W-Shingling for Wikipedia Document Evolution Study

CSCI 8790 Class Project

In this project, you will develop a system based on the W-Shingling technique discussed in the class to study the evolution patterns of Wikipedia pages. Specifically, you will track the similarity between the current version of a Wikipedia page and specific previous versions of the page.

Data Collection:

The students of the class will collectively create a corpus containing plain text dump of the Wikipedia pages of the cities in GA. Specifically, each of you will select 4 cities in the US (https://en.wikipedia.org/wiki/Category:Cities_in_the_United_States_by_state). For each city you will create a text-only dump containing the following versions. Let us represent the current version of a city "C" as W^C_T. You will create a dump containing the following versions -- $<V^C_T, V^C_{(T-3)}, V^C_{(T-6)}, V^C_{(T-9)}, \ldots. V^C_{(T-147)}>$. This corpus will be used for the project described below.

System Development:

You will develop a system that will incorporate the following functionalities:

1. Given a pair of W and λ values, calculate the (w, λ) shingles for each version of each page.
2. Calculate the Jaccard similarity between the current version and each of the $V^C_{(T-3)}$, $V^C_{(T-6)}$, $V^C_{(T-9)}, \ldots. V^C_{(T-147)}$ versions of the page.

Experiments:

In addition to developing the above system, you will conduct the following experiments:

1. Execute the system with following (w, λ) pairs – (25, 8), (25, 16), (25, 32), (25, 64), (25, ∞), (50, 8), (50, 16), (50, 32), (50, 64), (50, ∞), where ∞ signifies the case of using all shingles (rather than the minimum λ shingles). Identify the λ value that comes closest to the ∞ case.
2. Plot the similarities between the current version and each of the $V^C_{(T-3)}$, $V^C_{(T-6)}$, $V^C_{(T-9)}, \ldots. V^C_{(T-147)}$ versions of the page with X axis being the version and the Y axis being the Jaccard similarity.
3. Measure the time for calculating the shingles for the entire corpus for each of the above (w, λ) pairs. Plot the timings with x-axis as the λ value and Y axis being the timing. Different w values form different lines. Mention any interesting observations.

Things to Note:

1. You can use https://wikitext.eluni.co/ for extracting text from the Wikipedia page. Alternatively, you can write your own text extractor.
2. Use MD5 for hashing.
3. Submit the code and a report containing the experimental setup, graphs and observations/discussions.

4. Your project will be evaluated on the correctness of the code, the thoroughness of the experiments (you should perform multiple runs and report averages for timing experiments), appropriateness of the graph formats, and the intuitiveness and validity of the observations.