Lecture 6 Report

Ruochen Kong

1 Problem 1

The **cosine similarity** is **0.3975**, and **jaccard similarity** is **0.0369**.

2 Problem 2

The taggers used in this problem are :

1. Unigram tagger

2. Bigram tagger

3. Bigram tagger with the unigran tagger as the first backoff-tagger and the default tagger as the second backoff-tagger

The result shows that the unigram tagger has an accuracy of **0.8121** over validation dataset, the bigram tagger has an accuracy of **0.1021**, and the combined tagger has an accuracy of **0.8452**

Based on the performance, the trained combined tagger is used to tag the text1 file. The resulted tags showed no difference between the original text and the lower cased text.

3 Problem 3

Need vectorize embeddings, text tagging, vocabularies, models, and evaluation methods.