Report
Name: Cun Shi

EX1: Both lower-casing and lemmatization aim to reduce to the wordform and return the base or dictionary form of a word, which is known as lemma. This work can greatly improve the accuracy and efficiency of analyzing and programming.

EX2: The number of common words usually are the subset of N, therefore $dist_{jarcard} = \frac{common\ word \cap N}{commonword \cup N} = \frac{commonword}{N}$

EX3: Similarity(x,y) = 1 – distance(x,y). Because distance algorithm is symmetric. Thus Similarity algorithm is symmetric as well.