent k characters





* Smith–Waterman algorithm

## clustering text-based data

1)

text

bag-of-words count vectors

vector space using tf-idf

cosine similarity

clustering using the k-means

2)

text

bag-of-words count vectors

lda(Latent Dirichlet Allocation)/ Latent Semantic Analysis(Latent Semantic Indexing)

cluster according to probability

3)

text

bag-of-words count vectors

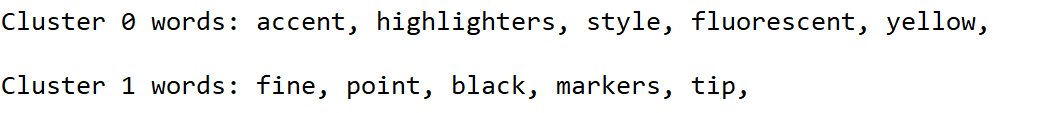
vector space using tf-idf

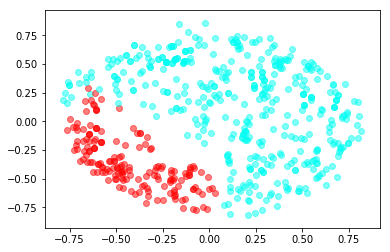
cosine similarity

hierarchical clustering

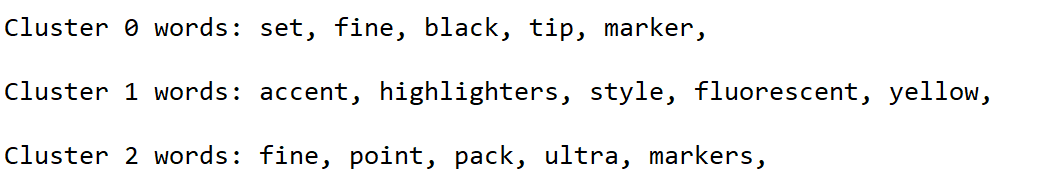
## cluster with k-means

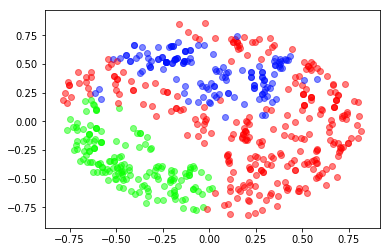
#### k=2



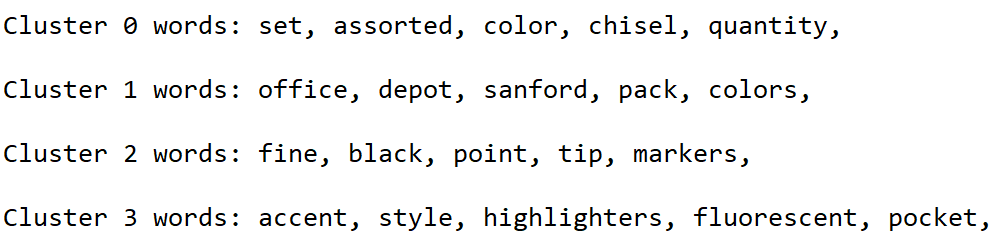


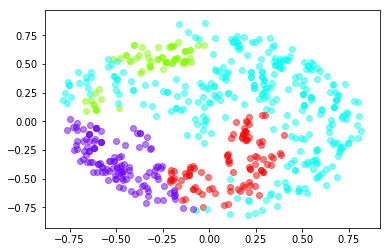
#### k=3



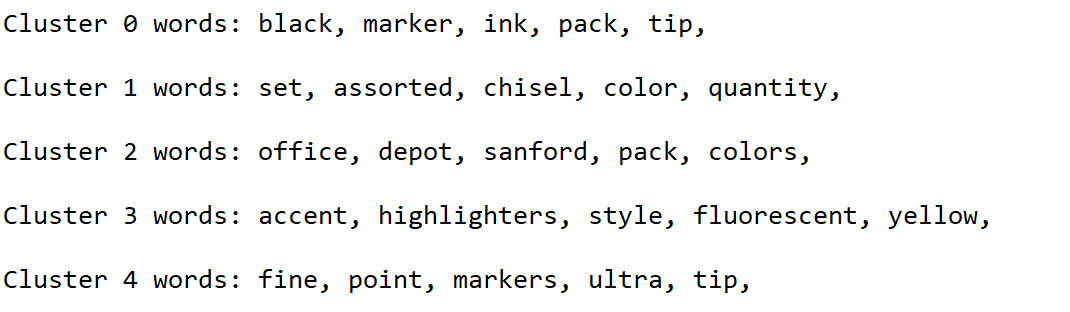


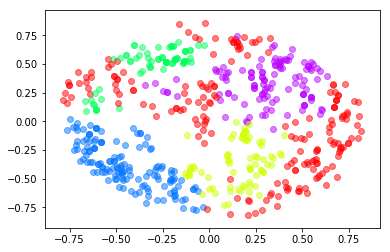
#### k=4



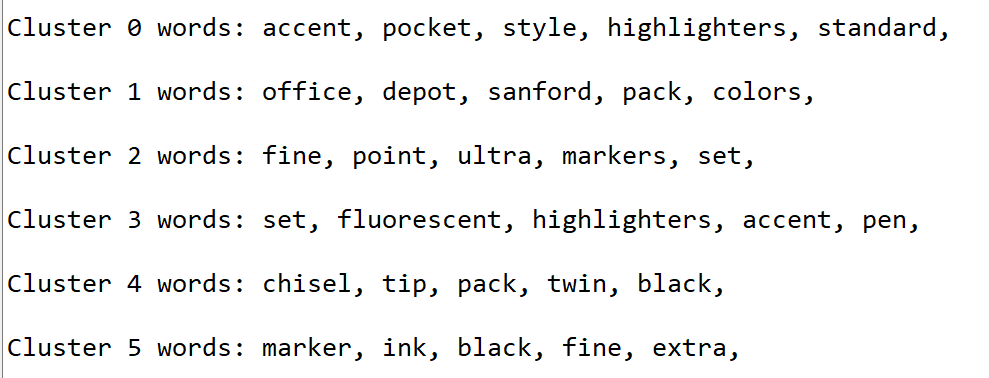


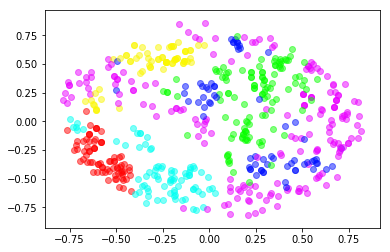
#### k=5



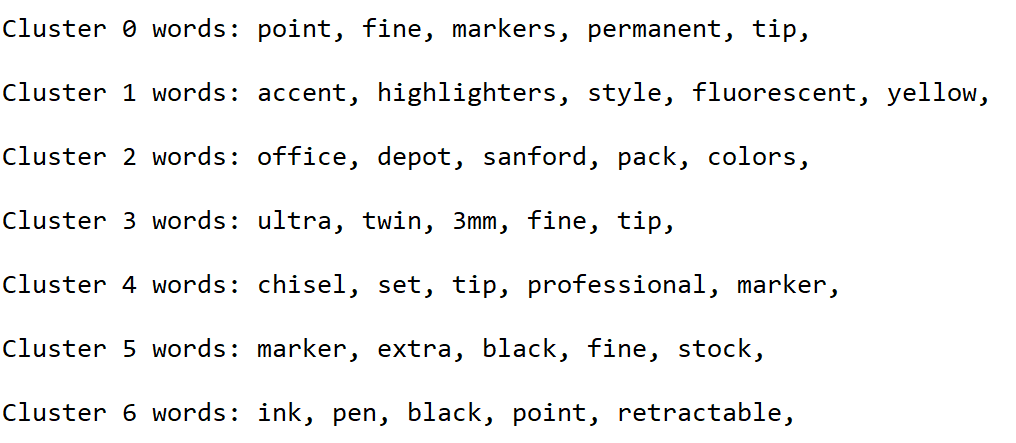


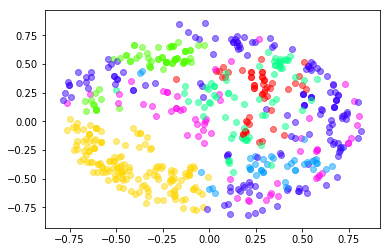
#### k=6





#### k=7





#### k=8

