Shreayan Chaudhary
March 1, 2023
Homework #2b Information Retrieval 601.666

The outputs of the writeup questions are attached as .tsv files:

1. List the top 20 retrieved documents for queries 6, 9 and 22 by their number, title and similarity measure, with the “relevant” documents starred: top_20_docs.tsv
2. For the top 10 retrieved documents, show the terms on which the retrieval was based (those with non-zero weights for both query and retrieved document) along with these weights: top_10_docs.tsv
3. List the top 20 documents that are most similar to Documents 239, 1236 and 2740, giving number, title and similarity measure: most_similar.tsv

Raw term frequency (TF) weighting, TF-IDF weighting, and Boolean weighting are three commonly used weighting techniques. Cosine similarity, Dice similarity, Jaccard similarity, and Overlap similarity are four commonly used similarity metrics. Raw, unstemmed tokens, tokens stemmed by the Porter stemmer, excluding stopwords from term vectors, and including all tokens, including punctuation, are all techniques used in document relevance ranking. Here's how they compare:

1. Raw TF weighting:
    - Assigns a weight to each term in a document based on its frequency in that document.
    - Does not take into account the distribution of a term across the corpus, which means it can be misled by common words that appear frequently across many documents.

2. TF-IDF weighting:
    - Assigns a weight to each term in a document based on its frequency in that document and its frequency across the entire corpus.
    - Reduces the importance of terms that appear frequently across the corpus (like "the", "and", etc.) and boosts the importance of terms that appear frequently in a document but rarely in other documents.

3. Boolean weighting:
    - Assigns a weight of 1 to a term if it appears in a document and 0 if it does not.
    - Does not consider the frequency of a term in a document, which means it cannot differentiate between documents that contain many occurrences of a given term and those that contain only one.
    - This technique can be useful for simple queries where only the presence or absence of a term matters.

4. Cosine similarity:
    - Measure of the similarity between two vectors of term weights.

    - Calculates the cosine of the angle between the two vectors, which ranges from -1 (perfectly dissimilar) to 1 (perfectly similar).

5. Dice similarity:
    - Measure of the similarity between two sets of terms.
    - Calculates the ratio of the number of common terms to the total number of terms in the two sets.
    - Suitable for short texts and tends to produce accurate ranking results.

6. Jaccard similarity:
    - Measure of the similarity between two sets of terms.
    - Calculates the ratio of the number of common terms to the total number of unique terms in the two sets.
    - Suitable for longer texts and tends to produce accurate ranking results.

7. Overlap similarity:
    - Measure of the similarity between two sets of terms.
    - Calculates the ratio of the number of common terms to the smaller of the two sets.
    - Tends to produce good results for shorter texts.

8. Raw, unstemmed tokens:
    - Easy to implement and tends to produce accurate ranking results for short texts or when the task requires exact matches of specific words.
    - Does not take into account variations of the same word, such as plural forms, verb tenses, or misspellings, which can hurt the accuracy of the ranking.

9. Tokens stemmed by the Porter stemmer:
    - Reduces variations of the same word to their root form, which can improve the accuracy of the ranking by grouping related words together.
    - Can also produce some false positives or false negatives due to overstemming or understemming.

10. Excluding stopwords from term vectors:
    - Can help reduce the noise in the data and improve the accuracy of the ranking by focusing on more informative words.
    - However, it can also remove some useful information, particularly for longer texts, where some stopwords may carry meaning.

11. Including all tokens, including punctuation:
    - Can capture some contextual information that may be lost by other techniques.
    - However, it can also introduce a lot of noise in the data and produce a less accurate ranking.

When to pick one permutation over another:

- If you are working with a small dataset and looking for simple queries, then Boolean weighting may be sufficient.
- If you are working with a large corpus and need to handle complex queries, then TF-IDF weighting is generally the best choice as it considers both term frequency and inverse document frequency.
- Raw TF weighting can be a good choice when you are experimenting with the algorithm, or when you want to see how terms behave in isolation without considering their corpus-wide distribution.
- Cosine similarity is the most commonly used similarity measure in information retrieval and text mining applications, and it tends to produce accurate ranking results in most cases.
- Dice similarity and Jaccard similarity are suitable for short and longer texts, respectively. Therefore, you can use Dice similarity for shorter documents or queries, and Jaccard similarity for longer documents or queries.
- Overlap similarity is simple and easy to implement. Therefore, you can use it when you have limited computational resources or when you need to process large datasets quickly.
- Raw, unstemmed tokens can be used for short texts or when exact matches of specific words are required.
- Tokens stemmed by the Porter stemmer can be used for longer texts or when variations of the same word need to be grouped together.
- Excluding stopwords from term vectors can be used for longer texts or when the focus is on more informative words.
- Including all tokens, including punctuation, can be used when capturing contextual information is important, but it should be used with caution to avoid introducing too much noise in the data.

In summary, the choice of similarity measure depends on the nature of the data and the specific requirements of the task. A combination of techniques may also be used to balance the benefits and drawbacks of each approach.