情報システム論実習 テキスト分析演習課題 07 月 09 日分

6930318812 沖野 雄哉

2020 年 7 月 9 日

1　利用した表現手法

ベクトルベースの手法の中から BoW と tf-idf を用いて表現した。

2　利用した距離尺度

以下の二つの距離尺度を利用した。

- ユークリッド距離
- コサイン類似度

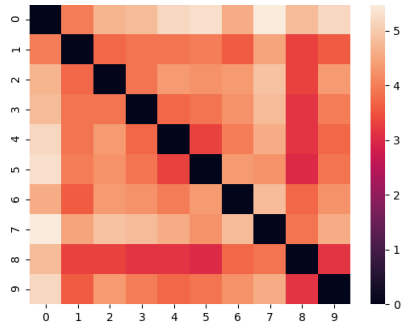
3　入力ドキュメント

以下の文書を対象に分析を行った。

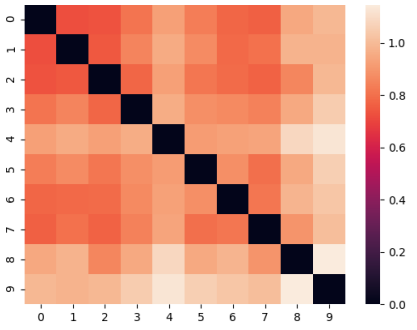
- View update is an important mechanism that allows updates on a view by translating them into the corresponding updates on the base relations.
- The existing literature has shown the ambiguity of translating view updates.
- To address this ambiguity, we propose a robust language-based approach for making view update strategies programmable and validatable.
- Specifically, we introduce a novel approach to use Datalog to describe these update strategies.
- We present a fragment of the Datalog language for which our validation is both sound and complete.
- We propose a validation algorithm to check the well-behavedness of the written Datalog programs.
- This fragment not only has good properties in theory but is also useful for solving practical view updates.
- Furthermore, we develop an algorithm for optimizing user-written programs to efficiently implement updatable views in relational database management systems.
- We have implemented our proposed approach.
- The experimental results show that our framework is feasible and efficient in practice.

4　実行結果

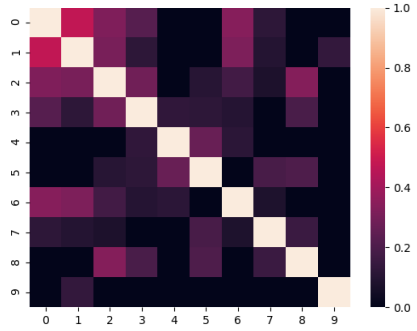
ドキュメント同士の距離を計算し、ヒートマップで表現した結果次のようになった。



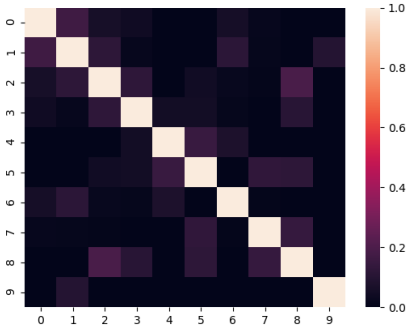
(a) BoW, ユークリッド距離



(b) tf-idf, ユークリッド距離



(c) BoW, コサイン類似度



(d) tf-idf, コサイン類似度

5　実行結果の考察

コサイン類似度は大きな違いはないが、全体的に類似度が下がっていると捉えられる。これは、類似度が大きい二つの文書は同じ語をドキュメントに含んでいるため、tfidf での表現はベクトルのノルムが小さくなり、コサイン類似度が小さくなったと考えられる。

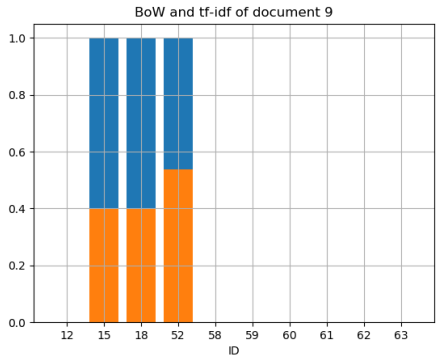
また、ユークリッド距離については、全体的な傾向として距離関係が逆転しているように見えた。文書 9、10 に着目して、それぞれの tfidf の値を出力した結果が以下の図である。文書 9 は文書が小さく 3 語しかないため、全体的に tf の値が大きくなり、tf-idf の値が大きくなっている。また文書 10 は文書 10 にしか含まれない単語を多く含んでいるが、語数が多いため、tf-idf 値が小さくなっている。BoW におけるユークリッド距離は単語数が少ない文書については小さくなる傾向があるが、上記の理由から tf-idf で表現した場合は単語数が少ない文書の方がユークリッド距離が大きくなり、距離関係の逆転が起こっているのではないかと考えられる。

6　感想

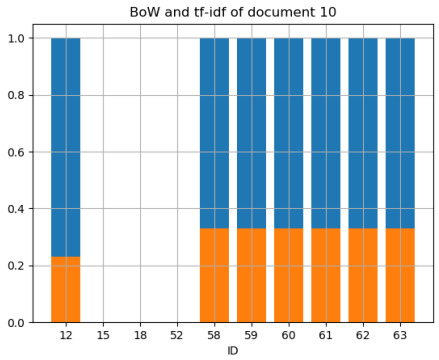
文書を単語の集合として表現することでおおまかに文書の類似度を測ることができると分かった。直感的には時系列的な情報が全てなくなっている状態でも大丈夫なのか気になる。文書表現に時系列情報を含める手法があるのか、その場合だとどのように変わるのか知りたい。ただ時系列を考慮すると、大量のデータを処理して類似する文書を取得したいケースなどを考えると、計算量の問題がありそうだとも思うので一長一短あるとも思う。また文書の類似度をヒューリスティックにでも評価することは難しいと感じたので、どのような評価指標が用いられているのか調べたい。

7　ソースコード

`https://github.com/oky-123/soc_info_mining/blob/master/1/documents_similarity.py`



(a) blue: BoW, orange: tf-idf



(b) blue: BoW, orange: tf-idf