Clustering 1

Important concepts

- The objective in this chapter are to understand text similarity and clustering.
- Some of the important concepts are
  - Information Retrieval (IR)
  - Document similarity measures
  - Machine learning algorithm

Similarity Measures

Text Similarity

- Is to analyze and measure how two entities of text are close or far from each other

- The text similarity can be classified broadly into the following two areas:

  - Lexical Similarity

  - Semantic Similarity

Lexical Similarity

- This method involved observing the contents of the text document with regard to syntax, structure and contents and measuring their similarity based on these parameters
- This is classified into two broad areas
    - Term Similarity
    - Document Similarity

Term Similarity

- Is the measure of similarity between individual words.
- The application of term similarity is with autocorrect.
- Some of the term similarity techniques are
    - Jaccard similarity
    - Cosine distance and similarity
    - Levenshtein Distance
    - Hamming Distance
    - Manhattan Distance
    - Euclidean Distance

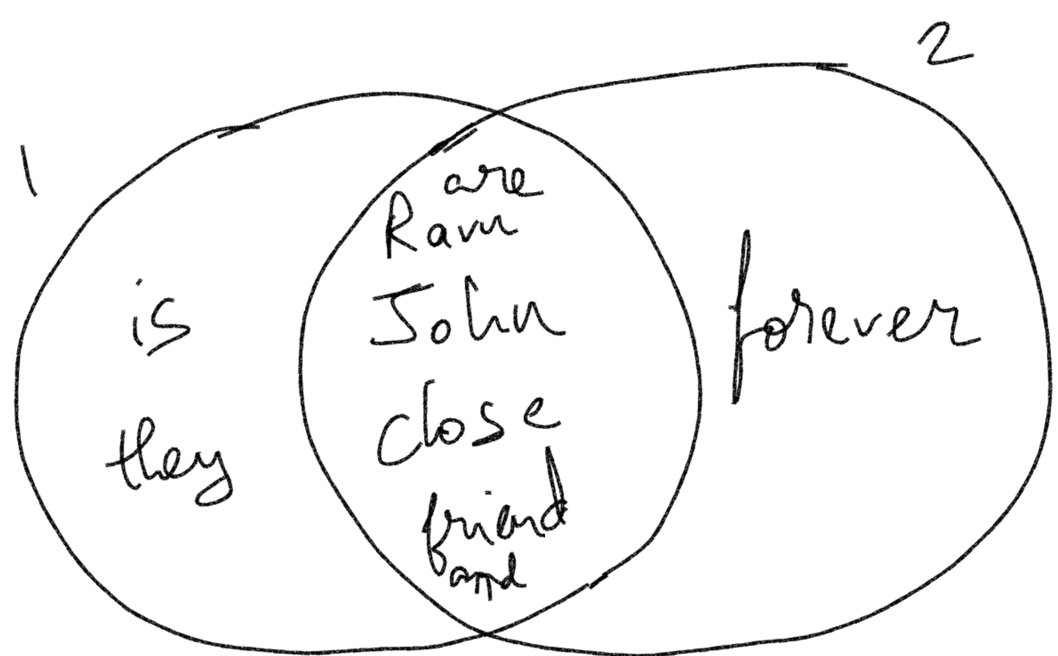
Text Similarity cont...

Jaccard Similarity

of two documents: $$js(u, v) = \frac{Size\ of\ intersection}{Size\ of\ union}$$

Sentence 1: Ram is John's friend and they are close

Sentence 2: Ram and John are close friends forever



Text Similarity cont...

Cosine Distance and Similarity

- Cosine similarity gives us the measure of the angle between the terms (in form of vector). Cosine distance is a metric derived from cosine similarity.
  - Cos 0 -> vectors are close to each other
  - Cos 90 -> indicated the terms are unrelated
  - Cos 180 -> indicates the terms are completely opposite

$$||u\,.\,v|| = ||u|| * ||v|| cos\ \theta$$

Text Similarity cont...

Levenshtein Edit Distance

- Is defined as minimum number of edits needed in the form of addition, substitution or deletions one convert from one word to another.

- Length of the strings need not be equal

$$ld_{u,v}(i,j) = \begin{cases} max(i,j) & if\ \ min(i,j) = 0 \\ min \begin{cases} ld_{u,v}(i-1,j)+1 \\ ld_{u,v}(i,j-1)+1 \\ ld_{u,v}(i-1,j-1)+C_{u_i \neq v_j} \end{cases} & otherwise \end{cases}$$

$$C_{u_i \neq v_j} = \begin{cases} 1 & if\ u_i \neq v_j \\ 0 & if\ \ u_i = v_j \end{cases}$$

Text Similarity cont...

Hamming Distance

- It is the distance measured between two strings under the assumptions that they are of equal length.
- Here we check the number positions that they have different characters or symbols

$$hd(u, v) = \sum_{i}^{n} (u_i \neq v_i) \qquad norm\_hd(u, v) = \frac{\sum_i^n (u_i \neq v_i)}{n}$$

Text Similarity cont...

Manhattan Distance

- Similar to hamming distance, instead of counting the number of matches, we subtract the mismatch at that position.

$$md(u, v) = ||u - v||_1 = \sum_i^n |u_i - v_i| \qquad norm\_md(u, v) = \frac{||u - v||_1}{n} = \frac{\sum_i^n |u_i - v_i|}{n}$$

Euclidean Distance

- Square of the distance between two vectors. Similar to Manhattan distance.

$$ed(u, v) = ||u - v||_2 = \sqrt{\sum_i^n (u_i - v_i)^2}$$