Information Retrieval

LSH IMPLEMENTATION

PLAGIARISM CHECK ON LEGAL CASE FILES

Aim

To build a plagiarism checker which can detect the similarity between 2 documents using the Locality Sensitive Hashing technique.

Dataset

This dataset contains Australian legal cases from the Federal Court of Australia (FCA). The cases were downloaded from AustLII (http://www.austlii.edu.au). We included all files from the year 2006,2007,2008 and 2009.

Data Structures Used

1. **total_doc:**

   List of documents with all the text in each xml file extracted using BeautifulSoup. Later modified to contain normalized sentences after tokenization,stemming and lemmatization.

2. **shingle_set :**

   A dictionary containing the list of k-shingles as its keys and a binary array of size equal to the number of documents which contains 1 if the shingle occurs in the doc and 0 if it doesn't.

3. **shingleMap:**

   A dictionary containing shingleID from shingle_set as key, mapped to the respective k-shingle.

4. **hash_funcs:**

   A list of N_HASHES (10) hash functions which contain a unique permutation of the document IDs in an array in each row.

5. **Signature Matrix:**

   A 2D array of size N_HASHES x size_of_total_doc which is filled using the LSH algorithm.

Specifications

- K-shingles = 2
- Number of Hash Functions = 10
- Number of documents = 400
- Number of rows per band = 2
- Bucket Size = 3

Distance Measures Used

- Euclidean Distance

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- Hamming Distance

$$\delta(x_i, y_i) = \begin{cases} 0 & x_i = y_i \\ 1 & x_i \neq y_i \end{cases}$$

Hamming Distance = $\sum \delta(xi,yi)$

- Jaccard Similarity

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

- Cosine Similarity

$$\frac{x \bullet y}{\sqrt{x \bullet x}\sqrt{y \bullet y}}$$

Pre-processing Done

- The documents were tokenized using a RegexpTokenizer
- The stopwords were then removed from these tokens using NLTK's stopset
- The resulting set was then stemmed using a PorterStemmer
- This was then lemmatized using a WordNetLemmatizer

Runtime

- Import: 6.0541746 s
- Preprocess: 2.5622439999999997 s
- Shingling: 0.046281300000000414 s
- Hashing: 0.06588400000000005 s
- Signature Matrix: 23.098033199999996 s
- Buckets and Bands: 0.006540499999999838 s
- Retrieval: 8.922462999999993 s

Team

- Raaed Ahmed - 2018A7PS0218H
- Ritika Reddy - 2018A7PS1224H
- Anurag Aribandi - 2018A7PS1218H
- Anvitha Nallan - 2018A7PS1214H