Study guide

String matching and similarity metrics

Why isn't exact string matching enough for IR? State at least 3 good reasons.

What does OCR stand for, and what does an OCR error look like? Can you give some realistic examples?

List and define the four types of similarity metrics.

What is the main difference between sequence-based and set-based similarity metrics?

What is the edit distance between the strings in the following pairs:

(john, jon)
(mighty, iffy)
(flight, frighten)
(jonathan, jon)
(mary-kate, marianne)

Which string similarity measure is a generalization of Levenstein? Where would it be more appropriate than Levenstein?

Complete the missing entries in the table below (in some cases, more than one word may be necessary but keep as brief as possible):

	Levenshtein	**Needleman-Wunch**
Costs	1	matrix
Operations		gaps
Result	distance	

Advanced:

Choose one of the well-known string similarity metrics (e.g., edit, Jaro-Winker etc.) that we studied in class, and write down an algorithm in pseudocode to compute the metric given two strings. What is the time complexity of the algorithm? Is this algorithm appropriate for computing long strings, or would you only save it for short strings?

We are trying to compute Levenstein distance between two strings x and y, represented as sequences of characters as shown below. Choose from the correct option in the third column of the table below.

$x_1x_2 \ldots x_{i-1}x_i$ is a prefix of x $\quad$ $y_1y_2 \ldots y_{j-1}y_j$ is a prefix of y

Case	Distance	Operation
$x_i = y_j$	d(i-1,j-1)	[keep, delete, insert, replace] x_i
x_i != y_j	d(i-1,j) + 1	[keep, delete, insert, replace] x_i
	d(i,j-1) + 1	[keep, delete, insert, replace] y_j [after, with] x_i
	d(i-1,j-1) + 1	[keep, delete, insert, replace] x_i [after, with] y_j

What are some important differences between Smith-Waterman and Needleman-Wunch? When would you use one over the other?

Give the formula for the Jaro similarity measure. What applications are best suited for it?

Calculate Jaro for the strings below:

(dickens, dixon)
(mighty, iffy)
(flight, frighten)
(jonathan, jon)

Three of the examples above were provided earlier also for edit distance. Comparing Jaro vs. edit distance scores on these examples, which one is more preferrable? Does it depend on the example?

What are some advantages of hybrid string similarity measures?

What is the generalized Jaccard measure? What practical problem can it address that the ordinary Jaccard has trouble with?

Describe the 'soft' component of soft tf-idf measure. What practical problem can it address that the ordinary tf-idf has trouble with?

Given two tf-idf vectors, how would you compare whether they are 'close' together or not? Give the formula and qualitatively describe the intuition.

Is using Euclidean distance as a measure of distance between two tf-idf vectors a good idea or bad idea (and why)?

Types of Data

What is the difference between nominal and ordinal variables? Give examples of each

What is the difference between interval and ratio variables? Give examples of each

Data processing

What is tokenization and what is another word for it? Give two realistic examples below, and show clearly how you are dealing with punctuation.

Give the formula for Jaccard. Is it sequence based or set based? Are you more likely to use it for comparing pairs of sentences or pairs of words?

Imagine you are processing social media data (you may assume Twitter). What are some other data preprocessing steps you need to execute after tokenization? How (if any) would you change the tokenization itself to accommodate the special characteristics of social media data?

Missing, Duplicate and Inconsistent Values

List two different ways of dealing with the missing value problem and the pros and cons of each. Use examples.

Similarly, give examples of cases where inconsistent and duplicate values might occur and how you would deal with them.

What do we mean by an 'imbalanced' data in machine learning parlance? What are some ways to deal with it?

Feature aggregation

Briefly discuss feature aggregation, using real-world examples. Why might we want to use it?

How is feature aggregation different from feature sampling?